UNIVERSITÉ DE NANTES
UFR DES SCIENCES PHARMACEUTIQUES ET BIOLOGIQUES

ÉCOLE DOCTORALE BIOLOGIE-SANTÉ

Année 2013

Contamination des logements par le plomb : Prévalence des logements à risque et Identification des déterminants de la contamination

THÈSE DE DOCTORAT

Discipline : Biologie, médecine et santé

Spécialité : Biostatistique

Présentée
et soutenue publiquement par

Jean-Paul LUCAS

le 30 octobre 2013, devant le jury ci-dessous

Président du jury	Pr.	Nathalie Seta, Université Paris Descartes
Rapporteur	Pr.	Nathalie Seta, Université Paris Descartes
Rapporteur	Pr.	Hervé Cardot, Université de Bourgogne
Examinateur	Dr.	Robert Garnier, Centre Antipoison de Paris
Examinateur	Dr.	Guillaume Chauvet, Ensai, Rennes
Invitée	Mme.	Corinne Mandin, CSTB, Marne-la-Vallée
Directrice de thèse	Pr.	Véronique Sébille-Rivain, Université de Nantes
Co-directrice de thèse	Dr.	Lise Bellanger-Husi, Université de Nantes

Table des matières

Avant-propos

Avant de commencer ce travail, le plomb m'était déjà très familier. En effet, quand on est pêcheur, le plomb est quelque chose que l'on manipule très souvent. En particulier lorsque l'on est pêcheur au coup ce qui est plus communément appelé « pêcheur à la ligne ». Lorsque l'on construit cette ligne, on utilise des petites boules fendues en plomb appelées « cendrées ». Les pêcheurs au coup ont une mauvaise habitude : ils serrent ces cendrées sur la ligne avec leurs dents. Bizarrement le pêcheur doit être plus souvent inquiet pour ses dents que par le fait d'être exposer à un risque sanitaire avec ce plomb qu'il manipule. Pourtant, la probabilité d'ingérer du plomb en fabriquant ces lignes doit être proche de 1. Il n'y a qu'à observer le bout de nos doigts recouverts d'une couche grise après avoir manipulé quelques cendrées. Ces doigts sont sans cesse alors mis à la bouche pour serrer les plombs. Les doigts dans la bouche... de vrais gamins !

L'enquête « Plomb-Habitat »a été initiée par Monsieur Emmanuel Briand au sein du Centre Scientifique et Technique du Bâtiment (CSTB). Je remercie Emmanuel d'avoir pensé à moi lorsque, à la constitution du dossier, mon nom est apparu sur la ligne « analyse statistique ». En dehors de cet aspect professionnel, Emmanuel a joué un autre rôle dans ma vie...

Je tiens donc à remercier Emmanuel Briand pour m'avoir permis de travailler sur une thématique intéressante. Cela a induit des recherches en statistiques que je n'aurais certainement pas eu l'occasion de faire par ailleurs. Je tiens à remercier le CSTB, et plus particulièrement Madame Séverine Kirchner d'avoir accepté que je m'inscrive en doctorat, bien qu'à l'époque, je fusse déjà ingénieur d'études au CSTB depuis trois ans.

Je souhaite remercier de plus Madame Lise Bellanger du laboratoire de mathématiques Jean Leray de Nantes, de m'avoir mis sur la route de ma future directrice de thèse puis de m'avoir encadré. Auparavant j'avais suivi les enseignements de Lise au début de mon cursus universitaire à la faculté des sciences et des techniques de Nantes, puis plus récemment dans le cadre du Master professionnel ingénierie mathématique en 2004-2005. Je remercie donc le Professeur Véronique Sébille, ma directrice, d'avoir accepté de m'encadrer au sein de l'équipe d'accueil 4275 « Biostatistique, Pharmacoépidémiologie et Mesures Subjectives en Santé » de la faculté de pharmacie de Nantes ; d'avoir accepté bien que quatre années furent planifiées pour réaliser la thèse et malgré le fait que je fusse basé en région parisienne.

Je remercie de plus les membres du comité de pilotage de l'enquête Plomb-Habitat pour leurs discussions enrichissantes, leurs relectures de mes articles et autres rapports scientifiques ainsi que pour leurs nombreux conseils. J'adresse un remerciement particulier à Monsieur Philippe Glorennec de l'EHESP. Je remercie particulièrement Madame Corinne Mandin, pour m'avoir encadré après Séverine au sein du CSTB, pour sa disponibilité et sa sympathie. Je tiens à dire merci à Yann le Strat de l'Institut de veille sanitaire (InVS) pour m'avoir initié aux sondages et pour m'avoir conseillé tout au long de la thèse. Je remercie de plus Alain le Tertre du même institut pour avoir réussi à rayer de mon vocabulaire l'adjectif « significatif » et pour m'avoir fait profiter de son expérience en statistiques appliquées.

Bien entendu je remercie les membres du jury d'avoir accepté d'évaluer ce travail au confluent de deux disciplines : la statistique appliquée à des données environnementales. Un remerciement particulier est destiné au Professeur Nathalie Seta de l'Université Paris Descartes pour avoir secouru un « doctorant en détresse ».

Je remercie enfin Marguerite, ma compagne, de m'avoir encouragé pendant ces quatre années estudiantines, ainsi que mes parents, Jean et Marie-Thérèse, pour m'avoir inculqué que seul le travail paye.

Introduction

Le plomb, métal lourd et reconnu toxique sans seuil, a des effets sanitaires graves chez l'Homme. Une baisse des capacités cognitives, un retard du développement psychomoteur ou des troubles du comportement par exemple sont connus pour des intoxications supérieures à 100 micro-grammes de plomb par litre de sang. Mais des effets délétères sur le développement neurologique en particulier sur le quotient intellectuel sont aujourd'hui constatés pour des intoxications plus faibles.
L'intoxication par le plomb se fait principalement par voie digestive et par voie pulmonaire dans une moindre mesure. Les enfants et les femmes enceintes constituent des populations très exposées. Par leur comportement spontané main-bouche, les enfants de moins de 6 ans sont particulièrement exposés via la poussière contaminée par le plomb à l'intérieur des logements. Leur absorption digestive est approximativement trois fois à cinq fois plus élevée que celle des adultes. Enfin leur système nerveux central, touché par l'intoxication, est en plein développement et l'intoxication peut commencer dans la vie intra-utérine. Cette exposition reste d'actualité en France à cause de l'accumulation passée et massive du plomb dans les sols, ce métal étant indestructible et peu mobile. En effet le plomb a été largement disséminé dans l'environnement par l'industrie et par l'utilisation des carburants plombés en France jusqu'à l'an 2000. La présence de plomb dans l'environnement résidentiel perdure à cause de la contamination de l'environnement extérieur, de l'utilisation passée de peintures à base de dérivés de plomb comme la céruse, le minium et certains siccatifs ainsi que de la présence de plomb dans certaines canalisations d'eau potable.

La loi du 9 août 2004 relative à la politique de santé publique a fixé dans ses objectifs de « réduire de 50% la prévalence des enfants ayant une plombémie $\geq 100\mu$g/L : passer de 2% en 1996 à 1% en 2008 » (objectif n° 18). L'indicateur retenu est « le nombre d'enfants de 1 à 6 ans ayant une plombémie $\geq 100\mu$g/L en population générale et dans les groupes à risque ».
Les connaissances sur l'imprégnation par le plomb de la population française sont principalement issues d'une enquête nationale menée par l'Inserm [1] et le RNSP [2] en 1995 et 1996 concernant les enfants de 1 à 6 ans. Cette enquête montrait que le taux de prévalence du saturnisme était de 2,1% dans cette classe d'âge, ce qui correspondait à 84 000 enfants sur l'ensemble du territoire.
Or les actions de dépistage du saturnisme infantile qui sont mises en œuvre depuis une quinzaine d'années n'ont permis d'identifier qu'une très faible partie des 84 000 enfants attendus : le nombre de cas de saturnisme de personnes mineures déclarés

1. Institut national de la santé et de la recherche médicale.
2. Réseau National de Santé Publique.

aux DDASS [3] chaque année sur l'ensemble de la France était de l'ordre de 500.
Par ailleurs, les expositions modérées sont un enjeux de santé publique grandissante, compte tenu de l'accumulation de preuves concernant les effets des faibles doses ($< 100\mu g/L$) en plomb chez l'enfant. Une actualisation des connaissances sur l'imprégnation des enfants et les facteurs de risque est donc indispensable.
C'est pourquoi la Direction Générale de la Santé (DGS) a demandé à l'InVS [4] de réaliser entre 2007 et 2009 une enquête de prévalence du saturnisme chez les enfants de 6 mois à 6 ans. L'enquête de prévalence du saturnisme offre une opportunité pour examiner, à travers une étude nationale complémentaire [5] au domicile de certains enfants, le lien entre la plombémie chez un enfant et les facteurs d'exposition de son environnement.

Afin de gérer le risque sanitaire, le ministère de l'emploi, de la cohésion sociale et du logement, et le ministère de la santé et des solidarités ont mis en place en 2006 le Constat de Risque d'Exposition au Plomb (CREP) [Ministère de la Santé et des Solidarités, 2006]. Le CREP a fait suite à l'État des Risques d'Accessibilité au Plomb (ERAP) mis en place en 1999. L'objectif premier du CREP est de mettre en évidence, dans un logement, le risque immédiat et futur pour ses occupants, d'une exposition au plomb liée aux revêtements contenant du plomb. Son protocole se base sur un mesurage surfacique du plomb et un repérage exhaustif des revêtements contenant du plomb (≥ 1 mg/cm^2). Il s'applique dans les logements construits avant 1949 vendus ou mis en location.
Par ailleurs à la suite d'un dépistage d'un cas de saturnisme (≥ 100 μg de plomb par litre de sang) chez une personne mineure, une enquête environnementale [6] est réalisée en partie au domicile de cette dernière. Cette enquête environnementale a pour but de déterminer l'origine de l'intoxication par le plomb de la personne intoxiquée. Pour cela des mesures et des prélèvements relatifs au plomb peuvent être réalisés au domicile de la personne intoxiquée.
À ce jour peu de données sont disponibles pour estimer l'exposition de la population nationale dans les logements. En effet par la réalisation d'un CREP ou d'une enquête environnementale à la suite d'un dépistage d'un cas de saturnisme, des données de contamination par le plomb dans les logements sont collectées. Néanmoins ces données n'ont jamais été centralisées dans une base de données nationale empêchant toute estimation de la contamination par le plomb en milieu résidentiel. Par ailleurs ces données ne concernant que des milieux résidentiels à risque, elles ne peuvent pas être utilisées pour estimer les niveaux en plomb dans les différents compartiments environnementaux de la population nationale de logements français.

La recherche des sources de contamination en plomb et de leur contribution à contaminer la poussière au sol ont été effectuées par différents travaux. Lanphear et Roghmann [Lanphear & Roghmann, 1997] ont montré que le plomb contenu dans le sol

3. Direction Départementale des Affaires Sanitaires et Sociales.
4. Institut de Veille Sanitaire.
5. Enquête Plomb-Habitat.
6. L'enquête environnementale a été mise en place par l'article 72 de la loi n° 2004-806 du 9 août 2004 relative à la politique de santé publique en modifiant l'article L1334-1 de la loi n° 98-657 du 29 juillet 1998 d'orientation relative à la lutte contre les exclusions.

extérieur contribuait à contaminer la poussière intérieure d'un logement. De plus les auteurs ont évalué la contribution du plomb des peintures à contaminer ces poussières intérieures. Dans leur analyse, seuls le plomb du sol extérieur et le plomb des peintures intérieures ont constitués des sources potentielles pouvant contaminer la poussière intérieure. L'étude a montré que le plomb des peintures contribuait à une plus grande contamination de la poussière intérieure que le plomb du sol extérieur. L'étude n'est cependant pas transposable au contexte français. En effet la mesure du plomb a été exprimée en plomb total alors qu'en France le dosage réglementaire en vigueur est le dosage du plomb acido-soluble. De plus l'étude n'a concerné que des logements américains d'une part, et uniquement des logements en zone urbaine d'autre part. Enfin leurs résultats ont été obtenus sur des données collectées il y a plus de 15 ans.
Des études plus récentes comme [Hunt et al., 2012] ont mis en évidence la contamination des poussières intérieures par la poussière ou le sol extérieur contaminé par le plomb. Cependant aucune étude n'a estimé dans quelle proportion le sol et la poussière extérieure pouvaient contribuer à contaminer la poussière intérieure déposée au sol.

Dixon *et al.* [Dixon et al., 2005b] ont montré que le plomb des poussières des parties communes d'immeubles migraient vers l'intérieur des logements et contaminaient leurs poussières. Cependant les auteurs n'ont pas indiqué dans quelle proportion les poussières des parties communes contribuaient à contaminer les poussières intérieures des logements. L'étude s'est basée sur des données collectées dans 14 états américains uniquement, dont les logements ont été investigués entre 1994 et 1996. Les résultats ont ainsi été obtenus sur des données américaines, géographiquement partielles et anciennes. Enfin le dosage du plomb a été réalisé en quantification totale.

Certaines études comme [Succop et al., 1998, Clark et al., 2004] ont été réalisées sur la base d'analyses prenant en compte plusieurs sources potentielles de contamination de la poussière intérieure. Succop *et al.* [Succop et al., 1998] ont effectué une méta-analyse à partir de 11 études afin d'étudier une modélisation des relations entre les niveaux en plomb de différents compartiments environnementaux résidentiels. Cependant les parties communes ne figurent pas parmi ces compartiments étudiés. De plus les données des 11 études sont relatives uniquement à des logements à risque (logements anciens) ou situés près de sites à risque. La démarche des auteurs ayant été focalisée sur leur modèle statistique, les relations entre les différents compartiments environnementaux ont été données en termes de pourcentages d'études (parmi les 11) ayant mis en évidence les dites relations. L'étude n'a donc pas estimé la contribution d'un compartiment environnemental à contaminer un autre, en particulier en ce qui concerne la contamination des poussières intérieures. De même Clark *et al.* [Clark et al., 2004] dans un contexte de post-réduction du risque plomb en milieu résidentiel, n'ont indiqué que partiellement la contribution de certaines sources à contaminer la poussière intérieure.

Concernant la variabilité spatiale de la contamination des poussières au sein d'un logement, le département américain de l'habitat et de l'urbanisme (HUD, *Department of Housing and Urban Development*) a recommandé de réaliser plusieurs prélève-

ments de poussière par lingette humide dans plusieurs pièces au sein d'un logement afin d'obtenir une mesure évaluant correctement la charge en plomb de la poussière du logement [U.S. HUD, 1995a]. De même l'Institut de Veille Sanitaire (InVS) a recommandé en France de réaliser au moins trois prélèvements de poussières par logement [Bretin, 2006]. Ces recommandations sont basées sur le fait que le prélèvement de poussières par lingette est jugé peu reproductible. Alors que *Wilson et al.* [Wilson et al., 2007] ont étudié les endroits où devaient se faire les prélèvements de poussière dans un logement afin de caractériser au mieux le risque pour l'enfant d'avoir une plombémie élevée ($\geq 100\,\mu g/L$), ils n'ont pas estimé la corrélation entre deux charges en plomb au sein d'un même logement. Aucune étude n'a semblé fournir cette corrélation permettant de confirmer si au sein d'un logement plusieurs prélèvements de poussière par lingette étaient nécessaires.

Ce rapide bilan fait apparaitre que les recherches menées depuis les années 1990 ont permis d'évaluer l'exposition au plomb dans l'environnement résidentiel. Ces évaluations ont permis de mettre en place des outils de réduction des expositions. De plus ces recherches ont mis en évidence l'intérêt particulier à apporter à la poussière contaminée déposée au sol à l'intérieur des logements. Ces poussières ont été incriminées comme étant le vecteur principal d'intoxication par le plomb chez l'enfant parmi les différents compartiments environnementaux résidentiels. Les travaux ont permis en outre d'identifier certaines sources à partir desquelles le plomb contenu dans ces poussières provenait. Ainsi la réduction des niveaux en plomb de ces sources est un moyen de réduire l'exposition au plomb en milieu résidentiel via la réduction préalable des niveaux en plomb dans les poussières intérieures.

Toutefois, les résultats apportés par ce domaine de recherche sont basés principalement sur des études américaines et reposent sur des données souvent anciennes. Dès lors il convient de s'assurer que ces résultats peuvent s'adapter au contexte actuel de l'habitat français. D'autre part la méthode de dosage du plomb en France n'est pas la même que la méthode utilisée aux États-Unis dans les principaux travaux publiés dans la littérature.
L'état de la contamination des logements par le plomb n'a jamais été réalisé en France à l'échelle nationale. Dès lors les niveaux en plomb dans les poussières intérieures des logements ne sont pas connus alors que les poussières intérieures constituent la voie principale d'exposition au plomb pour l'enfant. Les contributions respectives des sources potentielles pouvant contaminer les poussières intérieures ne sont donc pas connues en France ; elles ne sont d'ailleurs que partiellement connues outre Atlantique. De plus l'estimation des contributions conjointes de multiples sources potentielles pouvant contaminer la poussière intérieure ne semble jamais avoir été faite au sein d'une même étude. Par ailleurs, aux États-Unis comme en France, aucune étude n'a estimé la corrélation entre deux charges en plomb dans la poussière d'un même logement. Dès lors la recommandation de réaliser plusieurs prélèvements de poussière dans un logement afin d'évaluer sa contamination n'a jamais été confirmée.

Le travail présenté ci-après a pour ambition de répondre aux différentes interrogations dans le contexte français. Le travail consiste à estimer la prévalence des

logements possédant un risque plomb et à identifier les déterminants de la contamination des poussières intérieures déposées au sol. Ceci concerne les résidences principales abritant au moins un enfant âgé de 6 mois à 6 ans en France métropolitaine.

Après une revue de la littérature sur l'exposition au plomb en milieu résidentiel, le travail de recherche s'est basé sur les données collectées lors de l'enquête Plomb-Habitat réalisée en 2008 et 2009 dans 484 résidences principales françaises où 1834 pièces au total ont été investiguées. Chaque logement enquêté s'est vu affecté un poids de sondage[7] par la stratégie d'échantillonnage mise en place.

La démarche a consisté dans un premier temps à valider les données collectées. Pour cela des outils de *« data-management »* ont été utilisés pour obtenir des données valides, cohérentes entre elles, et pour travailler *in fine* sur une base de données avec le moins de données manquantes possible.

Dans un second temps un état de la contamination par le plomb à l'échelle nationale dans les différents médias environnementaux des résidences principales de France métropolitaine a été établi. Pour cela les outils de la théorie de l'échantillonnage et de l'estimation en population finie ont été utilisés pour estimer les niveaux en plomb. Afin d'améliorer les estimations, un redressement des poids de sondage des logements par post-stratification a été réalisé. En vue d'établir un état de la contamination par le plomb dans les logements, les niveaux en plomb dans l'eau du robinet, dans la poussière intérieure et celle des parties communes, dans la poussière et le sol des aires de jeu extérieures ainsi que dans les revêtements intérieurs ont été estimés. Un article scientifique décrivant la contamination par le plomb dans les logements français occupés par des enfants a été publié en 2012 [Lucas et al., 2012].

Dans un troisième temps, on s'est focalisé sur l'un des compartiments décrits précédemment, la poussière intérieure, dans le but d'étudier la provenance du plomb qu'elle contient. Puisque plusieurs pièces ont été investiguées au sein d'un même logement de l'enquête Plomb-Habitat, une modélisation multi-niveaux à 2 niveaux (pièces comme niveau 1 ; logements comme niveau 2) a été développée. Ceci afin d'estimer d'une part la contribution des sources potentielles à contaminer la poussière intérieure et d'autre part d'estimer la corrélation entre 2 charges en plomb à l'intérieur d'un même logement. Toutes les sources potentielles identifiées dans la littérature et disponibles à travers les informations collectées dans l'enquête Plomb-Habitat ont été sélectionnées pour figurer dans le modèle. Puisque les données étaient issues d'un plan de sondage, les différentes pondérations possibles pour le niveau 2 du modèle ont été testées afin de mettre en évidence l'impact des différents poids sur les estimations des effets des sources en plomb obtenues par pseudo-vraisemblance. Un modèle à 2 niveaux sans pondération au niveau 2 a été supposé à ce stade comme utile afin d'estimer la contribution de chaque source. Enfin, pour ne pas introduire de biais dans les estimations des effets des sources, les données manquantes ont été

7. Un logement avec un poids de 1000 représente 1000 logements dans la population.

imputées par imputation multiple avant de procéder à l'estimation.

Dans un quatrième temps une étude de simulation Monte Carlo a été réalisée afin de valider l'utilisation à l'étape précédente d'un modèle à 2 niveaux sans pondération. Elle a consisté à comparer les estimations des coefficients de régression obtenues selon les différentes pondérations possibles avec la valeur des coefficients fixée dans la population de logements simulée.

De l'exposition au plomb et de la présence du plomb en milieu résidentiel

Pour étudier le plomb dans un cadre particulier il est important de connaître les composés dans lesquels on peut le retrouver dans l'environnement, les usages du plomb par l'Homme et les effets du plomb sur la santé.

1 Les problèmes sanitaires liés à l'exposition au plomb

Le rapport de l'Inserm de 1999 [INSERM, 1999] constitue une expertise collective réalisée par des spécialistes de plusieurs disciplines. L'expertise a été réalisée sur la base d'une analyse des travaux scientifiques existant relatifs aux effets du plomb sur la santé des populations et aux contextes d'exposition. La première partie de ce rapport développe ainsi les thèmes suivants :

- Toxicocinétique et dosage du plomb
- Toxicité cellulaire du plomb
- Neurotoxicité cellulaire du plomb
- Effets du plomb sur le développement cérébral
- Effets du plomb sur les fonctions cognitives chez l'enfant
- Effets du plomb sur le système nerveux de l'adulte
- Effets du plomb sur l'appareil reproducteur
- Effets du plomb sur différents organes
- Effet cancérogène du plomb
- Diagnostic et traitement chez l'enfant

Le rapport Inserm contient un diagramme résumant les effets du plomb chez l'enfant et l'adulte, diagramme visible en figure 1. Ce diagramme date de 1990 et ést issu de l'agence américain ATSDR (*Agency for Toxic Substances & Disease Registry*). Il montre tout un panel de pathologies liées au degré d'intoxication chez l'adulte ainsi que chez l'enfant, à partir du taux minimum en plomb dans le sang où l'effet peut être observé. L'enfant est plus vulnérable à l'intoxication par le plomb pour différente raisons : son absorption digestive est trois [8] fois plus élevée que celle des

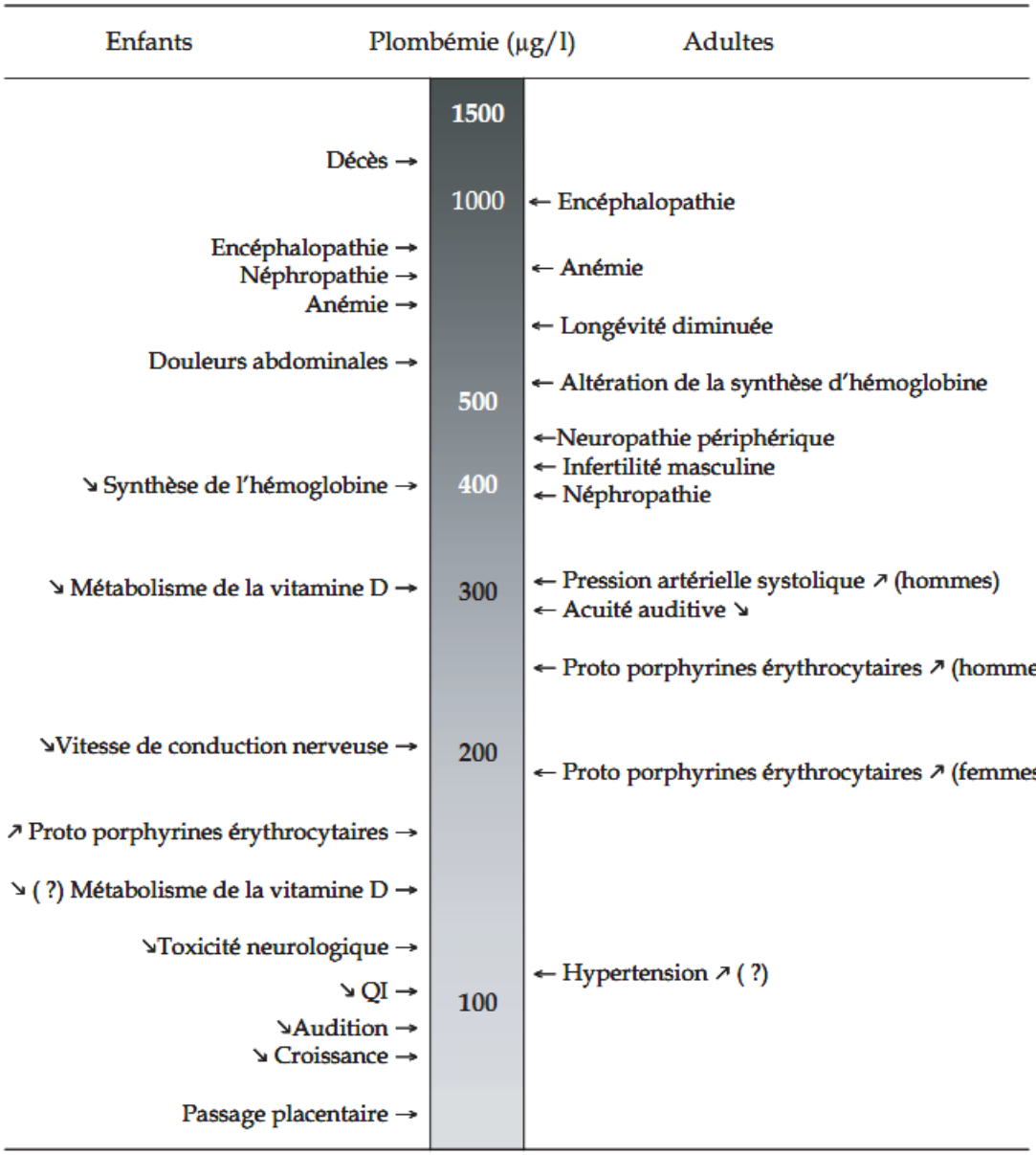
8. L'avis récent (du 14 juin 2013) du Haut Conseil de la Santé Publique (HCSP) sur l'analyse et l'évaluation de l'efficacité des actions engagées pour respecter la future limite de qualité de 10μg/L de plomb dans l'eau du robinet, indique cinq fois plus.

adultes ; il est en pleine croissance et son système nerveux central est en plein développement induisant que, pour une même imprégnation, les effets toxiques du plomb sont plus importants et plus sévères chez l'enfant que chez l'adulte ; le plomb passe la barrière placentaire et l'intoxication peut commencer dès la vie intra-utérine. Ainsi chez l'enfant un ralentissement de la croissance et une perte de l'audition peut être imputables à des taux de plomb dans le sang (plombémies) inférieurs à 100μg/L d'après les connaissances scientifiques de l'époque. Une perte de quotient intellectuel (QI) est associée à une plombémie au-dessus de 100μg/L. Des pathologies plus lourdes sont associées à des plombémies plus élevées : diminution de la synthèse de l'hémoglobine à partir de 400 micro-grammes de plomb par litre de sang ou, par exemple, une anémie pour des plombémies entre 500 et 1000μg/L. Dans une fourchette de 1000 à 1500μg/L le décès survient. L'apparition de certaines pathologies étaient incertaines : chez l'enfant une diminution du métabolisme de la vitamine D peut se situer autour de 150 ou bien de 300 micro-grammes de plomb par litre de sang ; chez l'adulte, une aggravation de l'hypertension est possiblement associée à des plombémies légèrement supérieures à 100μg/L.

En 2008 à la demande de la Direction Générale de la Santé (DGS), l'Inserm a publié les travaux réalisés au sein d'un groupe de travail avec l'InVS (Institut de Veille Sanitaire). Il s'agissait d'un second ouvrage prolongeant le premier cité précédemment, réalisé dans le cadre de la procédure d'expertise opérationnelle concernant les stratégies de dépistage du saturnisme[9] chez l'enfant [INSERM, 2008]. De plus cet ouvrage complète les connaissances sur l'intoxication au plomb et de ses effets sur la santé, en particulier en ce qui concerne les plombémies inférieures à 100μg/L. Cette complétion s'est faite à partir de nombreuses publications publiées à partir de l'an 2000, basées sur des études de cohortes, des études transversales ou longitudinales, et des méta-analyses. Ainsi des études publiées entre 2003 et 2005 indiquent un impact néfaste sur le QI chez des enfants n'ayant jamais eu une plombémie excédant 100μg/L. Alors qu'une plombémie de 100μg/L était considérée comme minimale pour observer des effets délétères sur le QI de l'enfant (figure 1), les nouvelles connaissances post-2000 ont remis en question ce seuil de plombémie égal à 100μg/L. Outre le QI, il a été montré une relation inverse entre plombémie et capacité à l'apprentissage par association pour des plombémies, là encore, inférieures à 100μg/L. Une relation inverse entre plombémie et performances cognitives a de même été montrée mais cependant sans pouvoir déterminer un seuil. De plus, plusieurs études ont rapporté que, même une exposition au plomb faible pendant l'enfance était associée à des troubles neuromoteurs, à des effets cognitifs néfastes, à un impact négatif sur la maturation de la balance posturale. Toutes les études citées par l'ouvrage Inserm publié en 2008 mettent donc en évidence des impacts délétères en lien avec des expositions au plomb en-deçà du seuil de 100μg/L, et ceci sans effet de seuil précis. Néanmoins certains effets restent difficiles à évaluer, comme l'effet à long terme d'une intoxication des jeunes filles pendant leur enfance sur le développement de leurs propres enfants par exemple.

9. Au 1er janvier 2013 le saturnisme est (encore) défini à partir d'un seuil de 100 microgrammes de plomb par litre de sang.

FIGURE 1 – Effets du plomb inorganique connus en 1990 chez l'enfant et l'adulte. Cités dans l'expertise Inserm de 1999.



2 La présence du plomb dans les logements et dans leur environnement

Le plomb a été utilisé par l'Homme depuis des millénaires et son l'utilisation a augmenté considérablement à partir de la révolution industrielle, entrainant une libération et une accumulation de ce métal dans l'environnement. Le plomb a été utilisé dans la peinture à partir du XIX^{e} siècle. Il a aussi été utilisé pour les canalisations d'eau. La dispersion du plomb causée par les alkyls de plomb utilisés dans l'essence dans la première moitié du XX^{e} siècle, a été un phénomène planétaire. Le plomb s'est donc accumulé dans les sols ; étant indestructible et peu mobile, ce métal a donc demeuré dans les couches superficielles où il reste accessible par l'Homme,

en particulier les enfants, ce qui représente une menace permanente pour la santé des populations.

2.1 Le plomb dans la peinture

J.L. Gibson en 1904 avança que la peinture au plomb était la cause de l'intoxication par le plomb de quatre enfants du Queensland en Australie [Gibson, 2005], atteints de troubles oculaires. Par ses découvertes, Gibson a marqué le début de l'ère où on allait se rendre compte de la dangerosité du plomb utilisé comme pigment dans les peintures, en particulier en France 80 ans plus tard... Gibson a fait une série d'observations qui sont toujours d'actualité, en particulier concernant le rôle joué par la pulvérulence de la peinture à base de plomb dans la contamination des poussières et par la même de la dangerosité de ces poussières contaminées. Ses découvertes qui ont été ensuite validées par de nombreux scientifiques avec des méthodes analytiques plus fines et par l'épidémiologie.

Le plomb dans la peinture est souvent synonyme de céruse dont on s'intéressera plus particulièrement. L'ouvrage du conservatoire national des arts et métiers (Cnam) dénommé « La céruse : usages et effets, Xe - XXe siècles » [Guillerme et al., 2003] fournit de précieuses informations relatives à ce composé du plomb, autrement appelé sous différents noms. Il est appelé hydrocarbonate de plomb, carbonate de plomb basique, hydrocerussite, blanc de plomb ou encore blanc de Saturne. La céruse a pour formule chimique $2(PbCO_3).Pb(OH)_2$ autrement notée $C_2H_2O_8Pb_3$ et a pour numéro CAS (*Chemical Abstracts Service*) 1319-46-6. Le chapitre écrit par L. Lestel, disponible dans un article à part [Lestel, 2002] est particulièrement intéressant lorsque l'on souhaite prendre connaissance de repères historiques quant à la production de céruse en France. Il fournit en annexe 1 de l'ouvrage du Cnam, une liste des producteurs français de céruse de 1800 à 1950, fort utile lorsque l'on veut rapprocher production de céruse et textes règlementaires légiférant sur son usage.
Une chronologie de l'usage de la céruse sous ces différentes formes est disponible via un article à part [Guillerme, 2002] ; cet article est l'introduction de l'ouvrage du Cnam. On y apprend l'origine du mot « saturnisme », origine étonnamment peu souvent rappelée en général dans les textes traitant du saturnisme et écrits en français : « La céruse, connue en Occident dans l'Antiquité, est réputée : aux temps des fastes de l'Empire romain, elle couvre de sa blancheur le visage et les mains des millions d'esclaves libérés pour les sept jours des Saturnales. » De manière annexe sur ce sujet on peut citer : « Dans l'alchimie, on a comparé le plomb à Saturne, non-seulement parce qu'on a cru ce métal comme le plus vieux et le père des autres ; mais encore parce qu'on le regardait comme très-froid ; parce qu'on lui attribuait la propriété d'absorber et de détruire en apparence presque tous les métaux, comme la fable disait que Saturne, le père des dieux, avait mangé ses enfants. »[10] La demande de céruse est très soutenue dès le milieu du XVIII^{e} siècle. Auparavant produit artisanal, importé, la céruse devient alors produit industriel. Grâce à ce pigment blanc utilisé dans la peinture mélangé à l'huile, les villes se parent de murs blancs immaculés.

10. Fourcroy A. F. Système des connaissances chimiques, et de leurs applications aux phénomènes de la nature et de l'art, Tome VI. Sect. VI. Art. 17. Du plomb, Paris, Brumaire an IX.

L'usage de la céruse se développe surtout au XIX^{e} siècle. Outre sa couleur blanche, la peinture à la céruse permet de résister à l'humidité salpêtreuse des murs et résiste mieux que le blanc de chaux qui s'écaille et se pulvérise. Plus les appartements étaient pauvres et sombres plus ils étaient cérusés. La céruse s'accumule d'abord dans les rez-de-chaussée et les basses cours pour ensuite monter dans les étages et dans les pièces de réception des appartements bourgeois. On s'y éclaire au gaz de houille qui noircit les murs et oblige à repeindre en blanc tous les ans. La céruse a donc concerné principalement le logement urbain.

Outre la céruse, d'autres composés de plomb ont été utilisés dans la peinture. En particulier le minium ou oxyde de plomb (Pb_3O_4) est un autre dérivé du plomb très largement utilisé jusqu'au milieu des années 1990. Contrairement à la céruse, il a été utilisé principalement à l'extérieur des bâtiments comme anticorrosif. Il a pour numéro CAS 1314-41-6. En outre, d'autres dérivés du plomb comme les siccatifs (permettant d'accélérer le séchage de la peinture) ont été largement employés ; ce sont des sels de plomb d'acides organiques. Contrairement à la céruse et au minium, les siccatifs sont beaucoup moins acido-solubles et donc moins bien absorbés par le tube digestif. De plus, ils ont été utilisés généralement à des concentrations plus faibles que la céruse et le minium pour qui, leur concentration représentait de 60 à 80 % de la peinture sèche.

Cet usage de la céruse et de certains dérivés du plomb a été réglementé en France. Une recherche bibliographique dédiée à la réglementation de l'usage dans la peinture de la céruse et des dérivés du plomb a été réalisée. En effet les analyses statistiques réalisées dans ce travail sont souvent basées sur la date de construction des logements, cette date étant étroitement liée à la réglementation relative à l'usage du plomb dans la peinture. La date de 1949 est communément considérée comme la date à partir de laquelle la céruse n'a plus été employée dans la peinture en France. Or les mesures prises à cette date l'avaient déjà été dès 1926 : 1949 ne peut donc pas être considérée véritablement comme date charnière. Cette réglementation ne concernait que l'usage de la céruse par les professionnels et n'a jamais concerné les particuliers ; l'interdiction de la vente de peinture au plomb ne s'est faite qu'en 1993. Un article recensant ce travail a été publié dans la revue Environnement Risques & Santé [Lucas, 2011]. L'article est disponible en annexe 12.

2.2 Le plomb dans l'eau du robinet

L'article de J. Baron [Baron, 1997] permet de prendre connaissance des déterminants du plomb dans l'eau du robinet.
L'origine du plomb dans l'eau est due aux matériaux en contact avec l'eau pendant son transport, en particulier le plomb des branchements entre le réseau public et le réseau des immeubles, ainsi que les canalisations à l'intérieur des immeubles. Les eaux de distribution ne contiennent qu'au plus quelques micro-grammes de plomb par litre, qu'elles soient d'origine souterraine ou superficielle. D'autres matériaux sont susceptibles de relarguer du plomb dans l'eau : l'acier galvanisé contenant 1% de plomb, le laiton et le bronze pouvant avoir des teneurs supérieures à 5% de plomb,

les soudures dites à l'étain contenant environ 60% de plomb, certains PVC d'origine étrangère stabilisés avec des sels de plomb.
Outre la source du plomb, d'autres paramètres jouent un rôle important sur la solubilité du plomb dans l'eau. En particulier les caractéristiques de l'eau elle-même : température, pH, alcalinité et sa teneur en phosphates. Des modèles thermodynamiques permettent d'estimer la solubilité théorique du plomb. À 25 °C, elle varie de plusieurs milligrammes de plomb par litre d'eau lorsque l'alcalinité et le pH sont faibles (titre alcalimétrique complet < 3 et pH $< 6{,}5$) jusqu'à moins de $100\mu g/L$ pour une alcalinité et un pH supérieurs (titre alcalimétrique entre 5 et 15, et pH $>$ 8).
Enfin les facteurs liés au réseau d'eau et son exploitation ont un rôle sur la solubilité du plomb. En particulier le temps de stagnation joue un rôle essentiel. Il semble que la concentration en plomb maximale soit atteinte entre 5 et 6 heures de stagnation pour des conduites en plomb de 10mm de diamètre et de plusieurs dizaines d'heures pour des conduites de 50mm de diamètre. Ce temps de stagnation dépend de la fréquence des soutirages par l'usager et leur importance. D'autres facteurs annexes peuvent impacter la solubilité du plomb : alternance des matériaux métalliques, vibrations, utilisation des conduites pour mise à la terre d'appareils électriques par exemple.

L'utilisation de canalisations en plomb [11] pour les installations nouvelles est interdite par le décret n° 95-363 du 5 avril 1995. Dans les faits il est communément admis que ces canalisations ont de moins en moins été utilisées à partir du milieu du XX^{e} siècle.

En 1994, la prévalence des branchements en plomb en France était estimée à 47%, allant de 20% en Rhône-Alpes jusqu'à 74% en région parisienne, à partir d'une étude de la Lyonnaise des Eaux-Dumez [Duguet et al., 1994]. Plus récemment, la Fédération Professionnelle des Entreprises de l'Eau indiquait que fin 2008, restaient encore plus de 1 200 000 branchements en plomb dans les services exploités par les opérateurs privés en France [12].

2.3 Le plomb atmosphérique

Les niveaux en plomb dans l'atmosphère aux États-Unis étaient étroitement liés à la consommation d'essence à base de plomb comme le souligne J. O. Nriagu [Nriagu, 1990]. Cet article permet de prendre connaissance de l'histoire de l'apparition du plomb dans l'essence grâce aux propriétés anti-détonnantes du plomb tétraéthyle ($C_8H_{20}Pb$). Le premier gallon de cette essence a été vendu le 2 février 1923 à Dayton dans l'Ohio aux États-Unis. Les industriels présentaient cette essence comme un véritable « cadeau de Dieu » car le plomb tétraéthyle a permis le développement

11. D'où le nom de « plomberie » du latin *plumbum*.
12. Rapport BIPE 2010 : les services publics d'eau et d'assainissement en France, données économiques, sociales et environnementales. Téléchargeable à l'adresse http://www.fp2e.org/userfiles/files/publication/etudes/12684096832_Rapport_BIPE_FP2E_2010.pdf au 07/06/2013.

du moteur à essence, essentiel pour la civilisation américaine selon eux. À travers le monde, la production d'essence au plomb a dépassé les 720×10^9 litres par an, ce qui en faisait en 1990 l'un des plus grands volumes de produits chimiques organiques jamais produits. Cependant de premières inquiétudes concernant l'insécurité sanitaire des additifs plombés sont apparues dès octobre 1922 après l'annonce des propriétés anti-détonantes du plomb tétraéthyle. Après d'intenses débats de plusieurs dizaines d'années, l'U.S. EPA (*Environmental Protection Agency*) imposa qu'au 1er juillet 1975 les plus grands revendeurs d'essence vendent au moins une catégorie d'essence sans plomb. Il s'agissait là de protéger en priorité les convertisseurs catalytiques, intolérants au plomb, qui devaient être installés dans les automobiles à partir de 1975. La protection de l'environnement et de la santé n'étaient donc pas les causes de cette réglementation. À partir de 1975 la consommation en essence plombée a chuté brutalement tout comme les niveaux de plomb atmosphérique aux États-Unis. Dans la communauté européenne (CEE) la concentration en plomb dans l'essence ne devait pas dépasser 0,4 g/L en 1984 (une essence était considérée sans plomb en deçà d'une concentration de 0,026 g/L). À partir du 1er janvier 1986 les pays de la CEE pouvait vendre de l'essence sans plomb de leur propre initiative. Au 1er juillet 1989 on devait pouvoir trouver de l'essence sans plomb dans tous les pays membres. À cette même date, la concentration maximale en plomb autorisée dans l'essence qui avait été planifiée était de 0,15 g/L.

L'essence au plomb a été définitivement interdite aux États-Unis par le *Clean Air Act* à partir du 1er janvier 1996. En Europe la directive 98/70/CE a interdit la commercialisation de l'essence à base de plomb à partir du 1er janvier 2000. Les données du Centre Technique Interprofessionnel d'Étude de la Pollution Atmosphérique (CITEPA) permettent de voir qu'entre 1990 et 2008 la concentration en plomb dans l'air a chuté en France de 98% [CITEPA, 2010]. Les émissions étaient de 4257 tonnes en France en 1990, et sont passées à 95 tonnes en 2008. De 1990 à 1999, le secteur prédominant d'émissions en plomb était le transport routier avec 91,2 % des émissions en plomb dans l'air, contre 68,7 % en 1999. À partir de l'année 2002 la part du transport routier devient très faible : 3,1 % en 2000 dues essentiellement aux traces subsistant dans les cuves entre le passage de l'essence au plomb à l'essence sans plomb. La contribution du secteur routier est devenue nulle par la suite. En 2008, l'industrie manufacturière était le secteur le plus émetteur en plomb dans l'air (on pourra se reporter à la section suivante sur ce sujet).

2.4 Le plomb dans le sol

Le plomb du sol peut provenir de deux sources selon le rapport du BRGM (Bureau de Recherches Géologiques et Minières) élaboré à la demande du Ministère de l'Écologie et du Développement Durable : sources naturelles et sources anthropiques [13] [Laperche et al., 2004].

Le plomb est un élément présent naturellement dans l'environnement. Des gisements de minerais de plomb ont existé et existent encore à travers le monde.

13. i.e. dues à l'homme.

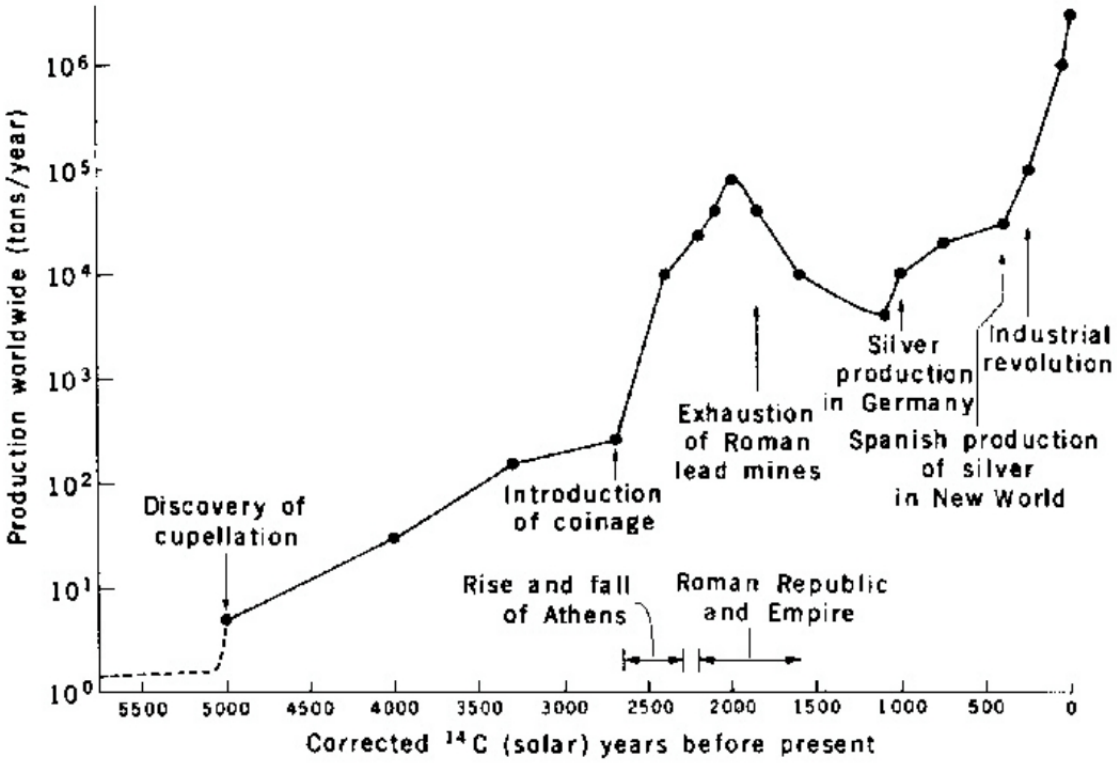
Les régions d'exploitation les plus importantes se situent en Australie, en Chine, aux États-Unis, en Russie, au Canada, en Afrique du Sud, au Mexique, au Pérou, dans l'ancienne Yougoslavie, en Bulgarie et au Maroc. En 1990, 70% des réserves mondiales se répartissaient entre l'Australie et l'Amérique du Nord. En France la production a été de 1800 kilotonnes entre 1880 et 1991. La dernière exploitation fermée en 1991 en France se situait dans le département du Gard (30).

Le plomb est le 36^{e} élément constituant l'écorce terrestre en termes d'importance. L'écorce terrestre contient 14,8 mg/kg de plomb. Le plomb entre dans la composition d'environ 240 minéraux naturels identifiés. Les teneurs en plomb dans les roches vont naturellement de quelques milligrammes à quelques dizaines de milligrammes par kilogramme (fond géochimique). Dans les sols agricoles la teneur en plomb varie de 2 à 200 mg/kg. Certains auteurs s'accordent à dire que les sols non contaminés contiendraient de 10 à 30 mg/kg en plomb. Selon d'autres, des teneurs en plomb au delà de 110 mg/kg ne devraient pas se rencontrer dans les sols naturels.

L'utilisation du plomb par l'Homme depuis des millénaires est source de dispersion du plomb dans l'environnement. La figure 2 montre la production mondiale du plomb depuis les cinq derniers millénaires. La métallurgie du plomb a débuté il y a environ 3000 ans. Sous Rome, on a utilisé ce métal pour les conduites d'eau, pour les poids-étalons, pour les tablettes d'écriture et pour la monnaie. L'oxyde de plomb était employé comme pigment. La révolution industrielle a propulsé la production de plomb à son apogée.

Ainsi comme élément de dissémination du plomb dans l'environnement et en particulier le sol, les différentes activités métallurgiques peuvent être citées : la métallurgie de première fusion consistant à faire subir au minerai plusieurs traitements afin d'extraire le plomb et les autres métaux ; la métallurgie de deuxième fusion est le recyclage, elle permet d'obtenir du métal à partir de la récupération des déchets (par exemple le recyclage des batteries). La production métallurgique de plomb a été de 163 000 tonnes en France en 1996 dont 72 000 tonnes pour le seul site de Métaleurop dans le département du Nord (59).
Concernant les productions industrielles induisant une libération directe ou indirecte du plomb dans l'environnement notamment le sol, il peut être cité : l'essence en plomb, le recouvrement de toit où l'étanchéité est assurée par des feuilles ou des bandes ou des tubes en plomb, les accumulateurs au plomb, les munitions de chasse et lests de pêche, la peinture au plomb. En 2008 l'industrie manufacturière (métallurgie des métaux non ferreux, des minéraux non métalliques, des matériaux de construction et la métallurgie des métaux ferreux) était la source principale émettrice avec 70 tonnes en plomb dans l'air en France métropolitaine (pour rappel, 4257 tonnes d'émissions en France en 1990) soit 73,4% des émissions totales ; l'industrie des métaux ferreux étant la plus responsable avec 55,6% des émissions de ce secteur. Le secteur résidentiel était responsable de 14% des émissions en 2008 et le transport aérien français de 5,9%.

FIGURE 2 – Production mondiale du plomb depuis 5000 ans. Citée dans le rapport du BRGM de 2004.



2.5 Le plomb dans la poussière en milieu résidentiel

La poussière au sol en milieu résidentiel ne contient du plomb par le biais de sa contamination par les autres compartiments environnementaux précédemment décrits. De ce fait, pour la poussière on se reportera à la section 4 de cette partie traitant du processus de contamination par le plomb entre les compartiments environnementaux.

3 La mesure des niveaux du plomb dans les compartiments environnementaux résidentiels

3.1 Dans la peinture

En France la mesure le plus répandue des niveaux en plomb en milieu résidentiel est réalisée par le Constat de Risque d'Exposition au Plomb (CREP) défini par l'arrêté du 25 avril 2006. Le CREP est un diagnostic technique et « doit être produit lors de la vente de tout ou partie d'un immeuble à usage d'habitation construit avant le 1er janvier 1949 ; depuis août 2008, il doit aussi être annexé à tout contrat de location d'un immeuble affecté en tout ou en partie à l'habitation, construit avant le 1er janvier 1949. De plus, ce constat doit être réalisé précédemment à tous travaux portant sur les parties à usage commun d'un immeuble affecté en tout

ou partie à l'habitation, construit avant le 1er janvier 1949, lorsque ces travaux sont susceptibles de provoquer une altération substantielle des revêtements » [Lucas, 2011]. La mesure du plomb est une mesure surfacique exprimée en milligrammes par centimètre carré (mg/cm^2) réalisée par un appareillage portatif à fluorescence X (« *XRF* »). La mesure est non destructrice. La mesure par fluoresence X est normalisée selon la norme française NF-X46-030 [AFNOR, 2008c].

Protocole CREP

Le protocole CREP consiste à réaliser des mesures du plomb surfacique des revêtements d'un logement, et ce de manière exhaustive. Le diagnostiqueur liste toutes les unités de diagnostic (UD) de chaque pièce ; une UD est un élément du bâti homogène en termes d'historique et de revêtements (e.g. deux murs d'une même pièce peints avec la même peinture et au même moment peuvent être considérés comme une même UD, alors que deux murs peints au même moment mais avec des couleurs différentes ne le seront pas). Il procède à une mesure du plomb et consigne les résultats dans un tableau, en indiquant la zone de la mesure (pour pouvoir identifier facilement *a posteriori* la zone plombée le cas échéant), le substrat (plâtre, bois, etc.), le revêtement apparent (peinture, papier peint, etc.) et l'état de dégradation du revêtement de l'UD s'il contient du plomb (figure 3).

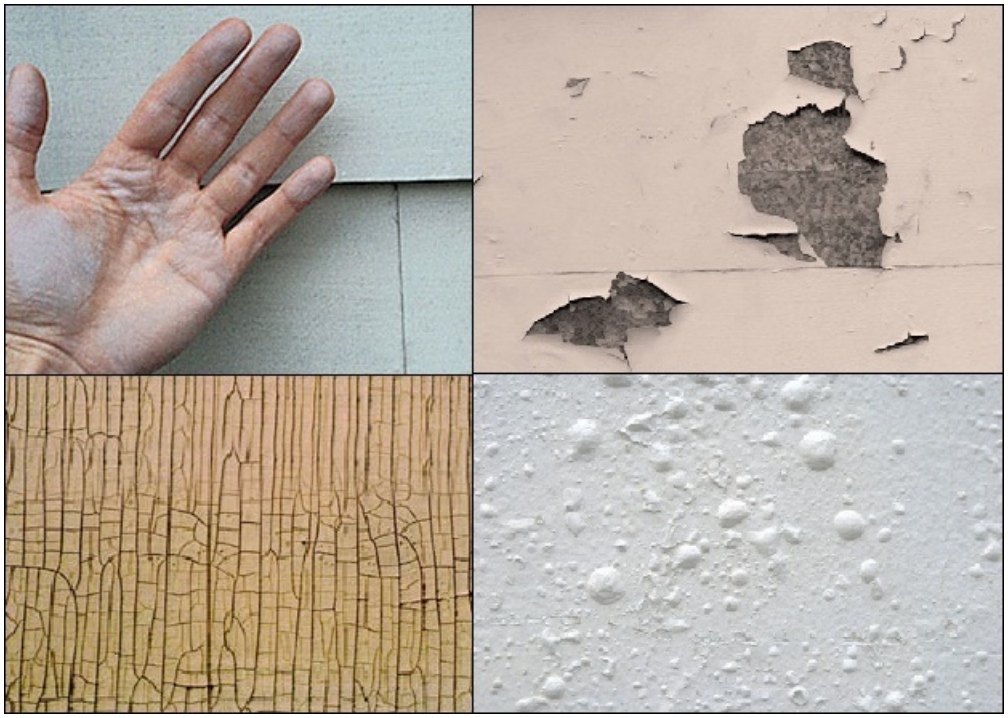
FIGURE 3 – Exemple de tableau de résultats de mesures XRF fourni par un rapport CREP. Extrait de l'annexe 1 de l'arrêté du 25 avril 2006 relatif au constat de risque d'exposition au plomb.

N°	LOCAL	ZONE	UNITÉ DE DIAGNOSTIC	SUBSTRAT	REVÊTEMENT apparent	LOCALISATION de la mesure (facultatif)	MESURE (mg/cm²)	NATURE de la dégradation	CLASSEMENT	OBSERVATIONS
4	Salon	A	Bâti porte	Bois	Peinture		0,05		0	
5							0,03			
6	Salon	A	Imposte	Plâtre	Papier peint		0,16		0	
7							0,12			
8	Salon	B	Plinthes	Bois	Peinture		0,13		0	
9		D					0,05			
14	Salon	C	Allège	Plâtre	Papier peint		0,06	D	3	
15							19,30			
16	Salon	C	Ouvrant fenêtre	Bois	Peinture		12,61	EU	2	
17	Salon	C	Bâti fenêtre	Bois	Peinture		13,10	EU	2	

Un revêtement contient du plomb dès lors que la mesure est supérieure ou égale à 1 mg/cm^2. L'état de dégradation a trois modalités : non dégradé ou non visible, état d'usage, dégradé. L'état d'usage correspond à des traces de chocs ou des micro-fissures considérés comme peu susceptibles de disséminer du plomb ; l'état dégradé,

comme par exemple de l'écaillage, de la pulvérulence, ou des lézardes, est considéré comme pouvant disséminer du plomb contenu dans le revêtement (figure 4).

FIGURE 4 – Exemples de dégradation de peinture. De gauche à droite et de haut en bas : pulvérulence, écaillage, lézardes, cloquage.



En fonction de la charge en plomb et de l'état de dégradation de son revêtement, une UD est classée selon la table 1. Le logement est alors classé de la manière suivante :

- Si toutes les UD sont classées en 0 alors le logement est classé 0.
- S'il n'y a que des UD classées au plus en 1, le logement est classé 1.
- S'il n'y a que des UD classées au plus en 2, le logement est classé 2.
- S'il y a au moins une UD classée en 3, le logement est classé 3.

En conclusion, le logement est dit sans revêtement à base de plomb si le logement est classé 0. S'il est classé 1 ou 2, l'auteur du constat rappelle au propriétaire l'intérêt de veiller à l'entretien des revêtements les recouvrant, afin d'éviter leur dégradation future. S'il est classé 3, l'auteur du constat rappelle au propriétaire l'obligation d'effectuer les travaux appropriés pour supprimer l'exposition au plomb et rapelle l'obligation de communiquer le constat aux occupants de l'immeuble ou de la partie d'immeuble concernée et à toute personne physique ou morale appelée à effectuer des travaux dans cet immeuble ou partie d'immeuble. Cette communication consiste à transmettre une copie complète du constat, annexes comprises.

La mesure du plomb dans la peinture peut aussi être faite par dosage. Cette méthode rendue utilisable par le protocole du CREP est très peu appliquée en

pratique [14]. Des écailles de peinture sont prélevées pour être analysées en laboratoire afin d'exprimer une concentration massique en milligrammes par gramme (mg/g). La norme NF-X46-031 [AFNOR, 2008a] définit l'analyse chimique. En France c'est la fraction acido-soluble du plomb qui est dosée (par opposition au plomb total utilisé ailleurs en particulier aux États-Unis). Elle consiste à simuler la solubilisation du plomb dans l'estomac et ainsi à s'approcher de la quantité de plomb pouvant passer dans le sang.

Table 1 – Classement des unités de diagnostic.

Charge en plomb	Type de dégradation	Classement
< 1 mg/cm^2	-	0
≥ 1 mg/cm^2	Non dégradé ou non visible	1
	État d'usage	2
	Dégradé	3

3.2 Dans la poussière

La mesure du plomb dans les poussières se fait principalement avec un prélèvement par lingette humide avec laquelle la surface est essuyée. Cette méthode a été comparée à des méthodes par prélèvement par aspirateur [Lanphear et al., 1995, Sterling et al., 1999]. Elle a été finalement recommandée par ces deux études pour des prélèvements à grande échelle. C'est en effet une méthode facile à mettre en œuvre, elle produit une mesure en plomb bien corrélée avec la plombémie et est moins onéreuse qu'un prélèvement par aspirateur. Ce dernier prélèvement a cependant l'avantage notable de fournir une concentration en micro-grammes de plomb par gramme de poussière (μg/g) et donc une mesure non liée à l'empoussièrement contrairement à la mesure fournie par le prélèvement par lingette exprimée en microgrammes par mètre carré (μg/m^2). De ce fait, la méthode par lingette est jugée peu reproductible par les experts et nécessite de réaliser plusieurs prélèvements pour aboutir à une bonne estimation de la contamination du local. Le prélèvement par lingette est normalisé selon la norme française NF-X46-032 [AFNOR, 2008b].

3.3 Dans le sol

La mesure du plomb dans les sols se fait généralement à partir de carottage. Il n' y a pas de norme officielle définissant ce prélèvement. Il existe une norme concernant les prélèvements en vue d'une interprétation agronomique d'après le Guide d'investigation environnementale des cas de saturnisme de l'InVS [Bretin, 2006] reprenant le guide du BRGM [Laperche & Mossmann, 2004]. Dans le cadre d'une enquête après la signalisation d'un cas de saturnisme infantile, il est conseillé de prélever plusieurs échantillons, de manière aléatoire sauf si des informations sont disponibles sur le lieu

14. Probablement à cause des résultats d'analyses laboratoires qui ne sont pas immédiats et surtout du surcoût qu'elle induit. Il est vraisemblable que l'analyse d'écailles de peinture n'est de plus pas perçue par les diagnostiqueurs comme étant nécessaire au rapport CREP.

de jeu de l'enfant. Le prélèvement est basé sur un échantillon composite d'au moins 3 échantillons jusqu'à 12 échantillons. Il est conseillé de prélever la couche 0-3 cm. La norme NF ISO 11464 [AFNOR, 2006] définit le pré-traitement des sols en vue d'analyses physico-chimiques. Aux États-Unis, l'U.S. HUD (*Department of Housing and Urban Development*) recommande de réaliser un prélèvement composite d'au moins 2 échantillons élémentaires avec une profondeur de 1 cm [U.S. HUD, 1995a].

3.4 Dans l'eau

Le prélèvement d'eau du robinet en vue du dosage du plomb peut être fait de différentes manières. L'article de J. Baron [Baron, 1997] décrit les différentes méthodes :

- Le prélèvement aléatoire est réalisé sur un site choisi au hasard. Il est aisé à réaliser car ils consiste simplement à prélever un volume d'eau. Avec ce type de prélèvement, pour un même point de prélèvement les résultats des concentrations peuvent varier de à 1 à 10 entre deux échantillons prélevés à différents moments.
- Le prélèvement au 2ème jet consiste à prélever après avoir purgé totalement le réseau. Le temps de séjour de l'eau étant minimal, la concentration mesurée devrait alors refléter la concentration minimale pour le point de mesure.
- Le prélèvement au 1ème jet consiste à prélever après avoir laissé stagner l'eau pendant une certaine période, généralement la nuit. Le temps de séjour de l'eau étant maximal, la concentration mesurée devrait alors refléter la concentration maximale pour le point de mesure.
- Le prélèvement après stagnation contrôlée est réalisé lorsque l'on souhaite se rapprocher de la durée moyenne entre deux soutirages afin d'évaluer une concentration en plomb reflétant la concentration moyenne de l'eau soutirée par l'usager.
- Le prélèvement proportionnel consiste a mettre un dispositif sur le robinet de l'usager permettant de collecter un volume d'eau égale à 5% de l'eau soutirée par l'usager. Ce mode de prélèvement permet ainsi de collecter un échantillon d'eau dont la concentration représente la concentration en plomb de l'eau utilisée par l'usager.

Ainsi la comparaison des concentrations en plomb dans l'eau du robinet est à réaliser avec prudence. En effet les méthodes de prélèvements d'eau utilisées dans différentes études n'étant pas toujours les mêmes, les différences peuvent être liées aux différentes méthodes mises en œuvre pour le prélèvement d'eau. Il faudra alors prendre soin d'interpréter les résultats, en particulier lors du calcul de la prévalence d'entités dépassant une certaine valeur règlementaire en plomb dans l'eau.
Une étude a été menée à travers les pays de l'Union Européenne (UE) afin de comparer les différents protocoles de prélèvement d'eau du robinet [European Commission et al., 1999]. Cette étude avait pour but de fournir une procédure de prélèvement d'eau permettant de satisfaire les besoins de la directive européenne 98/83/CE du conseil du 3 novembre 1998 relative à la qualité des eaux destinées à la consommation humaine. Cette dernière annonçait qu'à la fin 2013, une « valeur paramétrique »

de 10 μg de plomb par litre d'eau devait être respectée ; la valeur paramétrique à respecter étant de 25 μg/L à compter de fin 2003. Cette valeur de 10 μg/L doit s'appliquer à un échantillon d'eau destiné à la consommation humaine, prélevé au robinet par une méthode d'échantillonnage appropriée de manière à être « représentatif » d'une valeur moyenne hebdomadaire ingérée par les consommateur ; la « méthode d'échantillonnage appropriée » étant donnée *a posteriori* par l'étude européenne [European Commission et al., 1999].

Dans cette étude européenne, 11 régions tests ont été sélectionnées à travers les États Membres. Dans chaque région, des échantillons d'eau ont été réalisés dans 30 propriétés (logements), en faisant 3 prélèvements au cours d'une semaine dans chaque propriété. L'étude s'est intéressée aux performances du prélèvement aléatoire (RDT), du prélèvement après stagnation de 30 minutes (30MS) et du prélèvement au 2$^{\text{ème}}$ jet (FF) en comparaison avec le prélèvement proportionnel (COMP) considéré comme la référence.

Outre le fait d'être représentatif, c'est-à-dire de refléter une moyenne hebdomadaire de la concentration en plomb dans l'eau ingérée par l'usager, l'étude s'est intéressée à d'autres critères :

- la reproductabilité du prélèvement évaluée par le coefficient de variation des 3 prélèvements fait au sein d'une propriété ou par leur amplitude,
- le coût de la procédure de prélèvement,
- l'aspect pratique *in situ* (si la procédure est facilement mise en œuvre sur le terrain, si elle nécessite du personnel qualifié ou des outils de prélèvement spécifiques etc.),
- l'acceptabilité par l'usager.

Les conclusions du rapport sont présentées dans la table 2.

TABLE 2 – Performance des méthodes de prélèvement. COMP est la référence.

MÉTHODE	REPRÉ. ?	REPRO. ?	COÛT	ASPECT PRATIQUE	ACCEPTABILITÉ
COMP	-	-	158 à 176 ECU	faible	faible
RDT	oui	non	55 à 61 ECU	bon	bonne
30MS	oui	oui	105 à 124 ECU	moyen	moyenne
FF	non	oui	61 à 68 ECU	bon	bonne

Légende. REPRÉ. : Représentativité, REPRO. : Reproductibilité, COMP : Prélèvement proportionnel, RDT : prélèvement aléatoire, 30MS : prélèvement après stagnation de 30min, FF : prélèvement au 2$^{\text{ème}}$ jet, ECU : European Currency Unit - unité de compte de la Communauté Européenne avant l'adoption de l'Euro.

Au final, le rapport de l'étude européenne recommande le prélèvement aléatoire (RDT) ou après stagnation de 30min (30MS) à des fins de surveillance règlementaire c'est-à-dire ce que demande la directive européenne. Pour fournir à l'usager une valeur précise, répétable de sa moyenne hebdomadaire en plomb, la meilleure approche est le prélèvement après stagnation de 30min. Pour évaluer simplement le respect ou non d'une valeur réglementaire, ce même prélèvement ou le prélèvement aléatoire peut être utilisé.

3.5 Niveaux réglementaires en plomb

Concernant la peinture il n'y a pas de charge surfacique ou de concentration massique en plomb à respecter. Il ne semble pas avoir de valeur réglementaire dans d'autres pays et en particulier aux États-Unis à notre connaissance.
Il n'y a pas de valeur réglementaire française à respecter concernant la charge en plomb dans la poussière intérieure déposée au sol. La seule valeur réglementaire concernant ce média, est une valeur à respecter après travaux, égale à $1000\mu g/m^2$ en plomb acido-soluble (arrêté du 12 juillet 1999, arrêté du 25 avril 2006 et arrêté du 12 mai 2009) pour chaque échantillon de poussière prélevé. Aux États-Unis un seuil réglementaire existe. La valeur fédérale américaine de référence (*dust-lead hazard standard*) est de 430,5 $\mu g/m^2$ soit 40 $\mu g/ft^2$ [U.S. EPA, 2001].
Aucune valeur réglementaire à respecter n'existe en France concernant les sols meubles. En revanche un seuil réglementaire existe aux États-Unis et a pour valeur 400 mg/kg en plomb total [U.S. EPA, 2001]. Concernant les sols durs extérieurs, ni valeur réglementaire française ni valeur réglementaire américaine n'existe.
Concernant l'eau du robinet la valeur réglementaire européenne en vigueur jusqu'à la fin de l'année 2013 est égale à 25 $\mu g/L$ d'après la directive 98/83/CE du conseil du 3 novembre 1998. La valeur réglementaire en vigueur au 25 décembre 2013 sera de 10 $\mu g/L$. Aux États-Unis il n'existe pas de valeur réglementaire à respecter pour l'eau du robinet.

4 Les relations entre les médias environnementaux

Lanphear et Roghmann [Lanphear & Roghmann, 1997] ont montré, sur des données collectées dans les années 1990, que la peinture au plomb et les sols pollués par le plomb contaminaient les poussières intérieures. Les auteurs avaient néanmoins trouvé une plus forte contribution imputable à la peinture.
À partir de données collectées en 2000, une étude [Clark et al., 2004] a montré que les poussières extérieures situées à l'entrée du logement contaminaient les poussières intérieures.
Une étude plus récente, mais basée sur des données anciennes (1992-1996) [Hunt et al., 2006], a montré que le sol extérieur pouvait rapidement contaminer l'intérieur dès lors que l'introduction de sol via les chaussures était répétée. L'accumulation et la dispersion de sols pollués pouvait être rapide dès lors qu'un moyen de nettoyage (en particulier par voie humide) n'était pas ou peu mis en œuvre ; le nettoyage de l'intérieur semblant de plus échouer à supprimer toutes les particules de sols introduites dans le logement.
Concernant le sol extérieur, Lanphear *et al.* [Lanphear et al., 2003] avait de plus montré que la mise en place d'une réduction des niveaux en plomb dans les sols autour des logements conduisait à une diminution des niveaux en plomb des poussières intérieures. Les données avaient été collectées par une enquête menée en 1989 et une autre menée en 1998.
Enfin l'étude de Dixon *et al.* [Dixon et al., 2005b] a été montré à partir de données collectées entre 1994 et 1996, que les parties communes d'immeubles collectifs contenaient une quantité importante de peintures détériorées à base de plomb ainsi

qu'une quantité importante de poussières contaminées. Ces poussières contaminées contribuaient à la contamination de l'intérieur des logements par le plomb.

5 Sources d'exposition au plomb chez l'enfant

Le Guide d'investigation environnementale des cas de saturnisme de l'InVS [Bretin, 2006] résume en sa partie 2.3 les voies d'exposition au plomb chez l'enfant (on considère d'une manière générale, les enfants âgés de 6 ans au plus). On retiendra que les plus fortes expositions sont dues aux peintures anciennes. La contamination de l'enfant se fait par l'ingestion de poussières contaminées par ces peintures et plus rarement par l'ingestion d'écailles. Dans ce dernier cas, la contamination peut être rapidement très élevée. La contamination par les poussières se fait par le comportement naturel main-bouche de l'enfant. L'enfant peut également inhaler et ingérer des poussières émises par l'activité industrielle (extraction de minerai, fonderies de métaux, fabrication de batteries, *etc.*).

L'étude charnière en ce qui concerne l'exposition de l'enfant et les poussières résidentielles est sans doute la méta-analyse de B. Lanphear [Lanphear et al., 1998]. Douze jeux de données originales provenant de 12 études épidémiologiques ont été rassemblés afin d'estimer la contribution à la plombémie de l'enfant, du plomb contenu dans la poussière résidentielle au sol et contenu dans le sol extérieur. L'étude n'a retenu que les enfants de 6 à 36 mois car ces enfants étaient les plus à même de montrer une telle relation selon les auteurs, de par leur comportement jugé le plus à risque. Il a été nécessaire de « standardiser » certaines variables non homogènes à travers les 12 études (l'état des revêtements, le statut socio-économique, le comportement main-bouche de l'enfant) ; la standardisation n'est pas décrite en détails. Pour les études longitudinales se trouvant parmi les 12 études, un tirage aléatoire parmi les mesures répétées a été fait afin de ne garder qu'une seule mesure, tout comme les études où plusieurs enfants ont été enquêtées par ménage, un enfant a été sélectionné aléatoirement. Pour les études où certaines variables n'avaient pas été collectées, comme la concentration en plomb dans l'eau, ou qui avaient des données manquantes, la valeur a été imputée selon une loi-lognormale basée sur les données disponibles. La modélisation a été réalisée à travers une relation entre le logarithme de la plombémie et une fonction linéaire des logarithmes des variables environnementales donnant une concentration en plomb ; ce choix a été fait sur des travaux de la littérature jugeant qu'une telle modélisation était préférable à une modélisation où les variables n'auraient pas été transformées.
Les résultats montrent que la poussière intérieure est la source majeure d'apport en plomb chez l'enfant ayant une plombémie entre 100 et 250 μg/L. Elle indique de plus que la charge en plomb (« *loading* ») exprimée en masse de plomb par unité de surface (micro-grammes de plomb par pied carré dans les études américaines), est bien associée à la plombémie exprimée en micro-grammes par litre de sang (en micro-grammes par décilitre de sang dans les études américaines). Cette étude indique de plus une contribution du sol extérieur à la plombémie de l'enfant. L'auteur principal a montré en outre dans d'autres études que la réduction des concentrations en plomb dans le sol (proches de sites industriels polluants) permettait de faire dimi-

nuer la plombémie chez l'enfant. Les résultats de cette étude de 1998 seront utilisés par l'U.S. EPA pour établir une valeur fédérale de référence au début des années 2000. Cette valeur est alors passée en 1994 de 100 $\mu g/ft^2$ (environ 1 076 $\mu g/m^2$) en plomb dans les poussières intérieures au sol à 40 $\mu g/ft^2$ (environ 430 $\mu g/m^2$).
Des limites sont à signaler concernant cette étude : l'effet étude parmi les 12 études [15], la forme de la relation supposée, la non-prise en compte d'autres sources de contamination, les standardisations et les conversions de certaines variables, comme la concentration en plomb de la poussière de certaines études ($\mu g/g$) convertie en ($\mu g/m^2$).

L'enfant peut de plus s'intoxiquer en buvant l'eau du robinet en cas de présence de canalisations ou branchements en plomb ; cette eau peut contenir du plomb lorsqu'elle possède des caractéristiques physico-chimiques propices à la dissolution du plomb (cf. la section 2 précédente). À noter qu'en l'absence de source spécifique d'exposition, l'alimentation constitue la voie principale d'exposition, à plus de 80% de la dose hebdomadaire d'apport en plomb. La contamination des aliments se fait via les retombées atmosphériques ou par le sol (directe pour les produits végétaux ou indirecte pour les produits animaux via la chaîne alimentaire), ou par la préparation des aliments ou par leur stockage. Les groupes d'aliments qui sont les vecteurs de plomb les plus importants sont : le pain et les biscottes, les soupes, les légumes, les fruits, l'eau de boisson, les boissons non alcoolisées et les sucres et dérivés. La part importante de ces aliments vient avant tout de la quantité consommée, car les aliments les plus contaminés sont les crustacés et mollusques, ainsi que les abats.
Enfin, d'autre sources d'exposition non liées à l'habitat mais liées aux pratiques des occupants peuvent contribuer l'exposition des enfants. On peut citer de manière non exhaustive les cosmétiques traditionnels (certains khôls d'origine d'Afrique du Nord par exemple), la vaisselle (plats à tajine artisanaux vernis par exemple) ou la succion d'objets en plomb ou contenant du plomb (soldat en plomb, lest de rideaux par exemple).

15. Les différences observées pouvant être dues aux conditions expérimentales différentes pour chacune des 12 études (procédures de mesure par exemple ou des facteurs non identifiés différents selon les études et affectant le niveau de plombémie).

Spécificités des données d'enquête

Dans ce qui suit, les éléments de bibliographie utilisés sont les ouvrages suivants :
- [Tillé, 2001]
- [Lumley, 2010a]
- [Särndal et al., 1992]

Excepté pour les points particuliers, ils ne seront pas mentionnés de nouveau.

1 Introduction

Les données sur lesquelles le présent travail s'est basé, sont des données dites *données d'enquête*. Elles ont été obtenues à partir de l'investigation de logements. Ces logements figurent dans une liste qui a été construite à partir d'une procédure de tirage appelée *plan de sondage*. Cette liste définissant les logements à enquêter, est plus communément appelée *échantillon*, c'est-à-dire une portion de la population à laquelle on s'intéresse. La manière à partir de laquelle l'échantillon a été tiré constitue le plan de sondage. Par exemple, le tirage de 5 boules parmi 49 au Loto est un plan de sondage dit *aléatoire simple* ou chaque boule a la même probabilité d'être tirée à chaque tirage ; mais on pourrait décider de donner plus de chance à certaines d'entre elles d'être tirées en « pipant » ces boules. On aurait alors des *probabilités de sélection* dites *inégales*.

Ainsi, à partir de quelques logements observés, l'objectif est de fournir des résultats qui seront valables pour la population entière qui nous intéresse, à une incertitude près, si possible petite. Il s'agit là d'*extrapolation* ou encore d'*inférence*. Pour cela un outil appelé *estimateur* est construit pour chaque quantité qui nous intéresse, par exemple une moyenne. L'estimateur sera construit à partir d'une agrégation des données dont on dispose via l'échantillon et dont on étudiera les propriétés. On essaiera donc de construire l'estimateur avec les « meilleures propriétés ». Les bonnes propriétés seront généralement le fait que l'estimation soit la plus proche possible de la « vraie valeur » de la population (que l'on ne connaît pas) et que l'on puisse avoir confiance dans cette estimation. Ces outils appartiennent à une méthodologie appelée théorie de l'échantillonnage et de l'estimation en population finie ou théorie des sondages.

En théorie des sondages on revendique qu'**à partir d'un échantillon, on a la possibilité de faire presque aussi bien qu'un recensement, mais avec une**

durée de collecte et un coût beaucoup plus faibles.

Les éléments de méthode relatifs à la théorie des sondages utilisés dans ce travail seront présentés dans les sections 2 à 6 qui suivent. La collecte de données d'enquête produisant souvent des données manquantes pouvant dégrader la qualité des estimations, la section 7 considérera cette problématique. Enfin l'enquête Plomb-Habitat ayant permis de collecter les données utilisées dans ce travail sera présentée en section 8.

2 Terminologies

2.1 Population, échantillon et plan de sondage

On s'intéresse à une *population finie* U composée de N éléments distincts $e_1, e_2, \ldots, e_N$ qui peuvent être identifiés par un numéro d'ordre $1, 2, \ldots, N$. Par simplification on confond les éléments avec leur numéro d'ordre. On a donc :

$$U = \{1, 2, \ldots, N\}$$

U peut être par exemple la population des résidences principales en France. On s'intéresse au *caractère* Y sur ces N éléments : $\mathbf{y}_N = (y_1, y_2, \cdots, y_N)^\top$ où y_k est la valeur du caractère pour le k-ème élément. Les composantes y_k sont connues et sont donc des valeurs fixes et par conséquent non aléatoires. Y n'est alors pas une variable aléatoire, d'où le mot « caractère » employé pour Y plutôt que « variable » qui sous-entend souvent en statistique la notion de hasard. Il sera vu que l'aléa se situe uniquement au niveau de la procédure de tirage de l'échantillon. Un caractère peut être la concentration en plomb dans l'eau d'un logement par exemple. Dans l'approche développée ici, à un temps donné, on considère que le caractère est observable sur l'ensemble des N éléments de U à travers $\mathbf{y}_N$.

On note par $\mathbf{s}$ l'ensemble des n-uples non ordonnés sans remise possibles à partir de U c'est-à-dire l'ensemble des parties non vides de U^n défini par :

$$U^n = \overbrace{U \times U \times \cdots \times U}^{n \text{ fois}}$$

Un élément de $\mathbf{s}$ est appelé *échantillon* et noté s ; on a alors :

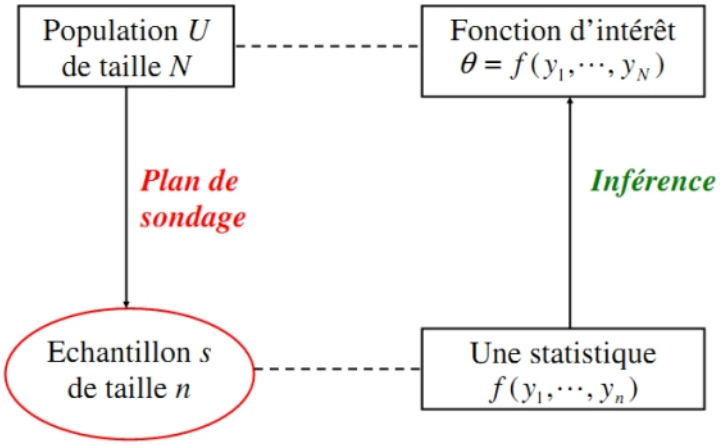
$$\mathbf{s} = \{s | s \subset U^n\} \setminus \emptyset$$

Par exemple si $U = \{1, 2, 3\}$ alors $\mathbf{s} = \{(1), (2), (3), (1, 2), (1, 3), (2, 3), (1, 2, 3)\}$.
$\mathbf{s}$ est de taille $\text{card}(\mathbf{s}) = 2^N - 1$.

On note $\mathbf{s}_n$ l'ensemble des échantillons de taille n, avec $n \leq \text{card}(U)$. Pour $U = \{1, 2, 3\}$ et $n = 2$ alors $\mathbf{s}_n = \{(1, 2), (1, 3), (2, 3)\}$.
$\mathbf{s}_n$ est de taille $\text{card}(\mathbf{s}_n) = C_N^n = N!/n!(N - n)!$.

L'objectif de l'*inférence* est de construire à partir de caractères une *statistique d'intérêt* sur un échantillon s tiré aléatoirement et d'extrapoler ensuite cette statistique à la population U comme illustré sur la figure 5. La notion de *statistique* sera précisée dans la section 2.2 qui suit. On pourrait calculer la valeur de cette statistique non plus sur une partie des éléments de la population, c'est-à-dire à partir d'un échantillon, mais à partir de tous les éléments de la population. On parlera alors de *recensement*. Lors d'un recensement on détermine la valeur exacte de la statistique (par exemple une moyenne) c'est-à-dire avec une incertitude nulle.

FIGURE 5 – Échantillonnage et inférence.



Pour disposer d'un échantillon s on met en œuvre un *plan de sondage* (non ordonné) *sans remise* noté $p(\cdot)$. Un plan de sondage est par définition une loi de probabilité sur $\mathbf{s}$ telle que

$$p(s) \geq 0, \text{ pour tout } s \in \mathbf{s}$$

et

$$\sum_{s \in \mathbf{s}} p(s) = 1$$

Si $U = \{1, 2, 3\}$ alors on peut définir par exemple un plan de sondage $p(\cdot)$ par

$$\begin{aligned} p(1,2) &= 1/3 \\ p(1,3) &= 1/3 \\ p(2,3) &= 1/3 \\ p(1) = p(2) &= p(3) = p(1,2,3) = 0 \end{aligned}$$

Il s'agit du plan de sondage consistant à tirer aléatoirement un échantillon de taille 2 avec la même probabilité de sélection.

s peut être alors vu comme la réalisation d'une variable aléatoire S dont la loi de probabilité est $p(\cdot)$. Ceci ancre la place de l'aléatoire dans la procédure d'analyse : l'aléatoire ne se situe qu'au niveau du plan de sondage $p(\cdot)$.

Pour pouvoir tirer aléatoirement des éléments dans une population U il est nécessaire de disposer d'une liste dans laquelle on puisse tirer. Cette liste s'appelle une *base de sondage* et couvre tous les éléments de la population, sans doublon, et en les identifiant sans ambiguïté afin de pouvoir les atteindre s'ils sont sélectionnés. S'il manque des éléments, on parle alors de *défaut de couverture* ce qui est problématique pour l'inférence.

Soit s un élément de $\mathbf{s}$. On définit la variable aléatoire *indicatrice* I_k par :

$$I_k = \begin{cases} 1 & \text{si le } k\text{-ème élément appartient à } s, \\ 0 & \text{sinon.} \end{cases}$$

I_k est en fait une notation simplifiée pour $I_k(S)$ car elle dépend de la variable aléatoire S.
L'*espérance* de I_k, notée $E(I_k)$ est égale à :

$$E(I_k) = \Pr(k \in s) = \sum_{s \ni k} p(s) \quad [16] \tag{2.1}$$

On note communément $E(I_k)$ par π_k la probabilité que le k-ème élément de U appartienne à s, autrement appelée *probabilité d'inclusion* d'ordre 1. On définit de façon similaire la probabilité d'inclusion d'ordre 2 comme étant la probabilité que les éléments k et l de U figurent communément dans l'échantillon :

$$\pi_{kl} = E(I_k I_l) = \Pr(k \in s \cap l \in s) = \sum_{s \ni k,l} p(s)$$

On peut montrer que sous n'importe quel plan de sondage $p(\cdot)$, la variance de I_k notée $V(I_k)$ est égale à $\pi_k(1 - \pi_k)$. De même on peut montrer que $Cov(I_k, I_l)$, la covariance de I_k et I_l, est égale à $\pi_{kl} - \pi_k \pi_l$.

Le *poids de sondage* de l'élément k appartenant à la population U est la quantité connue dès que le plan de sondage $p(\cdot)$ est défini ; il indique le nombre d'éléments représenté par k dans la population U. Ce poids de sondage est défini par l'inverse de la probabilité d'inclusion :

$$w_k = \frac{1}{\pi_k}, \forall k \in U$$

Cette expression requiert que tous les π_k soient non nulles (et c'est en fait aussi le cas pour les π_{kl}). Par exemple, si $\pi_k = 1/2000$ pour le logement k figurant dans l'échantillon s, le logement k représente 2000 logements dans la population U.

2.2 Estimateur, biais et variance

En théorie statistique générale, une *statistique* est une fonction à valeurs réelles dont les valeurs peuvent varier selon l'issue d'une expérience donnée. Elle doit pouvoir être calculée pour n'importe quelle issue.

16. La somme de l'équation (2.1) se fait sur les échantillons s (qui contiennent l'élément k) d'où la notation $s \ni k$.

Dans le cadre de la théorie de l'échantillonnage et de l'estimation en population finie, on s'intéresse à la manière dont une statistique d'intérêt varie en fonction des réalisations s (l'issue) de S.

Un *estimateur* est une statistique construite afin qu'elle produise, pour la majorité des échantillons, une valeur proche de la quantité d'intérêt (inconnue) de la population.

Une quantité d'intérêt est généralement appelé un *paramètre* et souvent noté θ. Des exemples de paramètres peuvent être la moyenne, la médiane ou un coefficient de régression. Si on ne s'intéresse qu'à un seul caractère, Y, on peut voir θ comme une fonction de $\mathbf{y}_N$:

$$\theta = f(y_1, y_2, \cdots, y_N)$$

Mais θ peut être aussi fonction de plusieurs caractères, par exemple lorsque l'on s'intéresse à un ratio. On donne en exemple de paramètre d'intérêt, de manière presque systématique, le *total* des valeurs d'un caractère Y :

$$\theta = \sum_U y_k = \sum_{k=1}^{N} y_k \tag{2.2}$$

La raison probable est que le total est en fait le seul paramètre en sondage qui soit d'intérêt, car à partir de son estimateur et de la variance de cet estimateur, il est possible de construire n'importe quel autre estimateur avec son incertitude.

Un estimateur est une statistique du paramètre d'intérêt θ noté par :

$$\hat{\theta} = \hat{\theta}(S)$$

Son *espérance* en théorie des sondages est définie par :

$$E(\hat{\theta}) = \sum_{s \in \mathbf{s}} p(s)\hat{\theta}(s)$$

et sa *variance* $V(\hat{\theta})$ par $E(\hat{\theta} - E(\hat{\theta}))^2$.

Un estimateur de θ est dit *sans biais* si la somme pondérée (par toutes les probabilités $p(s)$ possibles) de toutes ses réalisations, est égale à la valeur (inconnue) θ. Les réalisations de l'estimateur sont appelées *estimations*[17].
Un estimateur $\hat{\theta}$ est sans biais si $E(\hat{\theta}) = \theta$, le biais est alors $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ qui ne peut être calculé. Ce que l'on cherche est tout d'abord un estimateur si possible sans biais puisque on s'intéresse au fait que sa distribution d'échantillonnage soit étroitement concentrée autour de la valeur inconnue θ. On entend par *distribution d'échantillonnage* la distribution des réalisations de $\hat{\theta}$ sous le plan de sondage $p(\cdot)$; autrement dit la distribution de toutes les valeurs que peut prendre $\hat{\theta}$ à partir de tous les échantillons produits par $p(\cdot)$.

17. *estimate* en anglais ; et *estimator* pour estimateur.

Le caractère sans biais n'indique pas la manière dont se dispersent les valeurs prises par $\hat{\theta}$. Sa variance est cette mesure de dispersion et définit alors la notion d'*efficacité*. Dès lors pour choisir un estimateur on est souvent amené à considérer l'*erreur quadratique moyenne* (EQM) la plus petite possible, car si $EQM(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$ est petite, alors il y a des raisons de croire que l'échantillon tiré aléatoirement produira une estimation proche de la vraie valeur. Cependant si cette échantillon est malencontreusement « mauvais » (singulier) il pourra induire une estimation de θ se trouvant dans une des queues de distribution d'échantillonnage de $\hat{\theta}$. Il faudra néanmoins éviter d'utiliser des estimateurs fortement biaisés, car des intervalles de confiance valides ne peuvent être obtenus que si le biais n'est pas substantiel [Särndal et al., 1992, section 5.2]. Un estimateur approximativement sans biais avec une faible variance sera donc recherché.

Un *intervalle de confiance* $IC(s)$ est un intervalle aléatoire dépendant de l'échantillon s dont les bornes $a(s)$ et $b(s)$ sont telles que $a(s) \leq b(s)$. Le niveau de confiance $1-\alpha$ est la probabilité que le paramètre d'intérêt θ se situe dans cet intervalle : $\Pr[IC(s) \ni \theta]$. Si on note l'ensemble des échantillons s pouvant être tirés à partir du plan de sondage et contenant la vraie valeur θ par $\mathbf{s_{in}}$ et son complémentaire par $\mathbf{s_{out}}$ alors on aurait $\alpha = \sum_{s \in \mathbf{s_{out}}} p(s)$. Mais en pratique θ étant inconnu il est impossible de lister les éléments de $\mathbf{s_{out}}$ et de $\mathbf{s_{in}}$ et une méthode est alors nécessaire pour calculer $a(s)$ et $b(s)$ de manière à ce qu'un niveau de confiance $1-\alpha$ soit atteint. Pour les estimateurs utilisés en échantillonnage il est difficile de donner une méthode aboutissant à une valeur exacte $1-\alpha$ de niveau de confiance et des procédures approximatives de calcul doivent être utilisées. Un intervalle de confiance pour θ au niveau de confiance égal approximativement à $1-\alpha$ est donné par :

$$\hat{\theta} \pm z_{1-\alpha/2}[\widehat{V}(\hat{\theta})]^{1/2} \tag{2.3}$$

où $z_{1-\alpha/2}$ est le quantile d'une variable aléatoire Z suivant une loi $\mathcal{N}(0,1)$ tel que $\Pr(Z > z_{1-\alpha/2}) = \alpha/2$. L'intervalle (2.3) contiendra la vraie valeur θ dans approximativement $(1-\alpha) \times 100$ pourcents des répétitions de l'échantillon s tirées selon le plan $p(\cdot)$ si les deux conditions suivantes sont vérifiées :

1. La distribution d'échantillonnage de $\hat{\theta}$ est approximativement une distribution Gaussienne d'espérance θ et de variance $V(\hat{\theta})$.
2. Il existe un estimateur de la variance $\widehat{V}(\hat{\theta})$ de $V(\hat{\theta})$ tel que le ratio $\widehat{V}(\hat{\theta})/V(\hat{\theta})$ soit proche de 1 avec une probabilité proche de 1 quand la taille de l'échantillon est suffisamment grande.

La première condition assure que $(\hat{\theta}-\theta)/[V(\hat{\theta})]^{1/2}$ suive approximativement une loi $\mathcal{N}(0,1)$. Dès lors ajoutée à la deuxième condition

$$\frac{\hat{\theta}-\theta}{\sqrt{V(\hat{\theta})}} \left(\frac{V(\hat{\theta})}{\widehat{V}(\hat{\theta})} \right)^{1/2}$$

est une variable aléatoire suivant approximativement une loi $\mathcal{N}(0,1)$. Puisque

$$\frac{\hat{\theta}-\theta}{\sqrt{V(\hat{\theta})}} \left(\frac{V(\hat{\theta})}{\widehat{V}(\hat{\theta})} \right)^{1/2} = \frac{\hat{\theta}-\theta}{\sqrt{\widehat{V}(\hat{\theta})}} ,$$

on comprend l'équation (2.3) et en particulier l'utilisation du quantile normal $z_{1-\alpha/2}$.

Le principe extrêmement important en sondage pour obtenir un estimateur sans biais d'un total défini par l'équation 2.2, est le principe figurant dans l'estimateur dit « des valeurs dilatées » autrement appelé *π-estimateur* ou encore estimateur d'Horvitz-Thompson, ses créateurs. Horvitz et Thompson ont proposé pour le total d'une population U, $t_y = \sum_U y_k$, l'estimateur suivant :

$$\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k} = \sum_{k\in s} \frac{y_k}{\pi_k} \tag{2.4}$$

L'équation (2.4) peut être ré-écrite comme une fonction linéaire des variables indicatrices I_k :

$$\hat{t}_{y\pi} = \sum_U I_k \frac{y_k}{\pi_k}$$

ce qui permet de voir aisément que $\hat{t}_{y\pi}$ est sans biais pour $t_y = \sum_U y_k$ puisque $E(I_k) = \pi_k$ pour tout $k \in U$. On remarque de plus que cet estimateur ne pourrait être défini dès lors qu'une seule probabilité d'inclusion π_k serait nulle. L'expression y_k/π_k constitue une valeur dilatée. Elle permet de voir que lorsqu'un élément k est sur-représenté par rapport à un autre élément l i.e. $\pi_k > \pi_l$ alors y_k, le caractère Y associé à l'élément k, pèsera moins que y_l associé à l'élément l.

On peut montrer que pour un échantillonnage à taille fixe[18], la variance du π-estimateur peut être écrite de la manière suivante :

$$V(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}} Cov(I_k, I_l)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{2.5}$$

$$= -\frac{1}{2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}} (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{2.6}$$

Et un estimateur sans biais de $V(\hat{t}_{y\pi})$ est alors donné par :

$$\widehat{V}(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k\in s}\sum_{\substack{l\in s\\ l\neq k}} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

qui est appelé estimateur de *Sen-Yates-Grundy* [Yates & Grundy, 1953]. On voit donc que cet estimateur n'est défini que si les probabilités d'inclusion d'ordre 2, π_{kl}, sont toutes non nulles.

3 Les différents types de sondage

Dans cette section nous présentons uniquement les éléments de plans de sondage qui seront utiles pour la suite.

18. La taille de l'échantillon est fixée.

3.1 Sondage aléatoire simple

Sous un *sondage aléatoire simple* (sans remise), tout échantillon appartenant à $\mathbf{s}_n$, c'est-à-dire l'ensemble des échantillons de taille fixe égale à n, a la même probabilité d'être sélectionné :

$$p(s_1) = p(s_2), \forall s_1, s_2 \in \mathbf{s}_n$$

Puisque $\text{card}(\mathbf{s}_n) = C_N^n$, où N est la taille de la population U,

$$p(s) = \begin{cases} 1/C_N^n & \text{si } s \text{ est de taille } n \\ 0 & \text{sinon} \end{cases}$$

La *probabilité d'inclusion d'ordre 1* de l'élément k de U, π_k, est donc égale à

$$\pi_k = \sum_{s \ni k} p(s) = \sum_{s \ni k} \frac{1}{C_N^n} = \frac{1}{C_N^n} \times \sum_{s \ni k} 1$$

où $\sum_{s \ni k} 1$ dénombre le nombre d'échantillons contenant l'élément k.
Si l'élément k est sélectionné d'office, il ne reste qu'à tirer $n-1$ éléments pour construire s, dans une population ne contenant plus que $N-1$ éléments. Ainsi $\sum_{s \ni k} 1 = C_{N-1}^{n-1}$ et donc

$$\pi_k = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}, \forall k \in U$$

Par un raisonnement analogue on obtient que les *probabilités d'inclusion d'ordre 2* sont égales à :

$$\pi_{kl} = \frac{N(n-1)}{N(N-1)}, \forall k, l \in U, k \neq l$$

La proportion d'éléments présents dans l'échantillon au regard du nombre N d'éléments dans la population U s'appelle la *fraction de sondage* et est notée traditionnellement f :

$$f = \frac{n}{N}$$

Dans un sondage aléatoire simple sans remise, la probabilité d'inclusion est donc égale à la fraction de sondage.

Sous un sondage aléatoire simple sans remise, le π-estimateur du total dans la population, $t_y = \sum_U y_k$, peut être écrit par :

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_s y_k = N\overline{y}_s$$

avec $\overline{y}_s = \sum_s y_k/n$. À partir de l'équation (2.6) on a alors que sa variance est donnée par :

$$V(\hat{t}_{y\pi}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2 = N^2 \frac{1-f}{n} S_{yU}^2$$

où $S_{yU}^2 = \sum_U (y_k - \overline{y}_U)^2/(N-1)$. Un estimateur sans biais de la variance est :

$$\widehat{V}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} S_{ys}^2$$

où $S_{ys}^2 = \sum_s (y_k - \overline{y}_s)^2/(n-1)$.

3.2 Sondage proportionnel à la taille

S'il existe une information dans la base de sondage dont on dispose, appelée *information auxiliaire*, et qu'elle est reliée au caractère d'intérêt, alors il peut être judicieux de prendre en compte cette information pour tirer les éléments de l'échantillon.

Considérons le π-estimateur $\hat{t}_{y\pi} = \sum_s y_k/\pi_k$. Supposons qu'il soit possible de trouver un plan de sondage à partir duquel on ait :

$$y_k/\pi_k = c, \quad k = 1, \ldots, N$$

où c est une constante. Alors pour n'importe quel échantillon s de taille n on aura :

$$\hat{t}_{y\pi} = nc$$

Puisque $\hat{t}_{y\pi}$ est alors constant, sa variance sera nulle. Il est en pratique jamais possible de trouver une telle situation mais l'idée est là. Si on arrive à faire en sorte que π_k soit proportionnelle à une quantité X connue dans la base de sondage, avec X proportionnelle à Y, cela devrait alors induire des rapports y_k/π_k approximativement constants. Ainsi la variance du π-estimateur sera petite. L'équation (2.6) donnant l'expression de la variance du π-estimateur permet de comprendre algébriquement cela :

$$V(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k\in U}\sum_{\substack{l\in U \\ l\neq k}}(\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

Si les y_k/π_k sont approximativement constants alors le terme de la parenthèse sera proche de zéro et la variance sera alors petite.

Ainsi dans un *plan proportionnel à la taille*, taille donnée par une information auxiliaire, un élément de la population aura plus de chance d'être tirée qu'un autre. Mais ce déséquilibre sera rectifié dans l'estimation : le premier élément contribuera moins que le second puisque son poids de sondage est alors plus petit.

3.3 Sondage stratifié

Imaginons que la base de sondage dispose d'une information auxiliaire qualitative permettant de faire des groupes d'éléments dans la population U. Si ces groupes permettent de rassembler des éléments qui se ressemblent du point de vue du caractère d'intérêt Y, et donc de construire des groupes homogènes, alors cela permettra de diminuer naturellement la variance de l'estimateur fonction de Y.

Ces groupes sont appelées *strates* et divisent la population U en H sous-populations disjointes U_h :

$$U = \bigcup_{h=1}^{H} U_h \text{ avec } U_h \cap U_j = \emptyset \text{ si } h \neq j$$

Un *sondage stratifié* consiste alors à sélectionner un échantillon aléatoire s_h à partir de U_h selon un plan de sondage $p_h(\cdot)(h = 1, \ldots, H)$. Le tirage dans une strate est indépendant du tirage dans les autres strates.
L'échantillon final s n'est autre que la réunion des échantillons s_h :

$$s = s_1 \cup s_2 \cup \cdots \cup s_H$$

et parce que les tirages sont indépendants on a :

$$p(s) = p_1(s_1)p_2(s_2)\cdots p_H(s_H) \tag{3.1}$$

Le nombre d'éléments dans une strate h est appelée la *taille de la strate* dans la population et est notée N_h et supposée connue ; on a donc : $N = \sum_{h=1}^{H} N_h$. Un total dans la population peut alors être décomposé par :

$$t_y = \sum_U y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} t_{yh}$$

où t_{yh} est le total des valeurs prises par Y sur les éléments de U_h.
Le π-estimateur du total $t_y = \sum_U y_k$ peut s'écrire par :

$$\hat{t}_{y\pi strate} = \sum_{h=1}^{H} \hat{t}_{y\pi h}$$

où $\hat{t}_{y\pi h}$ est le π-estimateur de $t_{yh} = \sum_{U_h} y_k$.
La variance de $\hat{t}_{y\pi strate}$ est :

$$V(\hat{t}_{y\pi strate}) = \sum_{h=1}^{H} V(\hat{t}_{y\pi h})$$

avec $V(\hat{t}_{y\pi h})$ la variance de $\hat{t}_{yh}$. Un estimateur sans biais de la variance est donné par :

$$\widehat{V}(\hat{t}_{y\pi strate}) = \sum_{h=1}^{H} \widehat{V}(\hat{t}_{y\pi h})$$

en faisant l'hypothèse qu'il existe un estimateur sans biais $\widehat{V}(\hat{t}_{y\pi h})$ de la variance pour chaque h.

D'après l'équation (3.1) il est possible d'utiliser différents plans de sondage $p_h(\cdot)$ pour les différentes strates. Mais très souvent le même plan de sondage est utilisé dans chaque strate. Dans la plupart des applications, $p_h(\cdot)$ est un plan de sondage aléatoire simple. Outre la réduction de la variance évoquée en préambule, on peut aussi être intéressé spécifiquement à la précision dans des sous-populations de U, appelées *domaines*. Si ces sous-populations sont identifiables dans la base de sondage, elles peuvent alors définir des strates. Il est alors possible de définir un plan de sondage adéquate pour chacune des strates.

Sous un plan de sondage stratifié aléatoire simple $p(\cdot)$, la variance du π-estimateur $\hat{t}_{y\pi strate}$ du total dans la population est :

$$V(\hat{t}_{y\pi strate}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} S^2_{yU_h}$$

où n_h est la taille de la strate h dans l'échantillon, $S^2_{yU_h} = \sum_{U_h}(y_k - \overline{y}_{U_h})^2/(N_h - 1)$ est la variance des valeurs de Y dans la strate h et $\overline{y}_{U_h} = \sum_{U_h}(y_k/N_h)$.

Un estimateur sans biais de la variance du π-estimateur est :

$$\widehat{V}(\hat{t}_{y\pi strate}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} S^2_{ys_h}$$

où $S^2_{ys_h} = \sum_{s_h}(y_k - \overline{y}_{s_h})^2/(n_h - 1)$ est la variance des valeurs de Y dans l'échantillon dans la strate h.

En sondage stratifié, se pose la question de la taille (n_h) pour chaque strate de l'échantillon. Lorsque la fraction de sondage $f_h = n_h/N_h$ dans la strate h est égale à la fraction de sondage globale $f = n/N$ et ceci pour toutes les strates, on parle alors de plan de sondage stratifié avec *allocation proportionnelle*. Par exemple, si une population de $N = 1000$ logements est composée de $N_1 = 600$ logements collectifs et de $N_2 = 400$ logements individuels, et que l'on tire un échantillon de taille $n = 100$ selon un sondage stratifié sur le type de logement, l'allocation proportionnelle consistera à tirer 60 logements collectifs et 40 logements individuels. Ainsi dans ce type de plan, tous les individus ont la même probabilité d'inclusion. On parle alors de sondage *auto-pondéré*.

Lorsque l'on parle d'échantillon *représentatif*, dans l'esprit des gens cela indique qu'une allocation proportionnelle a été utilisée pour sélectionner l'échantillon. Autrement dit l'échantillon est une miniature de la population. Or cette allocation n'est pas l'allocation optimale en ce qui concerne la variance. En effet, si l'on considère le total des valeurs de Y dans une population, Neyman a montré en 1934 [Neyman, 1934], que sous un sondage stratifié aléatoire simple, l'allocation optimale dans une strate h est proportionnelle à la fois à la taille de la strate dans la population, N_h, et à la variabilité des valeurs de Y dans la strate, S_{yU_h} :

$$n_h = \frac{n N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}$$

Ce résultat est appelé *allocation de Neyman*. En pratique il est difficile d'utiliser directement ce résultat car il requiert de connaître la variance des valeurs de Y dans chacune des strates, ce qui n'est jamais le cas. Pour s'en rapprocher on peut être amené à utiliser les résultats d'enquêtes passées qui renseigneraient sur la dispersion des valeurs de Y.

En outre, ce résultat indique qu'il est nécessaire de sur-représenter les éléments appartenant à des strates où la dispersion est la plus grande, ce qui met en défaut

le concept de *représentativité* ancré dans les esprits. Voir invoquer dans un rapport d'enquête la qualité de l'échantillonnage justifiée par sa représentativité de la population, laisse à penser que l'échantillonnage a été réalisé avec une méconnaissance totale de la théorie des sondages. Malheureusement ce concept de représentativité permet de rassurer les observateurs puisqu'ils pensent alors que l'échantillonnage a été bien fait. Si cette représentativité possède un quelconque intérêt, c'est uniquement un intérêt de communication. Mais il faudrait alors au moins spécifié de manière transparente sur quelles variables auxiliaires l'échantillon est sensé représenter une miniature de la population.

3.4 Sondage à plusieurs degrés

Une base de sondage listant par définition tous les éléments de la population U à laquelle on s'intéresse, est pour certains types de population difficile à construire. Cette raison amène souvent à utiliser un *plan de sondage à plusieurs degrés* pour tirer l'échantillon.

Supposons que U soit l'ensemble des résidences principales en France métropolitaine. Il n'existe pas de base de sondage identifiant ces résidences principales avec le moyen de les atteindre facilement. Par exemple, si la prise de contact se fait par téléphone, cela nécessiterait qu'à chaque résidence principale soit associée un numéro de téléphone, ce qui n'est pas le cas, puisque certains logements n'ont pas de ligne téléphonique ouverte, et d'autres ont un numéro de téléphone caché. On peut alors procéder en indiquant que toutes les résidences principales se trouvent dans une des 22 régions administratives de la France métropolitaine. On décide alors de tirer aléatoirement, disons 5 régions parmi les 22. On construit ensuite une base de sondage de résidences principales, en supposant que cela soit possible, pour chacune des 5 régions tirées (disons que l'on a les moyens techniques et financiers de le faire pour 5 régions au plus). On tire ensuite un nombre fixé de résidences dans chacune des 5 régions. On aura alors procédé à un tirage à 2 degrés :

- au premier degré, tirage de régions ;
- au second, tirage de résidences principales.

Une autre raison d'utiliser un sondage à plusieurs degrés peut être technique et financière. Par exemple il existe une liste des écoles primaires et élémentaires françaises au ministère de l'éducation nationale. Si on souhaite enquêter *in situ* des écoles, on pourrait tirer directement des écoles à partir de cette liste. Cependant un tel tirage induirait un éparpillement (aléatoire) de ces écoles sur le territoire français. Cela impliquerait donc des déplacements nombreux à faire par les enquêteurs et induirait alors un temps de réalisation de l'enquête plus grand et plus coûteux que si les écoles sélectionnées se regroupaient par zones géographiques peu étendues. La solution pourrait alors être de nouveau de tirer un premier degré de zones géographiques.

Plus formellement, supposons que la population $U = \{1, \ldots, i, \ldots, N\}$ soit décomposables en N_{psu} éléments, appelées *unités primaires* (*primary sampling units*

(PSUs)) : $U_1, \ldots, U_j, \ldots, U_{N_{\text{psu}}}$. Par exemple U est l'ensemble des résidences principales en France métropolitaine et les unités primaires sont les régions administratives. Par simplification on confondra les unités primaires avec leur numéro d'identification. L'ensemble de la population des unités primaires est donc noté $U_{\text{psu}} = \{1, \ldots, j, \ldots, N_{\text{psu}}\}$. La taille de U_j est le nombre d'éléments N_j, appelées *éléments secondaires* (*secondary sampling units* (SSUs)) qu'elle contient. On a donc $N = \sum_{U_{\text{psu}}} N_j = \sum_{j=1}^{N_{\text{psu}}} N_j$.

On se limite à présenter un plan à 2 degrés, le raisonnement étant analogue pour un plan à strictement plus de 2 degrés.
Au premier degré, un échantillon s_{psu} d'unités primaires est tiré à partir de U_{psu} ($s_{\text{psu}} \subset U_{\text{psu}}$) selon un plan de sondage $p_{\text{psu}}(\cdot)$. Au second degré, pour chaque $j \in s_{\text{psu}}$, un échantillon s_j est tiré dans U_j ($s_j \subset U_j$) selon un plan de sondage $p_j(\cdot \mid s_{\text{psu}})$. L'échantillon souhaité composé d'éléments de U est alors :

$$s = \bigcup_{j \in s_{\text{psu}}} s_j$$

Cette présentation permet donc que l'on mette en œuvre un plan de sondage spécifique au deuxième degré selon l'échantillon obtenu au premier degré. Dans le cas contraire on dira que le plan à 2 degrés est *invariant* i.e. $p_j(\cdot \mid s_{\text{psu}}) = p_j(\cdot)$. De plus le sous-échantillonnage dans U_j pourrait dépendre du sous-échantillonnage fait dans U_l. Dans le cas contraire on dira que le plan à 2 degrés est *indépendant* (comme en stratification). On se limitera ici aux plans invariants et indépendants.

On note $\pi_j^{\deg_1}$ la probabilité d'inclusion (d'ordre 1) de l'unité primaire U_j au premier degré. La probabilité d'ordre 2 au premier degré, i.e. de tirer conjointement U_j et U_k est notée $\pi_{jk}^{\deg_1}$. On a :

$$Cov(I_j^{\deg_1}, I_k^{\deg_1}) = \Delta_{jk}^{\deg_1} = \begin{cases} \pi_{jk}^{\deg_1} - \pi_j^{\deg_1}\pi_k^{\deg_1} & \text{si } j \neq k \\ \pi_j^{\deg_1}(1 - \pi_j^{\deg_1}) & \text{si } j = k \end{cases}$$

On note par $\pi_{i|j}$ la *probabilité conditionnelle* de tirer au second degré l'élément i de U_j sachant que U_j a été tirée au premier degré ; et $\pi_{il|j}$ la probabilité de tirer conjointement les éléments secondaires i et l étant donné que l'élément j a été sélectionnée. On a :

$$Cov(I_i, I_l) = \Delta_{il|j} = \begin{cases} \pi_{il|j} - \pi_{i|j}\pi_{l|j} & \text{si } i \neq l \\ \pi_{i|j}(1 - \pi_{i|j}) & \text{si } i = l \end{cases}, \quad j = 1, ..., M.$$

La probabilité d'inclusion d'un élément i de U est simplement donnée par le produit de la probabilité d'inclusion de son élément primaire multipliée par sa probabilité conditionnelle :

$$\pi_i = \pi_j^{\deg_1}\pi_{i|j}, \quad i \in U_j$$

La probabilité d'inclusion (d'ordre 2) des éléments de U, i et l, est égale à $\pi_{il} = \pi_j^{\deg_1}\pi_{il|j}$ si i et l appartiennent à la même unité primaire U_j, et $\pi_{il} = \pi_{jk}^{\deg_1}\pi_{i|j}\pi_{l|k}$ si i et l

appartiennent à deux unités primaires distinctes, U_j et U_k respectivement.

Sous un plan de sondage à 2 degrés, le π-estimateur du total $t = \sum_U y_k$ des valeurs de Y dans la population U peut être écrit comme :

$$\hat{t}_{y\pi} = \sum_{j \in s_{\text{psu}}} \sum_{i \in s_j} \frac{y_i}{\pi_j^{\deg_1} \pi_{i|j}} = \sum_{j \in s_{\text{psu}}} \frac{\hat{t}_{y\pi j}}{\pi_j^{\deg_1}}$$

avec $\hat{t}_{y\pi j}$ le π-estimateur du total t_{yj} des valeurs de Y sur les éléments de l'unité primaire U_j :

$$\hat{t}_{y\pi j} = \sum_{i \in s_j} \frac{y_i}{\pi_{i|j}}$$

Concernant la variance de $\hat{t}_{y\pi}$ on peut montrer que :

$$V(\hat{t}_{y\pi}) = V_{\text{PSU}} + V_{\text{SSU}}$$

où V_{PSU} est le terme de la variance se rapportant aux unités primaires

$$V_{\text{PSU}} = \sum_{j=1}^{M} \sum_{k=1}^{M} \frac{t_{yj} t_{yk}}{\pi_j^{\deg_1} \pi_k^{\deg_1}} \Delta_{jk}^{\deg_1}$$

et V_{SSU} le terme de la variance se rapportant aux unités secondaires

$$V_{\text{SSU}} = \sum_{j=1}^{M} \frac{V(\hat{t}_{y\pi j})}{\pi_j^{\deg_1}}$$

et

$$V(\hat{t}_{y\pi j}) = \sum_{i \in U_j} \sum_{l \in U_j} \frac{y_i y_l}{\pi_{i|j} \pi_{l|j}} \Delta_{il|j}, \quad j = 1, \ldots, M$$

Enfin, sous un plan de sondage à 2 degrés un estimateur sans biais de la variance du π-estimateur $\hat{t}_{y\pi}$ est donné par :

$$V(\hat{t}_{y\pi}) = \widehat{V}_{\text{PSU}} + \widehat{V}_{\text{SSU}}$$

où $\widehat{V}_{\text{PSU}}$ est l'estimateur (non sans biais) de la variance associée au niveau des unités primaires

$$\widehat{V}_{\text{PSU}} = \sum_{j \in s_{\text{psu}}} \sum_{k \in s_{\text{psu}}} \frac{\hat{t}_{y\pi j} \hat{t}_{y\pi k}}{\pi_j^{\deg_1} \pi_k^{\deg_1}} \frac{\Delta_{jk}^{\deg_1}}{\pi_{jk}^{\deg_1}}$$

et $\widehat{V}_{\text{SSU}}$ est l'estimateur (non sans biais) de la variance associée au niveau des unités secondaires

$$\widehat{V}_{\text{SSU}} = \sum_{j \in s_{\text{psu}}} \frac{\widehat{V}(\hat{t}_{y\pi j})}{\pi_j^{\deg_1}}$$

et

$$\widehat{V}(\hat{t}_{y\pi j}) = \sum_{i \in s_j} \sum_{l \in s_j} \frac{y_i y_l}{\pi_{i|j} \pi_{l|j}} \frac{\Delta_{il|j}}{\pi_{il|j}}$$

avec $\pi_{ii|j} = \pi_{i|j}$.

Pour strictement plus de 2 degrés, on pourra se reporter à [Tillé, 2001, section 9.3] ou [Särndal et al., 1992, section 4.4].

Au moment de l'analyse il est parfois nécessaire de procéder à des regroupements de strates lorsque le plan est un plan stratifié au premier degré. Si la population possède une strate avec une seule PSU, il est alors nécessaire d'avoir une fraction de sondage de 100 % dans cette strate puisque sinon la fraction de sondage est de 0 %. La strate ne contribue alors pas à la variance du premier degré (mais peut contribuer aux calculs des variances des degrés suivants). Si dans la population, la strate possède strictement plus d'une PSU et qu'une seule PSU est présente dans l'échantillon, alors si 2 éléments de la strate se trouvent dans 2 PSUs différentes dans la population, leur probabilité d'inclusion d'ordre 2 est nulle, ce qui ne doit pas être le cas (cf. section 2.2 de cette partie). Dans les enquêtes, les strates avec une seule PSU surviennent à cause de la non-réponse ou lorsque l'on cherche délibérément à réduire la variance par une stratification très fine.
La meilleure méthode pour manipuler une strate avec une seule PSU est de la regrouper avec une autre strate similaire autant que possible dans la population. Les estimateurs de la variance seront alors plus précis que s'ils avaient été construits à partir des strates regroupées à la conception du plan.

3.5 Sondage en deux phases

La précision du π-estimateur peut être améliorée en stratifiant la population comme vu précédemment. Cela requiert de disposer d'une ou de plusieurs informations auxiliaires. Sans cela, si on devait alors mettre en œuvre un sondage aléatoire simple par exemple, cela demanderait une taille d'échantillon plus grande pour obtenir une précision acceptable, mais alors le coût de collecte serait certainement plus important. Il est possible de faire autrement. On pourra tout d'abord dans une première phase, tirer un échantillon s_a selon un plan aléatoire simple. Il sera alors collecté, sur les éléments de s_a, des informations auxiliaires par un moyen peu coûteux. Puis dans une seconde phase tirer un sous échantillon issu de s_a selon un plan $p(\cdot \mid s_a)$ en prenant en compte les informations auxiliaires collectées pour s_a ; le caractère d'intérêt, Y, n'étant observé que sur ce sous-échantillon. Un *sondage en 2 phases* aura alors été réalisé.

Plus formellement, notons s_a l'échantillon tiré en une première phase à partir de la population U, par un plan de sondage quelconque, $p_a(\cdot)$. s_b est le sous-échantillon tiré à partir de s_a par un plan de sondage $p(\cdot \mid s_a)$.
On note :

- $\pi_k^a = \Pr(k \in s_a)$ la probabilité d'inclusion de l'élément k d'appartenir à l'échantillon tiré en première phase ;
- $\pi_{kl}^a = \Pr(k \in s_a \cap l \in s_a)$ la probabilité d'inclusion conjointe.

De plus on note :

$$\Delta^a_{kl} = \begin{cases} \pi^a_{kl} - \pi^a_k \pi^a_l & \text{si } k \neq l \\ \pi^a_k(1-\pi^a_k) & \text{si } k = l \end{cases},$$

les valeurs des covariances des variables indicatrices I_k et I_l. Pour le sous-échantillon s_b on note de même par $\pi^b_k = \Pr(k \in s_b \mid s_a)$ la probabilité d'inclusion de l'élément k de s_a d'appartenir à l'échantillon s_b ; et $\pi^b_{kl} = \Pr(k \in s_b \cap l \in s_b \mid s_a)$ la probabilité d'inclusion conjointe conditionnelle. On note alors :

$$\Delta^b_{kl} = \begin{cases} \pi^b_{kl} - \pi^b_k \pi^b_l & \text{si } k \neq l \\ \pi^b_k(1-\pi^b_k) & \text{si } k = l \end{cases}$$

Les probabilités d'inclusion et les covariances relatives à la seconde phase, π^b_k, π^b_{kl} et Δ^b_{kl} dépendent de s_a et sont donc des variables aléatoires.

Comme précédemment la tâche est de trouver un estimateur sans biais du total $t_y = \sum_U y_k$ des valeurs d'un caractère Y dans U. Un candidat naturel est le π-estimateur :

$$t_{y\pi} = \sum_{s_b} \frac{y_k}{\pi_k}$$

Or π_k, la probabilité d'inclusion de l'élément k de U est égale à :

$$\pi_k = \pi^a_k E_{p_a}(\pi^b_k)$$

π^a_k est calculable mais pas $E_{p_a}(\pi^b_k)$ qui dépend de ce qui a été obtenu en première phase.

Le total est alors estimé par l'estimateur sans biais suivant :

$$\hat{t}_{y\pi^\star} = \sum_{k \in s_b} \frac{y_k}{\pi^a_k \pi^b_k}$$

La variance de l'estimateur $\hat{t}_{y\pi^\star}$ est donnée par :

$$V(\hat{t}_{y\pi^\star}) = \sum_{k\in U}\sum_{l\in U} \frac{y_k y_l}{\pi^a_k \pi^a_l}\Delta^a_{kl} + E_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a} \frac{y_k y_l}{\pi^a_k \pi^b_k \pi^a_l \pi^b_l}\Delta^b_{kl}\right)$$

et peut être estimée par :

$$\widehat{V}(\hat{t}_{y\pi^\star}) = \sum_{k\in s_b}\sum_{l\in s_b} \frac{y_k y_l}{\pi^a_k \pi^a_l}\frac{\Delta^a_{kl}}{\pi^a_{kl}\pi^a_{kl}} + \sum_{k\in s_b}\sum_{l\in s_b} \frac{y_k y_l}{\pi^a_k \pi^b_k \pi^a_l \pi^b_l}\frac{\Delta^b_{kl}}{\pi^b_{kl}}$$

On pourra voir [Särndal et al., 1992, section 9.4] pour l'étude de l'estimateur $\hat{t}_{y\pi^\star}$ lorsque l'on utilise un plan stratifié pour la deuxième phase.

Bien que théoriquement il y ait d'importantes différences entre un plan de sondage à deux degrés et à deux phases, il y a relativement peu d'impact dans les applications entre ces deux schémas de plan [Lumley, 2010a, p. 157].

4 Redressement par post-stratification

La terminologie *redressement* recouvre les méthodes d'amélioration des estimations au niveau de l'estimation et non plus au niveau du dimensionnement du plan de sondage. Par exemple la stratification est une méthode d'amélioration des estimations *a priori*. La stratification n'est possible que si on dispose d'une information auxiliaire sur toutes les éléments de la base de sondage. Cela peut ne pas être le cas, mais on peut connaître une information agrégée, souvent un total, sur la population, souvent à partir d'un recensement. Par exemple, on peut ne pas avoir l'information de la période de construction de chaque logement dans une base de sondage, mais on peut connaître le nombre total de logements dans la population par période de construction. Dans ce cas on pourra utiliser les périodes de construction comme strates, plus précisément comme post-strates, pour améliorer les estimations.
La *post-stratification* est la méthode de base permettant d'utiliser de l'information auxiliaire à l'étape de l'estimation.

Considérons une ou plusieurs informations auxiliaires qualitatives permettant de partitionner U en H sous-ensemble U_h, les post-strates. De même l'échantillon s est partionné en H sous-échantillons s_h.
Par exemple si les informations auxiliaires sont le type de logement (collectif/individuel) et l'environnement extérieur (urbain/rural), $2 \times 2 = 4$ post-strates sont construites. Chaque post-strate a un effectif N_h connu dans la population et naturellement connu dans l'échantillon par n_h ; $\sum_{h=1}^{H} N_h = N$ et $\sum_{h=1}^{H} n_h = n$. Le π-estimateur du total est donné par :

$$\hat{t}_{y\pi} = \sum_{s} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{s_h} \frac{y_k}{\pi_k}$$

On construit un estimateur dit *estimateur post-stratifié*, $\hat{t}_{y\pi post}$ par l'expression :

$$\hat{t}_{y\pi post} = \sum_{h=1}^{H} \sum_{s_h} \frac{y_k}{\widetilde{\pi}_k} = \sum_{h=1}^{H} \sum_{s_h} \widetilde{w}_k y_k$$

où $\widetilde{w}_k = 1/\widetilde{\pi}_k$ et $\widetilde{\pi}_k = \pi_k \times c_h$ pour k appartenant à s_h. c_h est appelé le *coefficient de redressement* pour la strate h.
N_h la taille de U_h peut être estimée par $\widehat{N}_h = \sum_{s_h} 1/\pi_k$ qui est différent de N_h si une stratification *a priori* n'a pas été faite sur les U_h. L'idée est donc de redresser cette estimation par un coefficient de redressement pour chaque post-strate de manière à ce que cette somme égale N_h. Ainsi :

$$c_h = \frac{N_h}{\sum_{s_h} 1/\pi_k}, \quad h = 1, \ldots, H$$

Dans le cas d'un sondage aléatoire simple, tous les π_k sont égales et valent $f = n/N$. Dans ce cas on aura :

$$c_h = \frac{N_h \times n}{n_h \times N}, \quad h = 1, \ldots, H$$

Les propriétés de l'estimateur post-stratifié peuvent être vues dans [Tillé, 2001, section 7.6] ou [Särndal et al., 1992, section 10.1.2]. On retiendra que l'estimateur

post-stratifié peut être biaisé si certaines post-strates sont vides. De plus il est plus intéressant en terme de réduction de la variance d'utiliser une stratification *a priori*. Comme pour la stratification *a priori*, les post-strates doivent être autant que possible reliées au caractère d'intérêt étudié pour que la diminution de la variance soit importante. Si la variable de post-stratification est indépendante du caractère d'intérêt, la variance de l'estimateur post-stratifié est alors supérieure à la variance du π-estimateur.

5 Approche plan - approche modèle

Dans ce qui suit l'élément de bibliographie utilisé est [Beaumont & Haziza, 2012].

Lorsque l'on s'intéresse à la précision d'un estimateur d'un paramètre d'intérêt, il a été montré dans la littérature qu'il n'existait pas de plan de sondage optimal. Autrement dit il n'existe pas de combinaison miracle « plan de sondage/méthode d'estimation », permettant d'obtenir une EQM (cf. section 2.2 de cette partie) plus petite que n'importe quel autre plan de sondage, et ceci quelque soit $\mathbf{y}_N$.

D'autre part les EQM des plans de sondage même classiques, se sont avérées d'expression suffisamment complexe pour que nous ne soyons pas capables dans la plupart des cas en pratique, de comparer deux plans de sondage pour déterminer lequel serait le meilleur ; même problème pour la comparaison de deux méthodes de redressement pour choisir la plus efficace.

C'est pour pouvoir palier à ces deux difficultés et surtout la seconde, que les théoriciens des sondages ont développé une autre approche que l'approche par le plan de sondage (*design-based*) jusqu'alors exposée ici. Cette approche est appelée approche modèle (*model-based*). On ne présente ici que la philosophie générale de cette approche dans la mesure où elle pourrait être le sujet d'un livre complet [Ardilly, 2006].

Dans l'approche basée sur le plan de sondage, on cherchait à inférer sur des caractéristiques d'une population finie U, par exemple sur un total. On parlera dans la suite d'*inférence descriptive*. Tous les éléments de la population étaient d'intérêt pour les inférences. Un échantillon s était sélectionné dans U selon un plan de sondage probabiliste $p(\cdot)$ donnant à s une probabilité $p(s)$ d'être sélectionné. Les propriétés de l'estimateur étaient évaluées par rapport au plan de sondage : 1) les valeurs du caractère d'intérêt $(y_k, k \in U)$ étaient traitées comme étant fixes et donc θ était non aléatoire. On faisait alors face à un problème d'estimation ; 2) les indicatrices d'inclusion $(I_k, k \in U)$ étaient aléatoires.

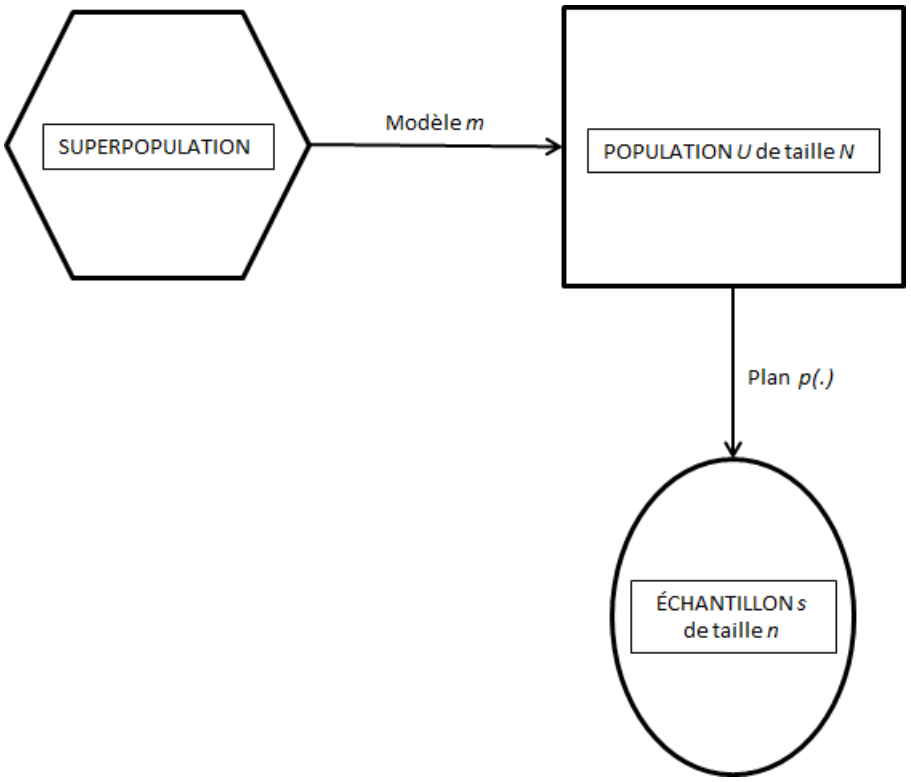
Dans l'approche basée sur le modèle, on cherche à inférer sur des caractéristiques d'une population finie U. Tous les éléments de la population sont d'intérêt pour les inférences. Dans cette approche, l'échantillon s peut être vu comme le résultat d'un

processus en deux étapes :

- Première étape : la population finie U de taille N est générée à partir d'une population infinie (souvent appelée *superpopulation*) selon un modèle m (similaire au modèle en statistique classique c'est-à-dire un ensemble d'hypothèses qui décrit la distribution des données d'une population infinie). Le vecteur $\mathbf{y}_N$ est ainsi généré, Y est donc une variable aléatoire (ou si l'on veut, les Y_k). On est alors face à un problème de prédiction. Autrement vu, la population U est un échantillon de taille N tirée (avec remise) dans la superpopulation.
- Deuxième étape : de la population U, un échantillon s de taille n est tiré selon un plan de sondage $p(\cdot)$. Le vecteur des indicatrices $\mathbf{I} = (I_1, I_2, \cdots, I_N)$ est ainsi généré et on travaille à partir de cet unique échantillon (les indicatrices ne bougent plus).

La figure 6 résume les deux approches.

FIGURE 6 – Deux approches pour l'inférence.



Il est donc supposé que, conditionnellement à une matrice de variables explicatives, $\mathbf{X}$, les Y_k sont des variables aléatoires. Le modèle le plus fréquemment considéré pour générer les Y_k est du type linéaire Gaussien :

$$
\begin{aligned}
&Y_k = \mathbf{X}_k^\top \beta + \epsilon_k \\
&E(\epsilon_k \mid \mathbf{X}_k) = 0 \\
&V(\epsilon_k \mid \mathbf{X}_k) = \sigma_k^2 \propto v_k \\
&\epsilon_k \text{ suit une loi normale}
\end{aligned}
$$

On note le plan de sondage $p(\cdot \mid \mathbf{Z})$ où $\mathbf{Z}$ est l'information du plan, par exemple l'identification des strates.
Le plan de sondage est dit non *informatif* ou *ignorable* pour Y si $p(\cdot \mid \mathbf{X}, Y) = p(\cdot \mid \mathbf{X})$ où $\mathbf{X}$ est une matrice de prédicteurs. Autrement dit le plan n'est pas informatif si l'information du plan $\mathbf{Z}$ est incluse dans $\mathbf{X}$ [19]. Pour satisfaire cette hypothèse il n'est pas nécessaire d'inclure toute l'information du plan dans $\mathbf{X}$. On inclut seulement l'information qui explique Y. Autrement dit, le modèle qui est valable pour la population l'est aussi pour l'échantillon. Ou bien encore : le plan de sondage n'a pas brisé la relation entre Y et $\mathbf{X}$. Dans l'approche modèle, l'inférence pour Y sera valide si le plan de sondage est non-informatif.
Le caractère informatif d'un plan de sondage n'est pas spécialement lié au contexte de l'approche modèle. D'une manière plus générale, un plan de sondage est informatif si la distribution des valeurs du caractère Y dans l'échantillon est différente de la distribution des valeurs du caractère Y dans la population. L'intégration des poids de sondage dans l'estimation permet alors d'effacer ce problème et de rendre l'inférence valide c'est-à-dire de pouvoir inférencer sur les éléments de la population qui ne figurent pas dans l'échantillon. Dans le contexte d'un modèle, un plan ignorable fera en sorte que la distribution de $Y \mid \mathbf{X}$ estimée à partir de l'échantillon sera valable pour la population.

Si $\hat{\theta}$ est un certain prédicteur de θ, l'*erreur de prédiction* est par définition $\hat{\theta} - \theta$. Les propriétés des prédicteurs sont évaluées par rapport au modèle m qui génère les Y_k et les indicatrices d'inclusion dans l'échantillon s ne bougent pas. Le biais par rapport à m peut être alors défini dans ce cadre par :

$$E_m[(\hat{\theta} - \theta) \mid s]$$

dans lequel θ figure dans l'espérance car θ est une variable aléatoire.
Le prédicteur $\hat{\theta}$ sera dit sans biais par rapport au modèle m si, étant donné s on a :

$$E_m[(\hat{\theta} - \theta) \mid s] = 0$$

De même une variance et une erreur quadratique moyenne sont définies respectivement par :

$$V_m[(\hat{\theta} - \theta) \mid s] \text{ et } E_m[(\hat{\theta} - \theta)^2 \mid s]$$

Ces moments s'interprètent de la manière suivante :

Les Y_k sont générées suivant un certain modèle m. En pratique on n'observe qu'une seule réalisation du modèle, $\mathbf{Y}^{(1)}$, et on fait l'inférence à partir de cette seule réalisation. On obtient l'erreur de prédiction $\hat{\theta}^{(1)} - \theta^{(1)}$ qui n'est pas observable puisque $\theta^{(1)}$ est inconnu.
En théorie on pourrait générer les $Y_k, k \in U$, un très grand nombre de fois indépendamment selon le même modèle m, disons R fois. On obtiendrait alors R erreurs de prédictions, $\hat{\theta}^{(1)} - \theta^{(1)}, \ldots, \hat{\theta}^{(R)} - \theta^{(R)}$ où l'échantillon s resterait le même à chaque répétition.

19. Ce qui peut se voir aussi à travers les variables indicatrices : $\Pr(I_i = 1 \mid x_i, y_i) = \Pr(I_i = 1 \mid x_i)$.

Dès lors le biais, $E_m[(\hat{\theta} - \theta) \mid s]$ se comprend simplement comme la moyenne des erreurs de prédiction quand $R \to \infty$. La variance et l'EQM s'interprètent de la même manière.

C'est dans ce contexte d'approche modèle que se situe la modélisation multi-niveaux sur données d'enquête décrite dans la section qui suit.

6 Modélisation à un niveau et modélisation multi-niveaux

6.1 Inférence dans l'approche modèle

Dans le cadre de l'inférence descriptive faite selon le plan de sondage, l'incertitude sur l'estimation du paramètre $\theta = f(y_1, \ldots, y_N)$ vient du fait que l'on ne dispose que d'un sous ensemble d'éléments de U qui est de taille N, c'est-à-dire que l'on dispose de l'échantillon s de taille n : $\hat{\theta} = f(y_1, \ldots, y_n)$. L'inférence est uniquement valide pour la population U et n'est en rien généralisable à d'autres populations.
Si tous les éléments de U sont échantillonnés, c'est-à-dire $s = U$, autrement dit la situation est un recensement, il ne peut y avoir d'incertitude sur l'estimation de θ. Autrement dit la variance de $\hat{\theta}$ est nulle.

Supposons que l'on s'intéresse à la relation entre un caractère Y et d'autres caractères $X_1, \ldots, X_p$, c'est-à-dire à un processus qui a généré les données que l'on a observées. L'idée est de dire que, puisque le modèle représente le processus qui a généré les données, il est possible d'établir des conclusions généralisables à d'autres situations où le même processus a opéré. Mais puisque le modèle ne peut être qu'une approximation, il est important (mais difficile) de connaître les écarts au modèle qui feront que l'analyse n'est pas valide.
Si on modélise les données observées sur une population entière (recensement), par exemple par un modèle de régression linéaire, on s'attend à ce que chaque coefficient de régression estimé ait une variance associée (non nulle) c'est-à-dire une incertitude puisque le modèle n'est qu'une approximation. Or il ne peut y avoir d'incertitude liée au plan de sondage puisque les données sont exhaustives. La variance provient d'autre part. Elle provient essentiellement de l'inférence entre la population et ce qui a été appelée la superpopulation. Cette situation tombe donc dans le cadre de l'inférence de l'approche modèle.

Pour résumer et comprendre facilement on peut écrire grossièrement [20] :

20. C'est en effet très grossier pour ne pas dire inexact. Il existe des estimateurs *« design-based »* des paramètres de modèles. Sur ce sujet on pourra se reporter à [Binder & Roberts, 2003]. Il existe de plus une approche hybride appelée approche assistée par un modèle (*model-assisted*).
Il s'agit ici de facilité la compréhension des choses sans entrer pour autant dans des considérations plus théoriques et exhaustives.

- θ est un coefficient de régression : approche modèle
- θ est un total, une moyenne etc. : approche plan

6.2 Modélisation à un niveau

Dans l'exemple d'une régression linéaire, un seul niveau d'information intervient. Par exemple, on peut tenter d'ajuster la concentration en plomb dans l'eau du robinet d'un logement en fonction de la température de l'eau et de la longueur des canalisations en plomb allant du robinet jusqu'au branchement au réseau public. Ces informations dans la table de données se présenteraient sur une même ligne pour un logement donné (un niveau).
Si ces données sont connues dans la population entière de logements, et si on suppose qu'il existe une relation linéaire entre les concentrations en plomb mesurées dans l'eau y_k, et la température de l'eau ainsi que la longueur des canalisations, notées matriciellement par $\mathbf{x}_k$, on serait amener classiquement à chercher le vecteur des coefficients de régression $\mathbf{b} \in \mathbb{R}^2$ qui minimisent le critère :

$$\sum_{k \in U} (y_k - \mathbf{x}_k^\top \mathbf{b})^2 \tag{6.1}$$

En dérivant par rapport à $\mathbf{b}$, on trouve l'équation estimante :

$$\sum_{k \in U} \mathbf{x}_k (y_k - \mathbf{x}_k^\top \mathbf{b}) = 0$$

et ce qui donne

$$\sum_{k \in U} \mathbf{x}_k y_k = \sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^\top \mathbf{b}$$

En notant $\mathbf{T} = \sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^\top$ et $\mathbf{t} = \sum_{k \in U} \mathbf{x}_k y_k$ et en supposant que $\mathbf{T}$ soit inversible on obtient le vecteur des coefficients de régression :

$$\mathbf{b} = \mathbf{T}^{-1} \mathbf{t}$$

qui peut être considéré comme un prédicteur des pentes dans la superpopulation. $\mathbf{T}$ et $\mathbf{t}$ sont des totaux et peuvent donc être aisément estimés par leur π-estimateur respectif si on ne dispose que d'un échantillon s :

$$\widehat{\mathbf{T}}_\pi = \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k}$$

et

$$\hat{\mathbf{t}}_\pi = \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\pi_k}$$

qui sont des estimateurs sans biais de $\mathbf{T}$ et $\mathbf{t}$. Enfin, $\mathbf{b}$ est estimé par :

$$\hat{\mathbf{b}} = \widehat{\mathbf{T}}_\pi^{-1} \hat{\mathbf{t}}_\pi$$

sous réserve que $\widehat{\mathbf{T}}$ soit inversible.

Le problème de construire un estimateur des coefficients de régression s'est donc résumé à pondérer chaque observation k par son poids de sondage $w_k = 1/\pi_k$.

Calculer la variance de cet estimateur est un problème moins simple dans la mesure où cet estimateur s'écrivant sous la forme d'un ratio, n'est pas une fonction linéaire des valeurs prises par les caractères. Une formule simple pour la variance d'un ratio ne peut pas être fournie même si les variances du numérateur et du dénominateur sont des expressions bien connues puisqu'elles sont linéaires. L'idée va être de calculer la variance d'une fonction linéaire approximant l'estimateur dont l'expression est non linéaire et donc trop compliquée.
Plus généralement, on considère un paramètre d'intérêt θ (par exemple comme ci-dessus $\theta = \mathbf{b}$) étant une fonction de totaux :

$$\theta = f(t_1, \ldots, t_j, \ldots, t_q)$$

où

$$t_j = \sum_U y_{jk}$$

c'est-à-dire le total des valeurs du caractère Y_j, dont y_{jk} est le k-ème élément. Le principe est alors assez simple : chaque total t_j inconnu est remplacé par son π-estimateur $\hat{t}_{j\pi} = \sum_s y_{jk}/\pi_k$ dans $f(\cdot)$ donnant l'estimateur suivant :

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{j\pi}, \ldots, \hat{t}_{q\pi})$$

Si f est non linéaire, une approximation de $\hat{\theta}$ est obtenue par *linéarisation de Taylor* donnant une expression linéaire à partir de laquelle une variance peut être facilement calculée. Pour les détails de ce procédé de linéarisation, on peut se référer à la section 5.5 de [Särndal et al., 1992] ou au chapitre 12 de [Tillé, 2001] ou encore à l'annexe A.2 de [Lumley, 2010a].

6.3 Modélisation multi-niveaux

Lorsque l'on souhaite étudier la relation entre des variables dans le même esprit qu'avec une régression linéaire, et que ces variables donnent l'information sur des unités de niveaux différents, il est possible d'utiliser une *modélisation* dite *multi-niveaux*[21]. Par exemple il peut s'agir de pièces échantillonnées à l'intérieur de logements tirés au sort préalablement : un premier niveau sera constitué de pièces et un second niveau sera constitué de logements. De nouveau, l'analyse de données issues de plans de sondage complexes a amené à développer des outils spécifiques pour ce genre de modélisation depuis les années 1990.

En données d'enquête, l'utilisation d'un modèle multi-niveaux se calque généralement sur un plan de sondage à plusieurs degrés, les degrés du plan de sondage étant similaires à ce qui est entendu ici par *niveaux* du modèle. Cependant, aussi

21. En statistique classique, ce genre de modélisation est aussi connue sous le nom de modélisation hiérarchique ou de modélisation mixte.

bizarre que cela puisse paraître, le jargon de la modélisation multi-niveaux inverse la numérotation des niveaux par rapport à celle des degrés. Dans l'exemple des logements et des pièces, les unités de niveau 1 du modèle seraient alors les pièces constituant le degré 2 du plan, et les logements seraient les unités de niveau 2 du modèle, c'est-à-dire les unités du premier degré du plan. Ainsi les unités du niveau 1 du modèle sont les dernières unités qui ont été échantillonnées suivant le plan de sondage.

En régression linéaire classique rappelée dans la section précédente, les erreurs du modèle sont supposées indépendantes. Dans l'exemple de données issues de plan de sondage à plusieurs degrés, cette hypothèse n'est pas réaliste de part la structure même de l'échantillonnage. Pour reprendre de nouveau l'exemple des logements et des pièces, on peut supposer que les caractéristiques des pièces issues d'un même logement soient corrélées entre elles induisant par la même une certaine corrélation entre les erreurs. En statistique classique cette corrélation se voit par exemple sur des données longitudinales où par définition des mesures (une plombémie ou une tension artérielle par exemple) seront répétées à plusieurs intervalles de temps sur le même individu.
Pour prendre en compte cette dépendance due à la hiérarchie il est possible de décomposer la variance en introduisant un effet dit aléatoire. *A contrario*, les effets sont dits fixes comme en régression linéaire standard, c'est-à-dire que l'effet d'un régresseur est vu comme un effet moyen à travers les observations. Dans la suite on se limitera au cas où seul l'*« intercept*[22] *»* est aléatoire : au lieu que la variable réponse Y des observations soit construite à partir d'une moyenne globale β_0 à laquelle est ajoutée la valeur des régresseurs X multipliée par un effet moyen fixe ψ (figure 7), la variable réponse des observations est construite à partir d'une moyenne β_{0j} propre à chaque groupe d'observations à laquelle est ajoutée la valeur des régresseurs multipliée par un effet moyen fixe (figure 8, ϵ note les erreurs). On se limitera dans notre cas à une modélisation à 2 ou à 3 niveaux pour traiter les données d'enquête utilisées dans le présent travail.

En supposant que les données proviennent d'un plan de sondage à 3 degrés (par exemple « région/logement/pièce » ≡ « niveau 3/niveau 2/niveau 1 »), un modèle à 3 niveaux à *« intercept »* aléatoire se décrit par les hypothèses et les relations suivantes :

$$\text{Niveau 1} : y_{ijk} = \beta_{0jk} + \sum_{m=1}^{q_1} \varphi_m x_{ijk}^{(m)} + \epsilon_{ijk} \text{ où } i = 1, \ldots, n_{jk}^{(1)} \tag{6.2}$$

$$\text{Niveau 2} : \beta_{0jk} = \beta_{0k} + \sum_{r=1}^{q_2} \psi_r x_{jk}^{(r)} + \xi_{jk} \text{ où } j = 1, \ldots, n_k^{(2)} \tag{6.3}$$

$$\text{Niveau 3} : \beta_{0k} = \beta_0 + \sum_{p=1}^{q_3} \delta_p x_k^{(p)} + \zeta_k \text{ où } k = 1, \ldots, n^{(3)} \tag{6.4}$$

avec $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_1^2)$, $\xi_{jk} \sim \mathcal{N}(0, \sigma_2^2)$ et $\zeta_k \sim \mathcal{N}(0, \sigma_3^2)$.
β_0 est la moyenne globale, ζ_k correspond aux effets aléatoires sur les unités de ni-

22. L'ordonnée à l'origine.

veau 3, ayant pour moyenne 0 et variance σ_3^2 représentant la dispersion autour de la quantité moyenne β_0. Les effets ζ_k sont supposés non corrélés entre les unités du niveau 3.
ξ_{jk} correspond aux effets aléatoires sur les unités de niveau 2, de moyenne 0 et variance σ_2^2 représentant la dispersion à l'intérieur d'une unité de niveau 3 autour de la quantité aléatoire moyenne β_{0k}. Leur variance est supposée constante entre les unités de niveau 3. Les ξ_{jk} sont supposés non corrélés entre les unités de niveau 2 et 3 et non corrélés aux covariables X.
Les ϵ_{ijk} sont les perturbations de moyenne 0 et de variance σ_1^2 représentant la dispersion de la variable réponse Y à l'intérieur d'une unité de niveau 2 et supposée constante entre les unités de niveau 2. Les ϵ_{ijk} sont supposés non corrélés entre les unités de niveau 1, 2 et 3 et non corrélés aux covariables.
On suppose de plus que les effets aléatoires ζ_k, ξ_{jk} et ϵ_{ijk} sont non corrélés.
φ_m, ψ_r et δ_p sont les coefficients associés aux covariables des niveaux 1, 2 et 3 respectivement et représentent les effets fixes.

Un modèle à 2 niveaux à *« intercept »* aléatoire est défini par deux équations sous des hypothèses analogues :

$$\text{Niveau 1} : y_{ij} = \beta_{0j} + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)} + \epsilon_{ij} \tag{6.5}$$

$$\text{Niveau 2} : \beta_{0j} = \beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \zeta_j \tag{6.6}$$

avec $j = 1, \ldots, n^{(2)}$ et $i = 1, \ldots, n_j^{(1)}$.

L'introduction d'un effet aléatoire pour l'*« intercept »* permet de modéliser la variabilité entre les unités de niveau 2. En effet on a dans le cadre d'un modèle à 2 niveaux :

$$\begin{aligned} V(Y_{ij}) &= V(\beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \zeta_j + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)} + \epsilon_{ij}) \\ &= V(\beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)}) + V(\zeta_j) + V(\epsilon_{ij}) \\ &= V(\zeta_j) + V(\epsilon_{ij}) \\ &= \sigma_2^2 + \sigma_1^2 \end{aligned}$$

étant donné que ζ_j et ϵ_{ij} sont indépendants par hypothèse et que les $X_j^{(r)}$ et $X_{ij}^{(m)}$ ne sont pas considérées comme aléatoires. Dès lors la proportion de variance expliquée par le niveau le plus haut i.e. le niveau 2 et appelée coefficient de partitionnement de la variance (VPC) est égal à

$$VPC = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}$$

Pour un modèle à *« intercept »* aléatoire cette quantité se confond avec le coefficient de corrélation intra-classe (ICC) souvent noté ρ, qui mesure la corrélation entre

FIGURE 7 – Schématisation d'un modèle à 1 niveau à effets fixes.

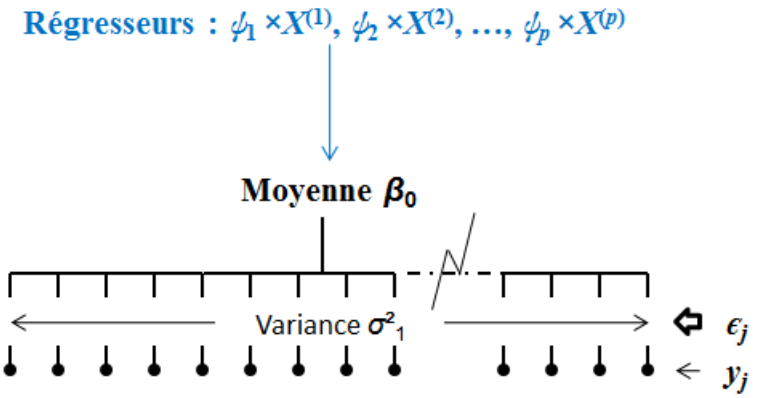
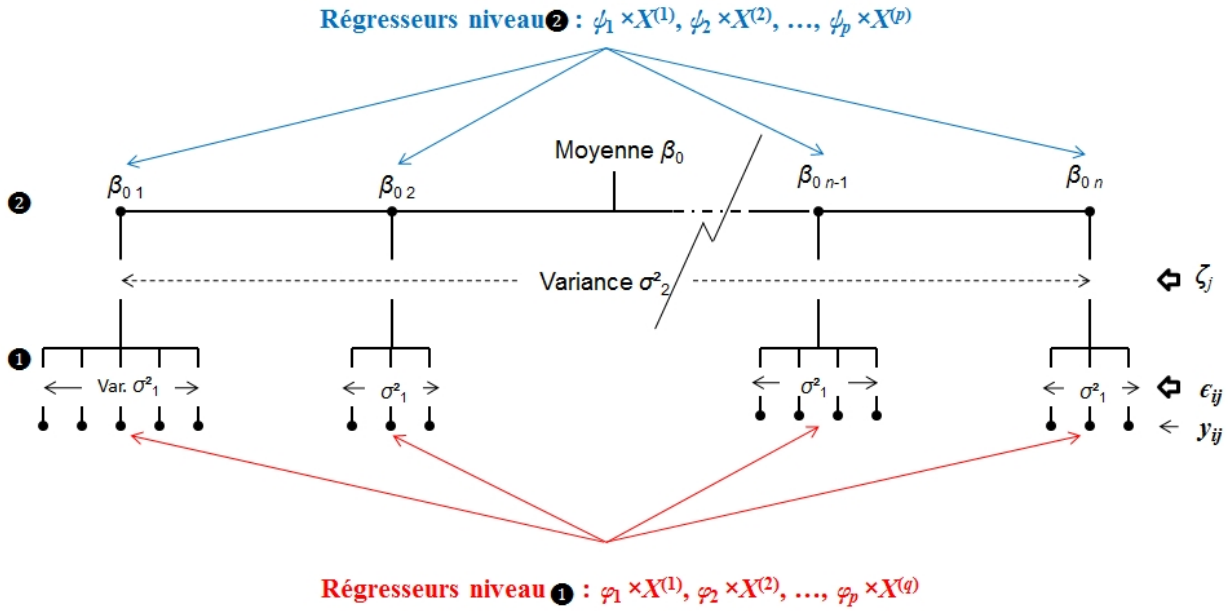


FIGURE 8 – Schématisation d'un modèle à 2 niveaux à *« intercept »* aléatoire.



deux unités de niveau 1 se trouvant dans la même unité de niveau 2.

Dans le cadre générale il existe différentes méthodes pour estimer les effets dans un modèle de régression. Par exemple en régression linéaire standard, on peut utiliser l'estimateur des moindres carrés comme dans l'équation 6.1. Cet estimateur se confond ici avec un autre estimateur, l'estimateur du maximum de vraisemblance (ML). Dans le cas d'un modèle de régression linéaire on fait classiquement les hypothèses suivantes :

$$y_k = \mathbf{x}_k^\top \mathbf{b} + \epsilon_k \quad , \epsilon_k \sim \mathcal{N}(0, \sigma^2)$$

de sorte que $y_k \sim \mathcal{N}(\mathbf{x}_k^\top \mathbf{b}, \sigma^2)$, avec les ϵ_k indépendants. Le paramètre d'intérêt à estimer est alors le vecteur θ composé des effets fixes et de la variance σ^2, $\theta = (\mathbf{b}^\top, \sigma^2)^\top$. Le principe consiste à maximiser une expression appelée vraisemblance de l'échantillon composé des y_k, quantifiant la vraisemblance d'observer un tel échantillon si les valeurs qui le composent sont issues de la loi supposée. Par indépendance des y_k

due à l'indépendance des ϵ_k, la vraisemblance est définie par :

$$L(\theta) = \prod_k f(y_k; \theta)$$

où $f(\cdot)$ est la fonction de densité de la loi de Laplace-Gauss :

$$f(y_k; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_k - \mathbf{x}_k^\top \mathbf{b})^2}{2\sigma^2}\right)$$

Pour des raisons algébriques il est plus commode de maximiser le logarithme de la vraisemblance :

$$\log(L(\theta)) = \sum_k \log(f(y_k; \theta)) \tag{6.7}$$

L'estimateur ML est l'expression qui permet de maximiser $\log(L(\theta))$.

Pour ajuster un tel modèle sur données d'enquête, les poids de sondage associés à chaque élément y_k de l'échantillon s peuvent être introduits dans la vraisemblance, produisant alors ce qui est appelée *pseudo-vraisemblance* [Skinner, 1989], et conduisent à *l'estimateur du pseudo maximum de vraisemblance* (PML). Ceci se base sur l'idée classique consistant à dire que la sélection de l'échantillon ne produira pas de biais si les valeurs sur toute la population avaient été observées, comme dans un recensement [Pfeffermann et al., 1998].

Lorsque les unités qui composent un échantillon s issu d'un plan de sondage, ont été sélectionnées à partir de probabilités d'inclusion inégales, il a été montré que ne pas tenir compte des ces probabilités dans l'ajustement d'un modèle multi-niveaux produisait des estimateurs biaisés [Pfeffermann et al., 1998]. L'inférence n'est pas valide car le plan est informatif. Pour contrer ce problème, Pfeffermann et al. [Pfeffermann et al., 1998] proposent de pondérer les unités de chacun des niveaux. Cependant la pondération en modélisation multi-niveaux n'est pas une extension triviale de la pondération en modélisation à un seul niveau comme il vient d'être vu en régression linéaire. Il y a deux raisons à cela. La première est que les valeurs dans la population finie ne sont pas indépendantes dans de tels modèles et donc la log-vraisemblance, vue dans le cadre d'un recensement, ne peut pas être une simple somme (comme l'équation 6.7) réalisée sur les éléments de la population finie. La seconde raison est que les poids de sondage des unités tirées en dernier par le plan de sondage ne comportent pas suffisamment d'information pour corriger le biais, contrairement à ce qui se passe en régression sur un seul niveau.

Pour palier à ce problème il a alors été proposé d'associer à chaque unité un poids égal à l'inverse de sa probabilité d'inclusion conditionnelle (cf. section 3.4 de cette partie). On nommera un tel poids un *poids conditionnel*. Dans le cadre d'une modélisation multi-niveaux à 2 niveaux cela donne :

$$w_{i|j}^{(1)} = 1/\pi_{i|j}$$
$$w_j^{(2)} = 1/\pi_j$$

où i indexe les unités du niveau 1 indiqué par $^{(1)}$ et j indexe celles du niveau 2 indiqué par $^{(2)}$.

La log-pseudo vraisemblance d'un modèle à 2 niveaux a alors pour expression [Rabe-Hesketh, 2007] :

$$\sum_{j=1}^{n^{(2)}} w_j^{(2)} \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j}^{(1)} \log \left(f(y_{ij} \mid \zeta_j) \right) \right\} g(\zeta_j) \mathrm{d}\zeta_j \tag{6.8}$$

La somme $\sum_{i=1}^{n_j^{(1)}} w_{i|j}^{(1)} \log \left(f(y_{ij} \mid \zeta_j) \right)$ est proche dans sa structure de l'expression dans l'équation 6.7. Elle représente la contribution à la log-vraisemblance des unités de niveau 1 conditionnellement à l'existence de l'effet aléatoire ζ_j au niveau 2. $g(\zeta_j)$ est la densité d'une loi de Laplace-Gauss de l'effet aléatoire ζ_j. $f(y_{ij} \mid \zeta_j)$ est une notation simplifiée de ce qui devrait être $f(y_{ij} \mid \zeta_j, \mathbf{b}, \sigma_1^2)$ et $g(\zeta_j)$ est une notation simplifié de $g(\zeta_j \mid \sigma_2^2)$, où $\mathbf{b}$ est le vecteur des coefficients de régression. Pour un modèle à 3 niveaux ou plus l'expression de la log-pseudo vraisemblance s'écrit dans la même logique mais devient vite compliquée à formaliser et nécessite une écriture par récurrence. On pourra se reporter à la section 4.2 de [Rabe-Hesketh & Skrondal, 2006] pour disposer d'une telle expression. Les variances des estimateurs des effets fixes $\mathbf{b}$ et des estimateurs des paramètres de variance σ_1^2 et σ_2^2 sont obtenues par linéarisation de Taylor [Rabe-Hesketh & Skrondal, 2006].

De plus il a été proposé, dans le cadre d'un modèle à 2 niveaux, de réduire le biais des estimateurs pouvant encore exister malgré l'introduction de poids conditionnels, en corrigeant les poids des unités de niveau 1 par ce qui est appelée une *mise à l'échelle*. Deux méthodes de mise à l'échelle ont été proposées dans [Pfeffermann et al., 1998].
La première consiste à faire en sorte que la somme des nouveaux poids des unités du niveau 1 égale la taille dans l'échantillon de l'unité du niveau 2.
La seconde consiste à faire en sorte que la somme des nouveaux poids des unités du niveau 1 égale la taille dans la population de l'unité du niveau 2.
Ceci a été étudié par une étude de simulation mais uniquement dans le cadre d'un modèle à 2 niveaux. Cependant même après ces corrections il ressort que si le plan de sondage est informatif au niveau 1 (cf. section 5), un biais peut persister dès lors que le nombre d'unités de niveau 1 est petit à l'intérieur de chaque unité de niveau 2.

Enfin, selon que l'on soit plutôt intéressé par les coefficients de régression ou par les paramètres de variance, on peut être amené à utiliser une méthode de mise à l'échelle plutôt qu'une autre [Carle, 2009].

7 Problématique des données manquantes

Cette section traite de la problématique des données manquantes et de leur effet délétère : l'incapacité d'obtenir des réponses utilisables peut affecter grandement la

qualité des estimations produites.

Dans cette section l'élément de bibliographie utilisé a été [Haziza, 2012].

7.1 Les sources d'erreur

Si on considère un paramètre d'intérêt θ et $\hat{\theta}$ son estimateur, *l'erreur totale* de $\hat{\theta}$ est définie par $\hat{\theta} - \theta$. Cette erreur totale est composée de *l'erreur due à l'échantillonnage* et de *l'erreur non due à l'échantillonnage*.

Les erreurs d'échantillonnage proviennent du fait que l'information désirée n'est observée que pour une partie de la population c'est-à-dire à travers un seul échantillon. Les valeurs non observées sont donc des données manquantes mais elles le sont de façon volontaire et ont été planifiées.

Les erreurs non dues à l'échantillonnage sont de trois natures :

- Les erreurs de couverture qui proviennent du fait que la base de sondage et la population ne coïncident pas.
- Les erreurs de mesure qui sont dues au fait que les valeurs obtenues sont différentes des vraies valeurs.
- Les erreurs de non-réponse qui proviennent du fait que l'information désirée n'a été observée que pour une partie de l'échantillon.

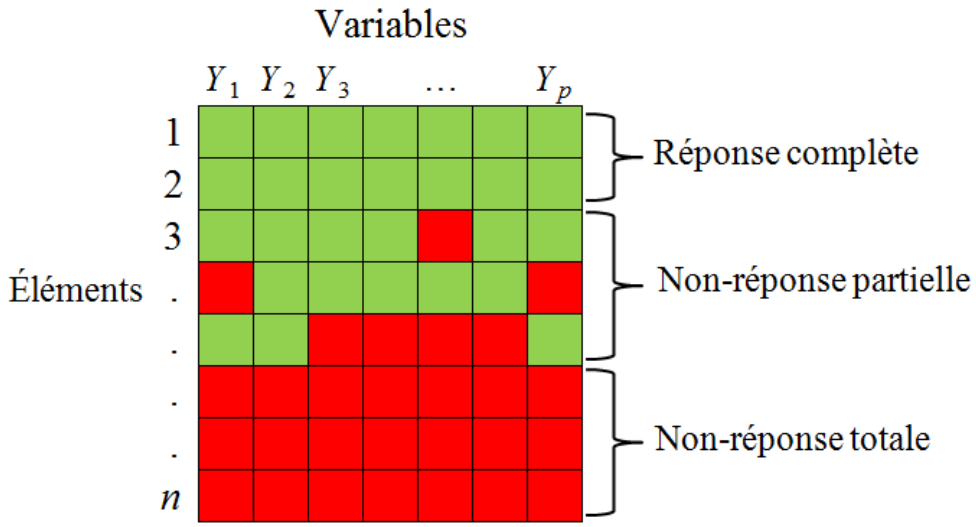
7.2 La non-réponse

On considère dans la suite que les erreurs de couverture et les erreurs de mesures sont négligeables (bien que ce ne soit vraisemblablement pas toujours le cas en pratique).

La présence de *non-réponses* si on parle de questionnaire, ou simplement de *données manquantes* dans un cadre plus général, est inhérente à la collecte de données. La prévalence d'au moins une donnée manquante augmente avec le volume d'informations à collecter.

On distingue deux types de non-réponse : la *non-réponse partielle* et la *non-réponse totale*. La non-réponse totale est un refus de participer (qui peut n'être en fait qu'un oubli) et induit qu'aucune information souhaitée ne peut être obtenue sur l'individu non répondant. La non-réponse partielle est un refus (ou une omission) de répondre à certaines des questions seulement. La figure 9 illustre le types de non-réponses.
La non-réponse a pour effet de produire un *biais de non-réponse* pour le paramètre d'intérêt, qui est du au fait que la distribution du caractère auquel on s'intéresse est différente chez les répondants et les non-répondants. Le biais de non-réponse s'interprète comme la différence entre l'estimation et la vraie valeur du paramètre

FIGURE 9 – Les types de non-réponses.



si l'échantillonnage, la non-réponse et le traitement de la non-réponse sont répétés un grand nombre de fois. Dans la pratique on ne peut pas savoir s'il y a un biais et le cas échéant on ne peut connaître son ampleur. La variance des estimateurs est généralement plus grande que celle des estimateurs que l'on aurait obtenu en l'absence de non-réponse. En l'absence de non-réponse, habituellement les plans de sondage et les estimateurs sont choisis de façon qu'il n'y ait pas de biais comme illustré en figure 10. La figure montre la moyenne (droite rouge) des pourcentages estimés de fumeurs par le π-estimateur, à partir de 1000 échantillons tirés de manière aléatoire simple dans une population de 100 000 individus. La droite bleu indique la proportion réelle de fumeur (égale à 27,4 %) ; n indique la taille de l'échantillon. Le biais est quasi nul quelque soit la taille de l'échantillon, n, et la variance diminue quand n augmente.

Pour réduire le biais de non-réponse un des éléments clés est l'utilisation de l'information auxiliaire disponible. Pour que cette réduction soit efficace il faut choisir les variables auxiliaires autant que possible liées aux caractères auxquels on s'intéresse et autant que possible liées à la probabilité de réponse.

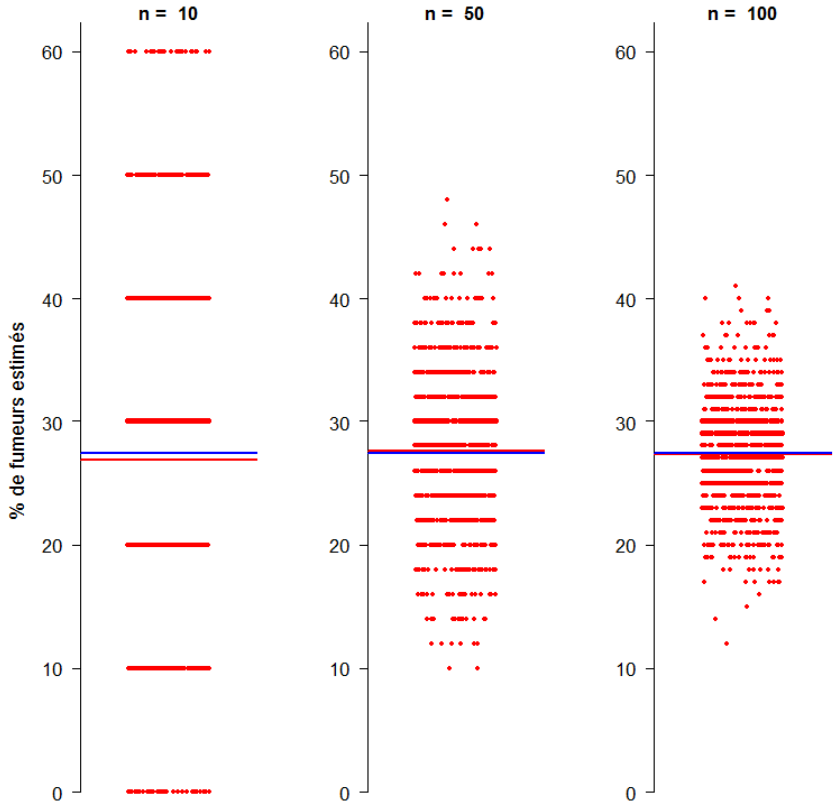
7.3 Mécanisme de non-réponse

Si on note par R_k la variable indicatrice de réponse valant 1 ($r_k = 1$) si l'élément k a répondu à un item ou une question donnée Y, le *mécanisme de non-réponse* peut être décrit par :

$$R_k \overset{\text{ind.}}{\sim} \mathcal{B}(1, 1 - p_k), \quad k \in U$$

Puisque p_k, la probabilité de non-réponse, est inconnue la distribution est inconnue. L'hypothèse d'indépendance entre les éléments k est habituellement satisfaite sauf dans certains cas, par exemple si on considère les personnes composant un même ménage. On parle de mécanisme car les causes de la non-réponse sont nombreuses. On peut considérer 3 types de mécanisme de non-réponse :

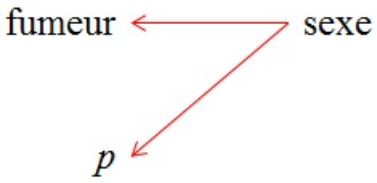
FIGURE 10 – Illustration du biais et de la variance d'un estimateur en absence de donnée manquante.



- *Uniforme* ou *complètement aléatoire* (*missing completely at random* : MCAR).
- *Ignorable* ou *aléatoire* (*missing at random* : MAR).
- *Non-ignorable* ou *confondu* (*missing not at random* : MNAR).

Le mécanisme de non-réponse est *uniforme* si $p_k = p$ pour tout élément de la population. Cela signifie que la probabilité de non réponse ne dépend d'aucune information, ni du caractère d'intérêt, ni de variable auxiliaire. Un tel mécanisme induit alors une absence de biais comme illustré en figure 11 traitant les mêmes données que la figure 10 : seule la variance diminue lorsque n augmente ou lorsque $1 - p$ augmente. En pratique un mécanisme uniforme est irréaliste mais il peut être plausible à l'intérieur de classes.

Le mécanisme de non-réponse est *ignorable* si $\Pr(r_k = 1|\mathbf{y}, \mathbf{z}) = \Pr(r_k = 1|\mathbf{z})$ où $\mathbf{z}$ sont des informations auxiliaires. Autrement dit après avoir pris en compte l'information auxiliaire adéquate, la probabilité de réponse ne dépend plus du caractère d'intérêt. Considérons de nouveau l'exemple de la proportion de fumeurs à estimer de la figure 11, et supposons que la probabilité de non-réponse est différente chez les femmes et chez les hommes. On a un mécanisme s'illustrant alors de la manière suivante :



Il y a un lien indirecte entre la probabilité de répondre ou non (p) et le fait d'être fumeur. Sur la figure 12 l'information du sexe n'a pas été prise en compte au niveau de l'estimation. La moyenne des répondants est biaisée et le biais augmente lorsque la probabilité de non-réponse augmente : les données demeurent non-ignorables. Sur la figure 13 l'estimateur, $\widehat{P}$, de la proportion de fumeurs a été ajustée sur le sexe (H,F) par :

$$\widehat{P}_{adj} = \frac{n_F}{n}\widehat{P}_F + \frac{n_H}{n}\widehat{P}_H$$

Après avoir pris en compte le sexe il n'y a virtuellement plus de biais car le lien indirect entre le fait de fumer et le fait de répondre ou de ne pas répondre a été éliminé : le mécanisme est ignorable.

FIGURE 11 – Illustration de l'absence de biais lorsque le mécanisme de non-réponse est uniforme.

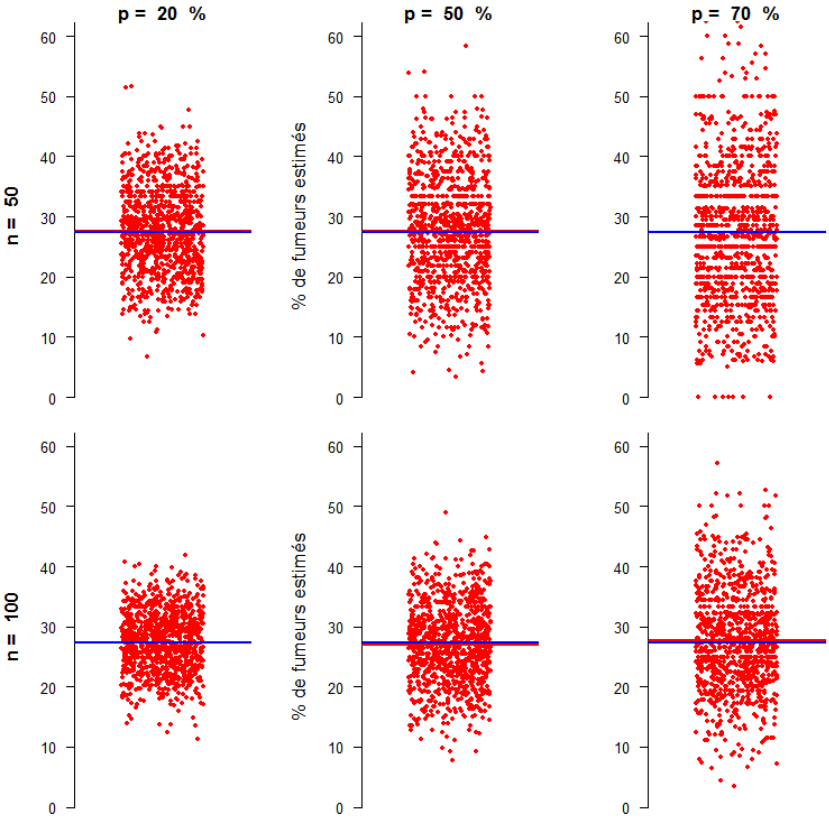
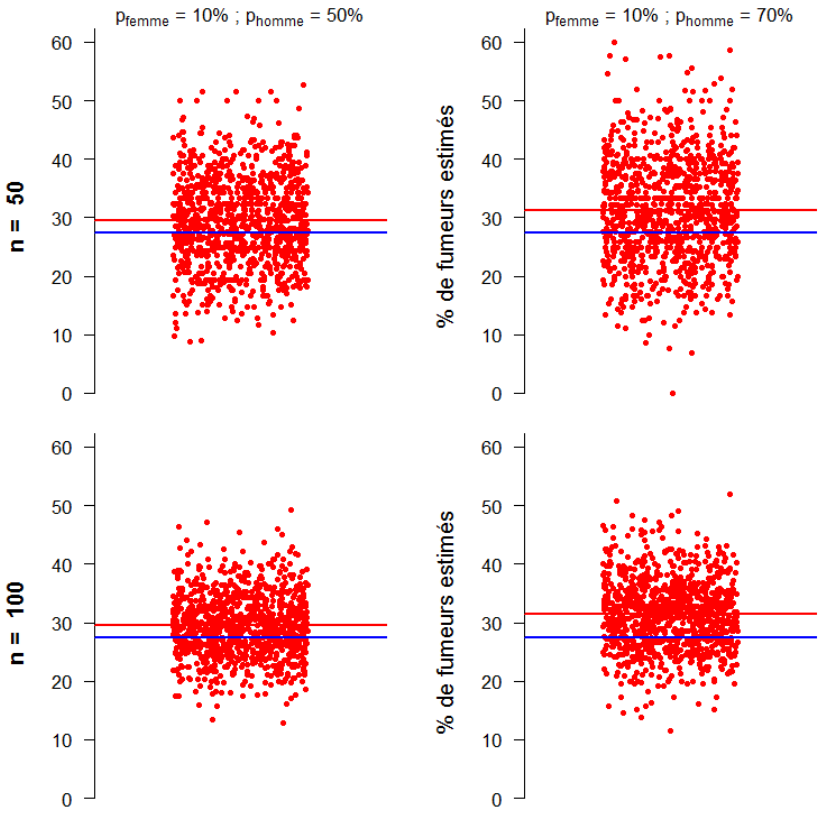


FIGURE 12 – Illustration de la présence de biais lorsque le mécanisme de non-réponse est aléatoire et non traité.



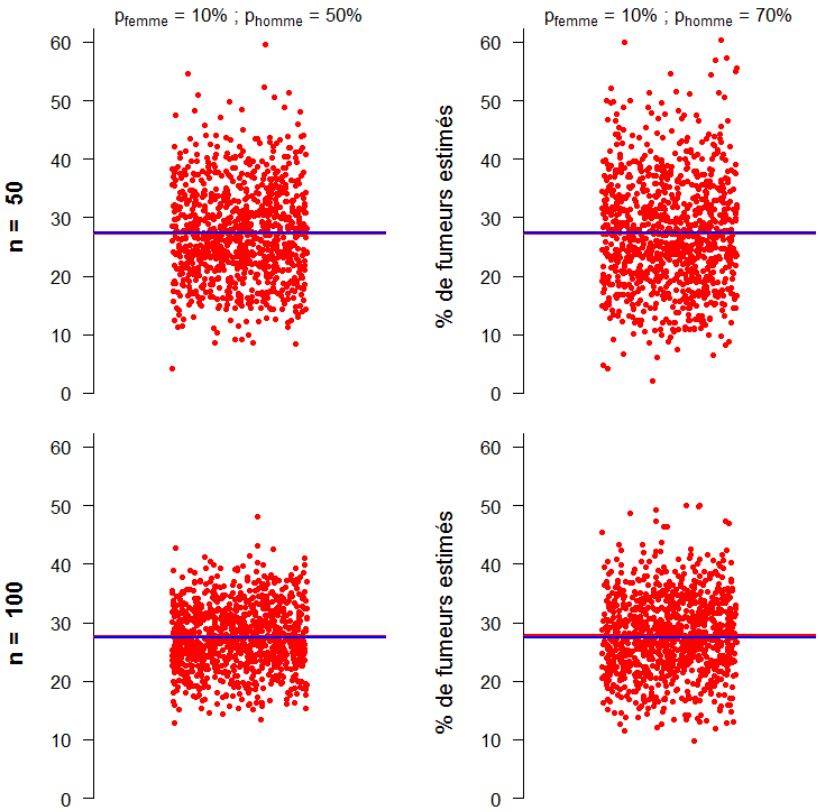
Lorsque le caractère d'intérêt est directement lié à la probabilité de non-réponse, c'est-à-dire avec une relation de causalité, le mécanisme est automatiquement non-ignorable :

$$\text{fumeur} \longrightarrow p$$

La moyenne des répondants et l'estimateur ajusté de la figure 12 et de la figure 13 respectivement, sont tous les deux biaisés comme illustré sur la 14. Cependant même si la relation ne pourra être brisée, on espère qu'en prenant en compte une information auxiliaire riche, on pourra réduire le biais le mieux possible.

En tout état de cause on ne peut savoir si on a brisé la relation existante, le cas échéant, entre le caractère d'intérêt et la probabilité de non-réponse.

FIGURE 13 – Illustration de l'absence de biais lorsque le mécanisme de non-réponse est aléatoire et traité.



7.4 Traitement de la non-réponse partielle par imputation multiple

Philosophie de l'imputation multiple et cas des données d'enquête

Les données manquantes peuvent être complétées par différentes méthodes. L'*imputation* consiste à prédire pour la donnée manquante une valeur à partir d'un modèle. L'imputation peut être *simple* lorsqu'une seule valeur est prédite et conduit alors à un (seul) jeu de données complet. Elle est *multiple* lorsque $M \geq 2$ valeurs sont prédites et conduit alors à $M \geq 2$ jeux de données complets.

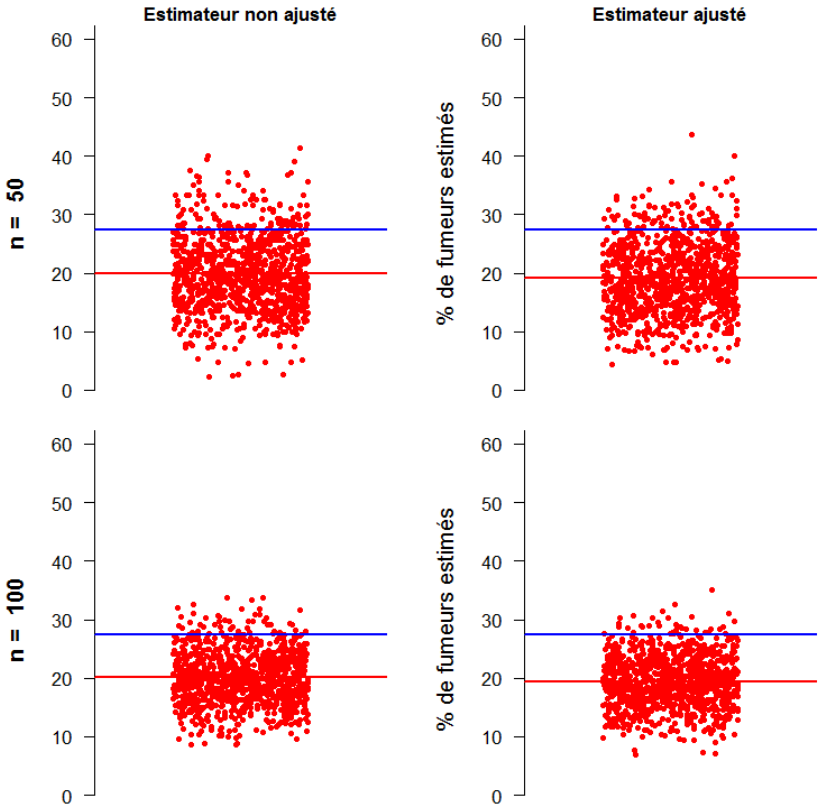
Derrière chaque méthode d'imputation il y a un ensemble d'hypothèses ou modèle. Toutes les méthodes d'imputation peuvent être motivées par le modèle suivant :

$$y_k = \eta(\mathbf{z}_k, \boldsymbol{\beta}) + \varepsilon_k \tag{7.1}$$

où $\mathbf{z}_k$ est le vecteur des variables auxiliaires pour l'élément k et $\boldsymbol{\beta}$ le vecteur de paramètres inconnus.

Le choix d'une méthode d'imputation requiert de faire un bilan de l'information auxiliaire disponible, d'examiner les données, de regarder pour quel type de

FIGURE 14 – Illustration de la présence de biais lorsque le mécanisme de non-réponse est non-ignorable et ceci même avec un traitement.



variable (continue ou catégorielle) l'imputation va se faire et de considérer le type de paramètre qui doit être *in fine* estimé (total, médiane etc.). Dès lors une certaine méthode d'imputation η dans le modèle (7.1) sera choisie. Il peut être cité à titre d'exemple : l'imputation par la moyenne, par le plus proche voisin, par la régression ou encore une méthode appelée *« predictive mean matching »* (cf. section 7.4 de cette partie).

Dans le cadre de l'imputation multiple on estime un paramètre d'intérêt θ sur chacun des M jeux de données alors produits par le modèle d'imputation. Le but est de combiner de manière adéquate ces M estimateurs et leur variance pour obtenir un estimateur « final » et sa variance.
On note par $\hat{\theta}_I^{(i)}$ l'estimateur (par exemple le π-estimateur) de θ sur le i-ème jeu de données complété $(i = 1, \ldots, M)$ et $U^{(i)}$ l'estimateur de sa variance. L'estimateur final noté $\hat{\theta}_{IM}$ est construit de la manière suivante :

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_I^{(i)} \tag{7.2}$$

Un estimateur de la variance de $\hat{\theta}_{IM}$ est donné par :

$$T = \widehat{V}(\hat{\theta}_{IM}) = \overline{U}_M + \left(1 + \frac{1}{M}\right) B_M \tag{7.3}$$

où $\overline{U}_M = \sum_{i=1}^{M} U^{(i)}/M$ et $B_M = \sum_{i=1}^{M} \left[\hat{\theta}_I^{(i)} - \hat{\theta}_{IM}\right]^2 /(M-1)$.

Le terme $\overline{U}_M$ est un estimateur de la variance due à l'échantillonnage (ou *variance intra*) et B_M est un estimateur de la variance due à l'imputation ou à la non-réponse (ou *variance inter*).

Si on considère un vecteur $\mathbf{\Theta}$ de paramètres $\theta_1, \ldots, \theta_j, \ldots, \theta_p$ à estimer, et $\{(\widehat{\mathbf{\Theta}}_I^{(i)}, \widehat{\mathbf{U}}_I^{(i)}) : i = 1, \ldots, M\}$ l'estimateur de $\mathbf{\Theta}$ et de la matrice de variance-covariance (VCE) associée $\mathbf{U}^{(i)}$ des estimateurs $\hat{\theta}_{jIM}$ définis chacun par l'équation 7.2, les équations analogues aux équations 7.2 et 7.3 peuvent s'écrire sous la forme :

$$\widehat{\mathbf{\Theta}}_{IM} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\mathbf{\Theta}}_I^{(i)} \tag{7.4}$$

et

$$\mathbf{T} = \widehat{\mathbf{V}}(\widehat{\mathbf{\Theta}}_{IM}) = \overline{\mathbf{U}}_M + \left(1 + \frac{1}{M}\right) \mathbf{B}_M \tag{7.5}$$

où $\overline{\mathbf{U}}_M = \sum_{i=1}^{M} \widehat{\mathbf{U}}_I^{(i)}/M$ est la matrice de variance-covariance *intra* et $\mathbf{B}_M = \sum_{i=1}^{M} \left(\widehat{\mathbf{\Theta}}_I^{(i)} - \widehat{\mathbf{\Theta}}_{IM}\right) \left(\widehat{\mathbf{\Theta}}_I^{(i)} - \widehat{\mathbf{\Theta}}_{IM}\right)^\top /(M-1)$ la matrice de variance-covariance *inter*.

L'utilisation de l'imputation multiple sur des données d'enquête est controversée. Alors que dans [Reiter et al., 2006] des recommendations sont faites dans le cadre de l'utilisation de l'imputation multiple, dans [Kim et al., 2006] il est indiqué que cela peut être risqué.

Dans le premier article [Reiter et al., 2006], les recommandations se basent sur une étude de simulation et compare différents modèles d'imputation. En particulier un modèle ne prenant pas en compte les variables du plan de sondage et un modèle de type régression les prenant en compte. Outre le fait d'introduire l'information auxiliaire adéquate pour rendre le mécanisme de non-réponse le plus ignorable possible, l'introduction dans le modèle d'imputation des variables du plan de sondage induit un faible biais des estimateurs des paramètres d'intérêt, $\hat{\theta}_{IM}$, en particulier lorsque ces variables sont reliées aux caractères d'intérêt. S'il n'est pas possible de prendre en compte ces variables du plan directement (trop de PSUs par exemple) il est conseillé d'introduire un proxy (la taille des PSUs par exemple). La non prise en compte des variables du plan dans la modèle d'imputation a pour résultat de produire des estimateurs des paramètres d'intérêt, $\hat{\theta}_{IM}$, sérieusement biaisés lorsque les variables du plan et les caractères étudiés sont liés. L'estimateur de la variance, $\widehat{V}(\hat{\theta}_{IM})$, produit des estimations plus grandes mais proches de ce que produit le même estimateur sur données complètes. En revanche lorsque les caractères étudiés ne sont pas liés aux variables du plan, le modèle d'imputation tenant compte des variables du plan produit toujours un estimateur du paramètre d'intérêt, $\hat{\theta}_{IM}$, faiblement biaisé mais il produit un estimateur de la variance, $\widehat{V}(\hat{\theta}_{IM})$, dont l'estimation

est plus grande que celle obtenue sur données complètes.

Le second article [Kim et al., 2006] porte sur l'étude du biais que peut comporter l'estimateur de la variance $\widehat{V}(\hat{\theta}_{IM})$. L'étude conclut que l'imputation multiple ne mène généralement pas à un estimateur sans biais de la variance car un terme de décomposition de la variance n'est pas pris en compte ; en l'occurence un terme fonction d'une covariance entre l'estimateur construit sur l'échantillon complet et l'estimateur construit sur données imputées. Comme ce terme peut être positif ou négatif, l'imputation multiple peut conduire à une sur-estimation ou sous-estimation de $V(\hat{\theta}_{IM})$.

Méthodes d'imputation

Dans ce qui suit les méthodes exposées ne sont pas forcément restreintes au cadre des données d'enquête. Cette section repose sur la référence [StataCorp, 2011] qui se base elle-même sur une nombreuse littérature que l'on ne citera pas car non explorée.

Généralités L'imputation (multiple) est dite univariée si une seule variable est à imputer ; elle est multivariée si plusieurs variables sont imputées à la fois. L'imputation univariée est à utiliser pour plusieurs variables lorsque ces variables sont indépendantes et seront utilisées dans des analyses séparées. Les logiciels, en particulier Stata V12[23], mettent à disposition différentes méthodes selon le type de variable à imputer : imputation par la régression, imputation *« predictive mean matching »* qui est similaire à l'imputation par la régression mais, en particulier, prédit une valeur parmi les valeurs observées ; elle est de plus utilisée lorsque la normalité du modèle sous-jacente est jugée suspecte. Une imputation tronquée est à utiliser pour une variable continue dont l'intervalle de valeurs est restreint. Il est aussi possible de traiter la censure dans l'imputation pour les variable à données censurées. Des méthodes d'imputations sont dédiées à l'imputation de variables catégorielles, dont les modalités sont ordinales ou non ordinales. Il existe enfin des méthodes traitant les variables dont les valeurs sont des comptages.

Dans le cas de plusieurs variables à imputer la méthode d'imputation dépend de la manière dont les données manquantes se répartissent « géographiquement » au sein de la table de données. La répartition peut être monotone, c'est-à-dire que les variables à imputer peuvent se réordonner dans la table de manière à ce qu'une variable (colonne) donnée possède des « trous » au moins sur les mêmes lignes que la variable précédente (*monotone-missing-pattern*). Dans ce cas une imputation univariée successive peut être utilisée basée sur les distributions conditionnelles marginales de chaque variable. Dans le cas contraire, la théorie requiert de procéder à une imputation multiple en considérant la distribution conjointe des variables à imputer.

23. StataCorp. 2011. Stata Statistical Software : Release 12. College Station, TX : StataCorp LP.

Imputation using chained equations Il existe une méthode alternative afin d'imputer plusieurs variables à la fois sans avoir la contrainte de la considération de la distribution conjointe de ces variables. Cette méthode appelée imputation selon des équations chainées (*imputation using chained equations*) (ICE) permet d'utiliser la flexibilité de l'imputation univariée, c'est-à-dire de spécifier un modèle (méthode et prédicteurs) pour chacune des variables à imputer, tout en réalisant une imputation multivariée requise pour un non *« monotone-missing-pattern »*. Elle peut être appelée sous d'autres expressions comme FCS pour *« fully conditional specification »* (*of prediction equations*) ou encore SRMI pour *« sequential regression multivariate imputation »*.

ICE est donc similaire à une imputation monotone dans le sens où la méthode se base sur une série de modèles d'imputation univariée. Cependant la méthode est itérative et ces itérations permettent de prendre en compte la dépendance des variables. La méthode repose sur le principe FCS c'est-à-dire que toutes les variables à imputer, excepté celle à imputer « en cours », sont utilisées comme prédicteurs dans l'équation de prédiction. Formellement si l'on considère $X_1, X_2, \ldots, X_p$ comme les variables à imputer et des prédicteurs sans données manquantes (variables indépendantes) notés $\mathbf{Z}$, le processus de l'imputation ICE est décrit par :

$$
\begin{aligned}
X_1^{(t+1)} &\sim g_1(X_1 \mid X_2^{(t)}, \ldots, X_p^{(t)}, \mathbf{Z}, \phi_1) \\
X_2^{(t+1)} &\sim g_2(X_2 \mid X_1^{(t+1)}, X_3^{(t)}, \ldots, X_p^{(t)}, \mathbf{Z}, \phi_2) \\
&\ldots \\
X_p^{(t+1)} &\sim g_p(X_p \mid X_1^{(t+1)}, X_2^{(t+1)}, \ldots, X_{p-1}^{(t)}, \mathbf{Z}, \phi_p)
\end{aligned}
$$

ceci pour les itérations $t = 0, 1, \ldots, T$ jusqu'à la convergence à l'étape $t = T$, où ϕ_j sont les paramètres du modèle. Les modèles d'imputations, $g_j(\cdot)$, peuvent être de différents types (densité normale, densité logistique, *etc.*), et considérés comme appropriés pour imputer X_j. Ces différents types de densités sont très utiles en pratique car ils permettent donc d'imputer en même temps et par différentes méthodes (régression logistique, linéaire, tronqué *etc.*) un ensemble de variables. C'est notamment cette flexibilité qui rend la méthode ICE populaire. Il est de plus possible de restreindre l'imputation à certaines lignes de la table de données (si « condition vérifiée » alors imputation).

Cependant une justification théorique générale de ICE est manquante. Le problème est que les densités conditionnelles $\{g_j(\cdot), j = 1, 2, \ldots, p\}$ peuvent ne pas correspondre en fait à une distribution conjointe (multivariée) de $X_1, X_2, \ldots, X_p$ sachant $\mathbf{Z}$. Dans cette situation la procédure ICE ne converge pas vers une distribution stationnaire ce qui peut poser la question de la validité de cette méthode statistique reposant sur ce principe de densités conditionnelles. Certains stipulent qu'en pratique cette problématique est peu importante mais ceci est encore du domaine de la recherche.

Modèle d'imputation

Outre le choix de la méthode d'imputation et des prédicteurs, la construction du modèle d'imputation requiert de prendre en compte la structure des données en particulier la structure *« clusterisée*[24] *»* que peuvent avoir par exemple des données d'enquête.

Après avoir imputer et avant de réaliser les analyses statistiques, il est une bonne pratique de vérifier que les résultats de l'imputation sont sensés à travers certains diagnostics.

Construction du modèle La construction du modèle d'imputation doit préserver les caractéristiques principales des données observées. Pour cela on doit :

- Utiliser tous les prédicteurs (utiles...par définition) autant que possible pour éviter de faire des hypothèses erronées relatives aux relations entre les variables.
- Inclure les variables du plan de sondage (poids de sondage, strates, *« clusters »* *etc.*). Cependant lorsque le plan est trop complexe il est difficile d'inclure directement certaines de ces variables, notamment en ce qui concerne les *« clusters »*. Dans ce cas il est recommandé d'inclure à défaut une variable de substitution (*proxy*) [Reiter et al., 2006].
- Respecter les dépendances des différents niveau d'information.

Si une covariable apparaît dans le modèle à estimer, elle doit alors apparaître dans le modèle d'imputation. Par conséquent la variable réponse du modèle d'analyse devrait toujours apparaître dans le modèle d'imputation. Toutes informations auxiliaires donnant de l'information sur le mécanisme de non-réponse doit figurer dans le modèle d'imputation afin de rendre le mécanisme de non-réponse ignorable autant que possible et ainsi d'améliorer la qualité des valeurs imputées.

Cas de la variable réponse Dans le cas où la variable réponse possède elle-même des données manquantes, il y a un débat autour de ce qui doit être fait pour la variable réponse : si elle doit être imputée ou bien si les observations correspondant à ses données manquantes doivent être écartées de l'analyse. Il ne semble pas avoir de réponse définitive à ce sujet. Il faut juste noter que plus le pourcentage de données manquantes est important, plus l'inférence sera sensible à une mauvaise spécification du modèle. De même si des variables ont une importante proportion de données manquantes, leurs données imputées auront plus d'influence sur les résultats.

Transformation de variables Bien que différentes méthodes d'imputation soient disponibles dans les logiciels, elles ne peuvent recouvrir toutes les distributions que les variables à imputer peuvent avoir. Appliquer une transformation à ces variables est alors possible, quitte à faire la transformation inverse après imputation.

24. C'est-à-dire le fait que des unités se regroupent au sein d'unités plus grosses e.g. des pièces au sein d'un logement.

Diagnostics après imputation Bien que cela soit encore du domaine de la recherche, il est utile de réaliser certains diagnostics après l'imputation, en particulier comparer la distribution des données imputées et celle des données observées. Un trop grand écart est généralement synonyme d'anomalies et le modèle d'imputation doit alors être révisé.

8 L'enquête Plomb-Habitat

8.1 Objectifs de Plomb-Habitat

L'enquête Plomb-Habitat est une enquête environnementale réalisée à partir d'un sous-échantillon d'enfants de l'enquête de prévalence du saturnisme infantile (6 mois-6 ans) menée par l'InVS dénommée Saturn-Inf et réalisée entre 2007 et 2009 en France. L'enquête Plomb-Habitat a été pilotée par le Centre Scientifique et Technique du Bâtiment (CSTB) et réalisée entre octobre 2008 et août 2009. La population cible initiale était constituée du parc de résidences principales en France métropolitaine. Du fait du sous-échantillonnage réalisé à partir des enfants de l'enquête de prévalence, la population pouvant être décrite était le parc de résidences principales en France métropolitaine, où au moins un enfant âgé de 6 mois à 6 ans était présent. Le déroulement de l'enquête ainsi que les exploitations des données ont été suivies dans le cadre d'un comité de pilotage appelé COPIL regroupant les différents partenaires de l'enquête. À savoir : le CSTB, l'École en Hautes Études en Santé Publique (EHESP), l'InVS, l'Hôpital Lariboisière AP-HP (Assistance Publique – Hôpitaux de Paris) et l'Institut Supérieur d'Agriculture de Lille (ISA).

Les objectifs de l'enquête Plomb-Habitat étaient :

- d'améliorer les connaissances sur les déterminants des plombémies ;
- d'identifier les sources et les compartiments environnementaux responsables des plombémies modérées (comprises entre 30 et 100 μg/L) ;
- de comparer la pertinence des analyses en plomb total et en plomb acido-soluble comme éléments explicatifs et/ou prédictifs des plombémies ;
- d'établir un modèle empirique de prédiction des plombémies en fonction des concentrations en plomb dans l'environnement ;
- de fournir un premier panorama de la contamination par le plomb dans le parc de logements français ;
- d'estimer la proportion de cas de saturnisme infantile (plombémie $\geq 100\mu$g/L) pour laquelle l'analyse des ratios isotopiques du plomb dans le sang et dans les compartiments environnement apportait une plus-value pour identifier la source.

8.2 Plan de sondage

L'enquête InVS Saturn-Inf a constitué une première phase. Le plan de sondage mis en œuvre est un plan à 2 degrés stratifié au premier degré.

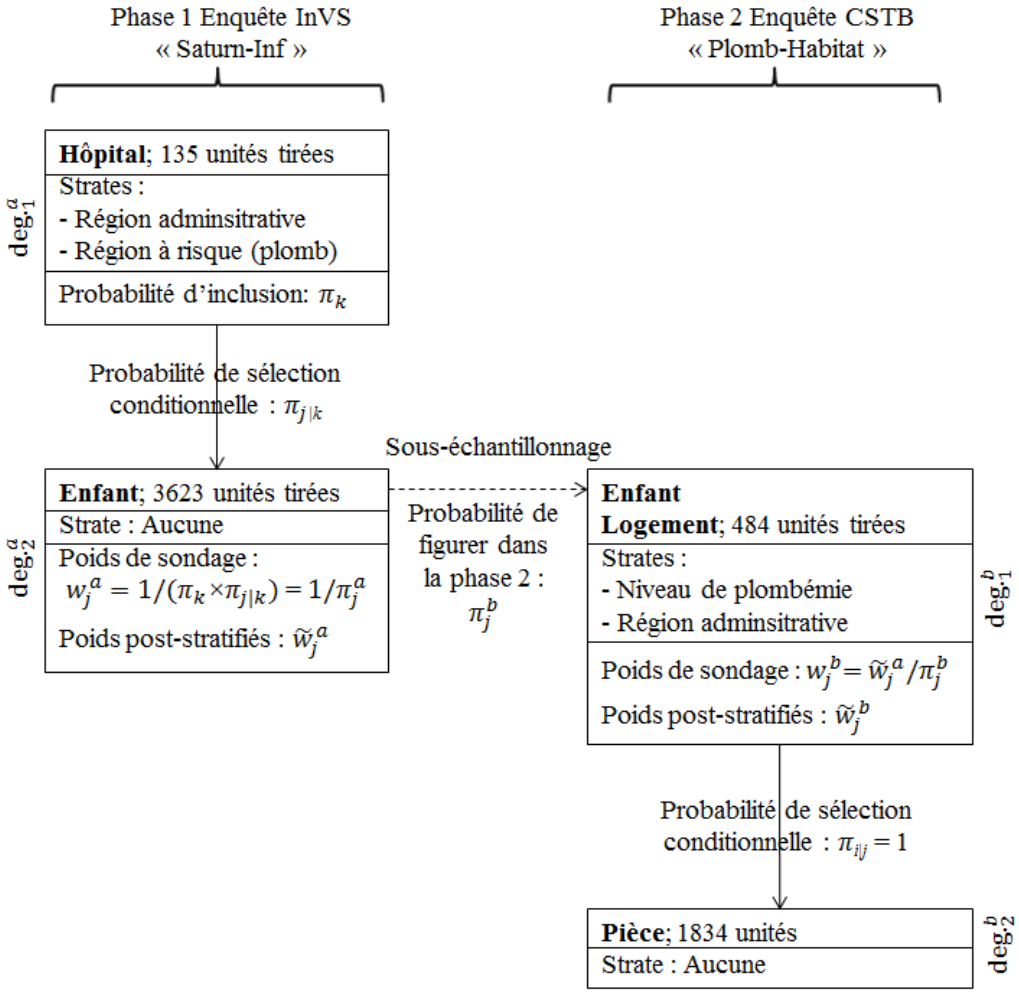
Dans une deuxième phase, l'échantillon de logements a été sélectionné selon un plan de sondage stratifié.
La taille de l'échantillon de logements visée était intialement de 500 logements. Cette taille est basée sur un calcul réalisé par l'InVS sous contrainte de puissance statistique, afin de mettre en évidence un certain risque pour l'enfant de dépasser le seuil de 100μg/L. Ce calcul a été fait à partir des données de l'étude américaine *National Survey of Lead and Allergens in Housing* de 2001 réalisée par l'U.S HUD :
Ainsi on souhaitait observer l'impact sur la plombémie, de la poussière contaminée par le plomb ainsi que l'impact de la peinture au plomb. Pour le risque lié à la peinture, l'indicateur d'exposition retenu était celui utilisé dans l'enquête américaine du HUD : plus de 10% de la surface totale des peintures intérieures dégradée. L'exposition mesurée en population générale (ici les non-malades) dans cette étude concernait 4% des logements avec la présence d'enfants de moins de 6 ans. Sous ces hypothèses, pour observer un *« odds ratio »* de 5 il était nécessaire de recruter 42 cas de saturnisme et 420 enfants avec une plombémie inférieure à 100μg/L dans l'enquête environnementale Plomb-Habitat pour obtenir une puissance d'enquête de 80%.
De manière analogue pour la poussière contaminée, 3% des logements montraient la présence de poussières avec des concentrations en plomb supérieures à 440μg/m^2 dans l'enquête HUD. Sous cette hypothèse, en supposant que le risque d'avoir une plombémie supérieure à 100μg/L était 5 fois plus important chez les enfants exposés que chez les non exposés, le calcul de la taille d'échantillon montre qu'un effectif de 594 logements à enquêter était nécessaire pour obtenir une puissance d'enquête de 80%.
Un compromis sur la taille d'échantillon a abouti à retenir 500 logements à enquêter.

Le plan de sondage de l'enquête résumé par la figure 15, est décrit ci-après.

Première phase (enquête de prévalence Saturn-Inf) :
135 hôpitaux ont été tirés comme unités primaires au premier degré selon une stratification construite à partir des régions administratives (22 en France métropolitaine) et du groupe à risque plomb (« à risque » versus « non à risque ») des bassins de population auxquels appartenait chaque hôpital. Ces groupes à risque ont été construits et intégrés dans la base de sondage par l'InVS. Les hôpitaux dans les régions administratives présentant une plus forte proportion de sites potentiellement pollués par le plomb (Île-de-France, Provence-Alpes-Cote-d'Azur, Nord et Haute-Normandie) ont été sur-représentés. La base de sondage du premier degré était composée des hôpitaux publics de plus de 300 hospitalisations par an (observés en 2007). Les services (non tirés aléatoirement), en général le service de pédiatrie (qui reçoit la plupart des enfants dans les CHR) et parfois le service de chirurgie infantile (dans certains CHU) ont été choisis pour représenter la base de sondage du deuxième degré.

Au second degré, l'inclusion des enfants a été aléatoire au cours d'une période variable selon les établissements, en fonction des disponibilités des médecins investigateurs.
Les critères d'inclusion des enfants étaient les suivants :

FIGURE 15 – Plan de sondage de l'enquête Plomb-Habitat



- Enfants âgés de 6 mois à 6 ans ;
- hospitalisés en hospitalisation complète ou de jour pendant la période d'étude ;
- pour lesquels une prise de sang avait été prescrite dans le cadre des soins ou pour lesquels un système de prélèvement (cathéter) était déjà en place ;
- résidant en France au moment de l'inclusion.

Ont été exclus :

- Enfants hospitalisés spécifiquement pour un bilan ou un traitement du saturnisme (chélation).
- Enfants atteints de pathologies mettant en jeu le pronostic vital.
- Enfants immunodéprimés.
- Enfants atteints de maladies chroniques influant sur l'immunité humorale et cellulaire.
- Enfants transfusés ou ayant reçu des gammaglobulines dans les 6 mois précédents.
  Seul un enfant par fratrie pouvait être inclus.

À ce stade, par strate, la probabilité d'inclusion de l'enfant j de l'hôpital k était :

$$\pi_j^a = \pi_k \times \pi_{j|k}$$

où π_k est la probabilité d'inclusion de l'hôpital k et $\pi_{j|k}$ la probabilité conditionnelle de l'enfant j égale au nombre d'enfants inclus dans l'hôpital k divisé par le nombre d'enfants hospitalisés dans le service pendant la période d'étude. Le poids de sondage de l'enfant était donc $w_j^a = 1/\pi_j^a$. L'exposant ^{a} est la notation pour la première phase (cf. section 3.5 de cette partie) et évite la confusion due à l'indexation par j faite aussi pour les logements de la 2^e phase.

Un coefficient de redressement (post-stratification), c, a été appliqué au poids de sondage de l'enfant de manière à ce que :

- La somme des poids des enfants de l'hôpital soit égale à l'effectif d'enfants hospitalisés dans l'hôpital en 2008. Pour ce faire, la base nationale du PMSI [25] 2008 a été utilisée.
- La somme des poids des enfants inclus dans une strate soit égale au nombre d'enfants réellement hospitalisés dans la strate en 2008. Pour ce faire, la base nationale du PMSI 2008 a été utilisée.
- La somme des poids des enfants égale le nombre d'enfants par ZEAT [26] par classe d'âge, selon le sexe et selon le fait de bénéficier de la CMUc [27]. Pour ce faire, la base de la Cnam [28] a été utilisée.
- La somme des poids des enfants inclus dans une ZEAT donnée soit égale au nombre d'enfants recensés dans cette ZEAT selon la base du recensement 2006 de l'INSEE [29] ; les données de la Cnam utilisées ci-dessus ne couvrant pas l'ensemble des enfants français mais seulement ceux inscrits au régime général de l'assurance maladie (soit 92 %).

Le poids final de l'enfant j dans l'enquête de prévalence a donc été $\widetilde{w}_j^a = w_j^a \times c$.

Deuxième phase (enquête environnementale Plomb-Habitat) :
Un plan de sondage stratifié aléatoire simple a été utilisé pour obtenir le sous-échantillon d'enfants (ou de manière analogue, de logements) de l'enquête Plomb-Habitat. Les critères d'inclusion étaient les suivants :

- Accord des parents (sauf cas de saturnisme - dans ce cas inclusion systématique).
- Le logement habituel était celui du parent présent à l'hôpital.
- L'enfant devait résider dans le logement depuis au moins 6 mois.

25. Programme de Médicalisation des Systèmes d'Information.
26. Zone d'Études et d'Aménagement du Territoire, correspondant à un regroupement de régions administratives.
27. Couverture Médicale Universelle complémentaire.
28. Caisse nationale d'assurance maladie.
29. Institut National de la Statistique et des Études Économiques.

Les strates ont été construites à partir de la région d'hospitalisation et du niveau de plombémie :

- $< 30\mu$g/L, tirage aléatoire ;
- [30 ; 100[μg/L, inclusion systématique ;
- $\geq 100\mu$g/L, inclusion systématique.

La taille du sous-échantillon obtenu a été de 484 au lieu de 500 [30].
Étant donné l'échantillon obtenu en première phase, la probabilité d'inclusion, π_j^b, de l'enfant j dans l'enquête Plomb-Habitat, a été calculée par l'InVS comme étant égale au nombre d'enfants inclus dans l'enquête Plomb-Habitat divisé par le nombre d'enfants éligibles à l'enquête Plomb-Habitat. Le poids de l'enfant de l'enquête Plomb-Habitat était donc $\omega_j^b = \widetilde{w}_j^a/\pi_j^b$. Un dernier coefficient de redressement a été appliqué par l'InVS sur ce poids ω_j^b afin que la somme des poids des enfants inclus dans une région d'habitation donnée soit égale au nombre d'enfants recensés dans cette région. Le poids post-stratifié est $\widetilde{\omega}_j^b$.
Ce poids final $w_j^b = \widetilde{\omega}_j^b$, pour l'enfant/le logement j de l'enquête environnementale, a été fourni au CSTB.

Les pièces à investiguer dans un logement donné n'ont pas été tirées aléatoirement mais automatiquement incluse dès lors qu'elles étaient d'un certains type (voir section suivante). Leur probabilité d'inclusion conditionnelle tout comme leur poids de sondage conditionnel sont donc égaux à 1.

8.3 Protocole de prélèvement et de mesure

In situ, le déroulement de l'enquête Plomb-Habitat s'est fait en deux parties : une partie questionnaire qui sera exposée dans la section suivante, et une partie mesures & prélèvements décrite ici. Le protocole de l'enquête a été détaillé par ailleurs dans un rapport CSTB [Chaventré et al., 2009]. De manière plus succincte, le protocole de prélèvement et de mesure est le suivant :

Prélèvement d'eau du robinet

L'enquêteur a réalisé un prélèvement d'eau froide au robinet de la cuisine. Si le prélèvement en cuisine n'était pas possible, à défaut le prélèvement s'est fait en salle de bain. Le type de prélèvement mis en œuvre a été le prélèvement après stagnation contrôlée de 30 minutes décrit en section 2.2 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel ». La procédure était la suivante :

30. En fait plus de 500 enquêtes ont été réalisées mais certaines d'entre elles ont été écartées au moment de la validation post-collecte. On pourra se reporter à la section 2 du chapitre 2 sur ce sujet.

- rinçage d'un flacon de 2 litres par trois fois (au robinet de l'évier) ;
- attente de 30 minutes avant d'effectuer le prélèvement, et rappel de la consigne aux occupants du logement de ne pas utiliser l'eau d'un quelconque point d'usage du logement (wc, lavabo, bain, douche ou machine à laver...) ;
- prélèvement au premier jet au robinet de l'évier en position eau froide, dans le flacon de 2 litres, avec un débit faible et en inclinant le flacon afin d'éviter le dégazage et rebouchage du flacon ;
- homogénéisation par agitation du flacon de 2 litres ;
- transvasage dans un flacon « métaux » de 250 mL contenant de l'acide nitrique (250 μL de HNO_3 (65–70%)) jusqu'au rétrécissement du goulot.

Prélèvement de poussière intérieure déposée au sol

Le prélèvement a été réalisé systématiquement dans les pièces suivantes dès lors qu'elles existaient dans le logement :

- la chambre de l'enfant ;
- le salon, séjour ou salle à manger ;
- l'entrée ;
- la cuisine ;
- la salle de jeu de l'enfant ;
- une chambre d'un autre enfant (en théorie l'enfant immédiatement le plus âgé).

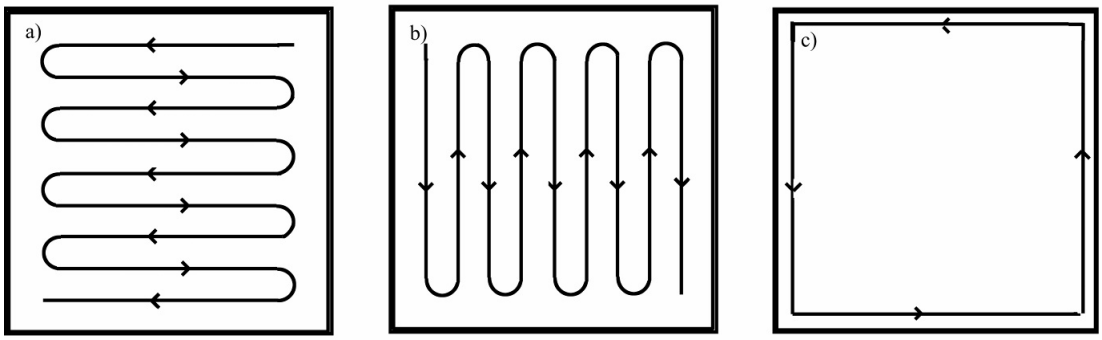
Un maximum de 5 pièces intérieures ont été investiguées, et constituaient l'ensemble des pièces intérieures dites pièces enquêtées pour la suite. En cas de parties communes (immeuble collectif) le palier de l'appartement et le hall de l'immeuble étaient de plus enquêtés.

Le prélèvement de poussière s'est fait par lingette humide (cf. section 3.2 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel »). Pour cela, un gabarit de 33cm × 33cm est mis par terre et l'enquêteur essuie selon la procédure décrite en figure 16. La lingette mise en tube est envoyée ensuite au laboratoire pour analyse.

Mesure du plomb surfacique des revêtements

Dans chacune des pièces enquêtées, l'enquêteur a réalisé des mesures du plomb surfacique des revêtements de la pièce par fluorescence X selon le protocole CREP décrit en section 3.1 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel ». Toutefois, la description de l'état de dégradation du revêtement s'est faite à partir d'un seuil de 0,5 mg/cm^2 et non à partir du seuil de 1 mg/cm^2, qualifiant un revêtement contenant du plomb selon le protocole CREP. Une première mesure du revêtement d'une UD, supérieure à 0,5 mg/cm^2 devait être

FIGURE 16 – Procédure d'essuyage du sol avec une lingette humide pour prélever la poussière dans une pièce, en 3 étapes a), b) et c).



confirmée par 2 autres mesures. Les mesures étaient réalisées autant du possible sur la partie inférieure à 1 mètre de l'UD puisque cette partie constitue la zone d'accessibilité de l'enfant. Les dimensions de chaque UD mesurée (longueur × largeur) devaient être renseignées.

En cas de revêtement dégradé avec une charge surfacique en plomb $> 0,5$ mg/cm^2 un prélèvement du revêtement (écaille) pouvait être réalisé afin de pouvoir doser le plomb en masse contenu dans le revêtement (une valeur exprimée en mg/g est alors fournie par le laboratoire). Ce type de prélèvement étant destructeur, il n'était réalisé que si l'occupant avait donné son accord.

Prélèvement sur l'aire de jeu extérieure de l'enfant

Si l'enfant jouait à l'extérieur du logement, un prélèvement de sol devait être réalisé sur l'aire de jeu principale de l'enfant (voir aussi la section 3.3 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel »). Si l'aire de jeu était sur un sol meuble, par exemple de la pelouse, un prélèvement de terre était réalisé par carottage sur une épaisseur de 2 cm puis envoyé au laboratoire. En cas d'aire de jeu sur sol dur, par exemple une cour bitumée, un prélèvement de poussière a été réalisé de la même manière que le prélèvement de poussière à l'intérieur. Un seul type d'air de jeu a été prélevé le cas échéant et ainsi une seule valeur de plomb sur l'aire de jeu extérieure, exprimée en mg/g ou en μg/m^2, est associée à chaque logement.

Prélèvement d'autres sources d'intoxication

Si le ménage occupant le logement enquêté utilisait des sources inhabituelles pouvant possiblement être une source d'intoxication de l'enfant (cf. section 5 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel »), un prélèvement de cette source pouvait être réalisé avec l'accord de l'occupant. Les

sources visées étaient la vaisselle, les cosmétiques et les remèdes traditionnels[31].

Mesure du plomb dans les prélèvements

Le dosage du plomb par le Laboratoire d'Étude et de Recherche en Environnement et Santé (LERES) a fourni pour chaque échantillon solide une valeur en plomb total et en plomb acido-soluble (cf. section 3.1) à partir d'une méthode développée par le laboratoire [Le Bot et al., 2011]. Brièvement la méthode d'analyse consiste en une minéralisation en 2 étapes : premièrement est ajoutée à l'échantillon de l'acide chlorhydrique 0,15 N à 37 °C pour solubiliser le plomb acido-soluble ; puis de l'eau régale (3 :1 HCl/HNO_3) est ajoutée à une partie aliquote à 95 °C pour solubiliser le plomb résiduel i.e. non acido-soluble. Le plomb dissout est alors analysé par ICP-MS (*Inductively-coupled plasma mass spectrometer*) selon la norme ISO 17 294-2.
Pour les prélèvements de sols meubles, l'échantillon a préparé avant le dosage du plomb : il a été séché à l'air à une température inférieure à 40 °C puis tamisé avec une maille inférieure à 250μm par un broyeur centrifugeuse.
Pour les sources d'intoxication type vaisselle, le plomb a été dosé à partir d'une solution acide solubilisant le plomb contenu dans le revêtement ; une valeur de plomb exprimée en μg/L est alors fournie. Pour les cosmétiques généralement sous forme de poudre, une concentration en plomb exprimée en mg/g a été fournie par le laboratoire.

8.4 Collecte de données descriptives par questionnaire

Un questionnaire regroupant différents thèmes a été élaboré pour l'enquête Plomb-Habitat par le COPIL. Approximativement 350 items étaient à renseigner par l'enquêteur pour chaque logement investigué autour des thèmes recoupant en particulier :

- la description du logement (type de logement, nombre de pièces et leur caractéristiques, environnement extérieur, *etc.*) et de ses équipements (système de ventilation, présence de WC, *etc.*) ;
- les travaux et l'entretien du logement (travaux de peinture, déshumidification, fréquence et moyen de nettoyage du sol, *etc.*) ;
- le ménage (statut d'occupation, métier ou loisir à risque d'exposition au plomb, utilisation de source à risque d'exposition au plomb, revenu, *etc.*) ;
- emploi du temps et comportement de l'enfant (temps hors du logement, comportement main-bouche, fréquentation des pièces, *etc.*).

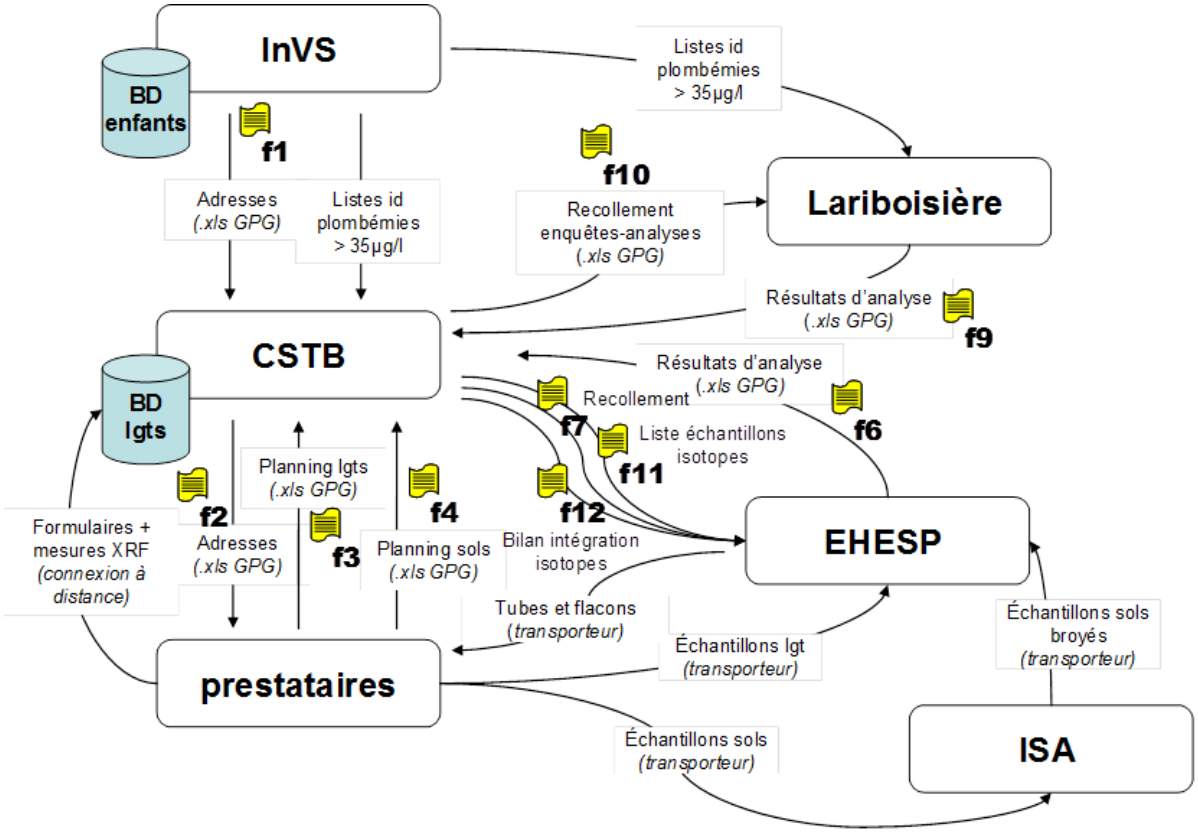
8.5 Système d'information

Le système d'information entre les partenaires et les enquêteurs ou prestataires est résumé dans la figure 17. Le CSTB collectait toutes les informations de l'enquête

31. Dans les faits, aucun remède n'a été prélevé.

Plomb-Habitat provenant des différents acteurs. L'InVS en charge des informations concernant l'enfant et de la base de données de l'enquête hospitalière Saturn-Inf, transmettait les adresses des logements pouvant être inclus dans l'enquête environnementale Plomb-Habitat. L'hôpital Lariboisière fournissait au CSTB les résultats des analyses sanguines des ratios isotopiques. Le LERES sis à l'EHESP recevait tous les prélèvements faits par les enquêteurs dans chaque logement, excepté l'échantillon de sol qui transitait par l'ISA pour préparation.

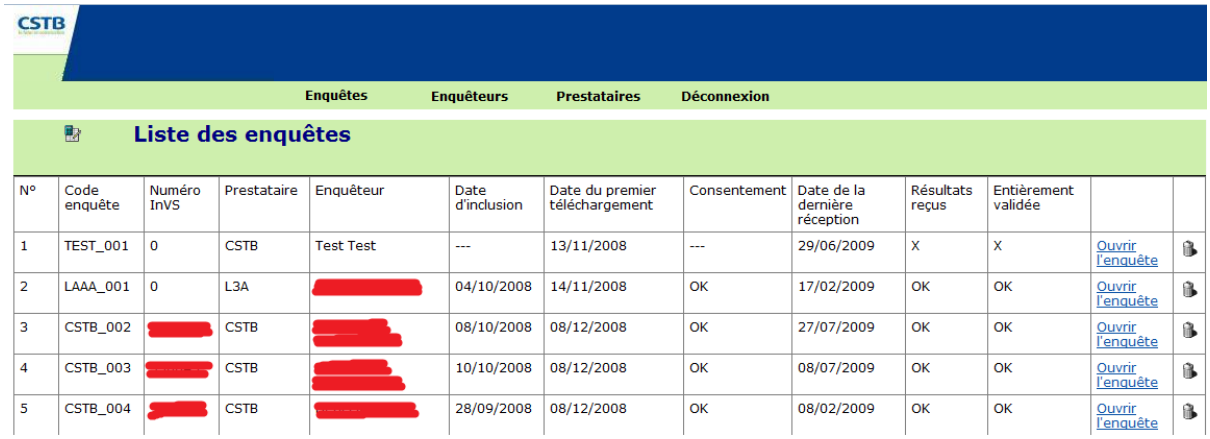
FIGURE 17 – Flux des données entre les différents partenaire de l'enquête Plomb-Habitat



Les données ont été stockées dans une base de données Microsoft SQL Server™développée par le pôle informatique du département Énergie-Santé-Environnement du CSTB. En parallèle, une application informatique a été développée pour le recueil des données de l'enquête Plomb-Habitat. Elle est composé d'une partie dite « client » installée sur les machines (ordinateurs ou tablettes) des enquêteurs, dans laquelle les enquêteurs saisissent les réponses à chaque item, et d'une partie « serveur » située au CSTB collectant les données stockées sur chaque poste client puis transmise via internet. L'enquêteur avait donc la possibilité de compléter directement les informations demandées *in situ*. L'application client ou serveur étaient analogues dans leur présentation. La figure 18 présente le tableau de bord du suivi des enquêtes à disposition du pilote de l'étude. La figure 19 montre la structure de l'application client : un volet à gauche avec l'arborescence des thèmes organisés en rubriques, à

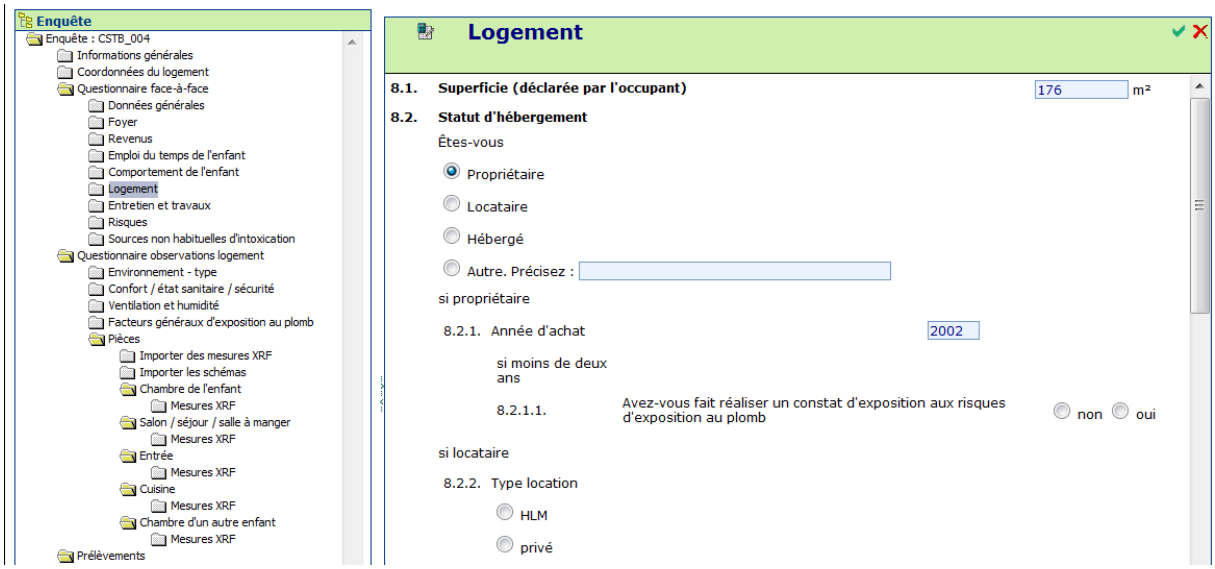
droite les items avec le moyen de renseigner la réponse par différents moyens selon le type d'information demandée (radio bouton, champs ouverts etc.).

FIGURE 18 – Extrait du tableau de synthèse de suivi des enquêtes de l'application serveur.



N°	Code enquête	Numéro InVS	Prestataire	Enquêteur	Date d'inclusion	Date du premier téléchargement	Consentement	Date de la dernière réception	Résultats reçus	Entièrement validée		
1	TEST_001	0	CSTB	Test Test	---	13/11/2008	---	29/06/2009	X	X	Ouvrir l'enquête	
2	LAAA_001	0	L3A	[redacted]	04/10/2008	14/11/2008	OK	17/02/2009	OK	OK	Ouvrir l'enquête	
3	CSTB_002	[redacted]	CSTB	[redacted]	08/10/2008	08/12/2008	OK	27/07/2009	OK	OK	Ouvrir l'enquête	
4	CSTB_003	[redacted]	CSTB	[redacted]	10/10/2008	08/12/2008	OK	08/07/2009	OK	OK	Ouvrir l'enquête	
5	CSTB_004	[redacted]	CSTB	[redacted]	28/09/2008	08/12/2008	OK	08/02/2009	OK	OK	Ouvrir l'enquête	

FIGURE 19 – Extrait de questionnaire de l'application client.



Chapitre 1

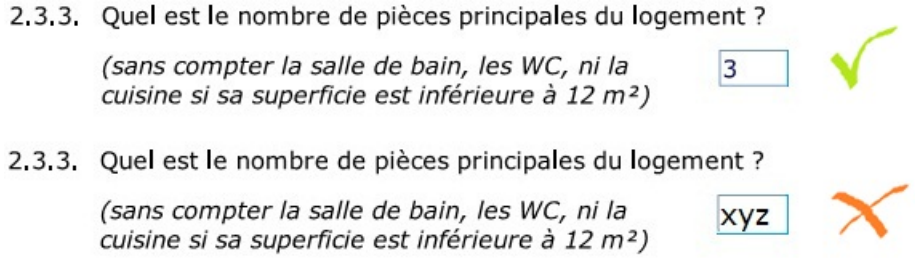
Validation des données

L'information réside essentiellement dans les données. Ainsi la première étape a été de s'assurer de la validité des données avant de réaliser les analyses statistiques d'une part et de fournir les données aux partenaires de l'enquête Plomb-Habitat d'autre part. Afin de disposer d'un jeu de données propre, des procédures de *« data-management »* ont été appliquées. Ces procédures ont permis de vérifier que les valeurs pour une variable donnée, étaient des valeurs possibles. Ces procédures ont permis également de contrôler la cohérence entre les valeurs de plusieurs variables collectées. Le bon enchainement des questions a de plus été contrôlé par ces procédures. Enfin ces procédures ont permis de construire des tables d'exploitation valides dans le sens où les diverses informations contenues dans différentes tables pour une entité statistique donnée, pouvaient être mises en correspondance.

1 Validation au niveau de l'application client et de l'application serveur

Afin de limiter les erreurs de saisies et les incohérences des données lors de leur collecte, des cadenas ont été placés dans l'application client de saisie des données au moment de son développement informatique.

Les champs devant recevoir une date ou un chiffre ont été dimensionnés afin de produire une erreur si autre chose que ces deux types de données était saisi :



Les radio-boutons ○ non ◉ oui ont été utilisés. Ils ont permis de détecter la non saisie de réponse car ils étaient initialement vierges, c'est-à-dire qu'aucune des réponses possibles n'étaient saisies par défaut. Sur l'exemple du Oui/Non, cela signifie

que ni le Oui, ni le Non était mis par défaut : ○ non ○ oui. Un comptage des radio-boutons renseignés par page permettait de signaler à la personne saisissant les données que certains radio-boutons n'avaient pas été renseignés.

La personne saisissant les données devaient valider chacune des pages de l'application client, en cliquant sur un bouton valider. De même, une fois toutes les informations saisies pour une enquête (un logement), le questionnaire rempli devait être validé par l'enquêteur avant d'envoyer les données dans la base du CSTB (l'enquêteur devait contractuellement relire ses données saisies afin de transmettre des données valides).

Afin de limiter les erreurs de saisie des mesures XRF qui étaient nombreuses pour chaque enquête, une procédure d'import automatique en base de données à partir du fichier de l'appareil de mesure utilisé a été développée. Ceci a été rendu possible par le fait qu'un seul type d'appareil à fluorescence X a été utilisé par les enquêteurs. Ainsi les fichiers de type Excel fournis par les appareils étaient homogènes. Seul un enquêteur a utilisé un appareil différent. Il a du se conformer par ses propres moyens au format de fichier demandé afin d'éviter de développer une procédure d'import spécifique.

À partir de l'application serveur, le coordinateur de l'étude[1] avait accès aux informations saisies pour chaque enquête (= chaque logement enquêté). Outre la réalisation du suivi du bon déroulement des enquêtes à partir de l'application serveur, le coordinateur a géré les erreurs les plus grossières ou facilement décelables à partir des données saisies. Un retour vers l'enquêteur pour corriger ces erreurs était alors fait par le coordinateur le cas échéant.

2 Validation post-collecte

La validation et le cas échéant les corrections n'ont pas été faites sur la base de données SQL Server mais sur des tables SAS[2] afin de garder la traçabilité des données. Ainsi, une fois qu'une enquête dans un logement était terminée et ses informations envoyées dans la base de données SQL Server, les données étaient importées via le logiciel SAS dans des tables dimensionnées de manière à rendre la validation possible.

2.1 Validation de la structure de la base

On appelle *validation de structure* tout ce qui touche à l'identification des enregistrements, à la correspondance entre les différentes tables et aux informations clés relatives aux critères d'inclusion d'un logement dans l'enquête environnementale Plomb-Habitat.

1. Monsieur Franck Chaventré, CSTB.
2. SAS System for Windows, version 9.1.3 ; SAS Institute Inc., Cary, NC.

La vérification du respect des critères d'inclusion a porté sur l'âge de l'enfant. Il devait être supérieur ou égal à 6 mois et strictement inférieur à 84 mois[3] à la date d'inclusion dans l'enquête Saturn-Inf. Cette vérification a également porté sur la durée de résidence dans le logement qui devait être d'au moins 6 mois à cette même date d'inclusion. Cette vérification a induit la suppression de 23 enquêtes qui avaient été réalisées bien qu'il ait été convenu contractuellement que l'enquêteur devait vérifier ces critères avant la réalisation de l'enquête.

Afin de procéder à cette validation, un programme codé sous le langage SAS était lancé après chaque import de base SQL Server sous SAS. Puisque les données arrivaient en base SQL Server au fur et à mesure que les enquêtes étaient réalisées, l'import s'est fait tout au long des enquêtes. Afin de faciliter la validation puis ensuite l'exploitation des données, des tables dites plates étaient construites à chaque import. Cette étape permettait de regrouper les données d'une même entité statistique dans une même table et ainsi de limiter le nombre de tables. Cela évitait de procéder à des concaténations de tables lors de la validation et de l'exploitation des données (en particulier par les partenaires de l'étude après que les données leurs ont été fournies) ; cette concaténation pouvant être source d'erreurs. Autrement dit, les informations d'une entité statistique figurant dans plusieurs tables SQL Server étaient regroupées dans une même table. Par exemple, la figure 20 montre deux tables, l'une nommée « Tenquetes » renseignant sur des informations telles que le « code_enquete » officiel pour chaque enquête (logement), la date de réalisation de l'enquête, la date d'envoi des données au CSTB etc. La seconde table nommée « Tqfaf_enfants » comportaient des informations telles que le sexe de l'enfant, sa date de naissance, ou encore si l'enfant était scolarisé ou non par exemple. Le regroupement des informations de ces deux tables s'est fait via la concaténation à partir de l'identifiant de l'enquête commun aux deux tables. Certains regroupements d'informations ont nécessité des concaténations plus complexes lorsque les informations étaient réparties dans plusieurs tables de données.

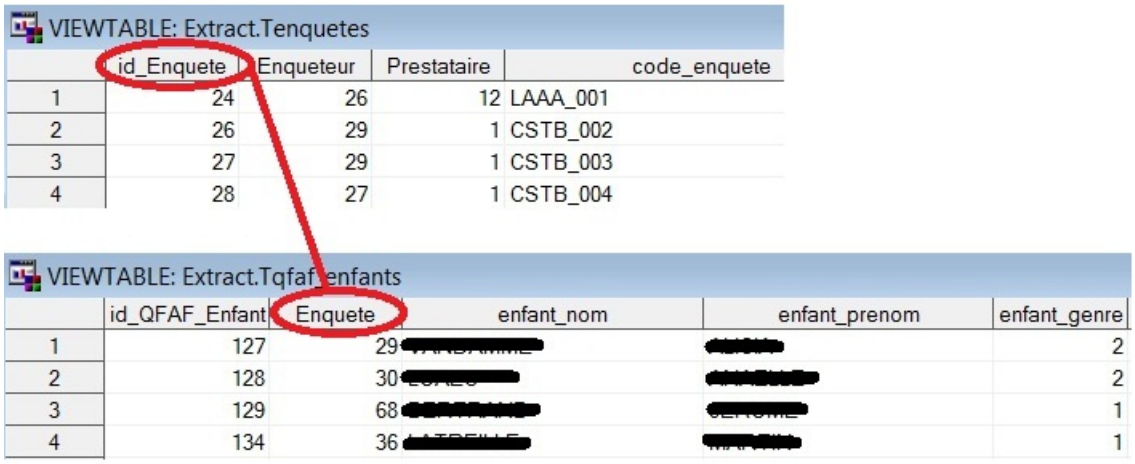
Après créations des tables plates, 30 contrôles de structure identifiés ont été lancés. Le résultat était constitué d'une table dans laquelle pour chaque entité statistique concernée un « *query text* » renseignait sur le problème, tel que par exemple « L'âge de l'enfant semble être en dehors des limites 6 mois-6 ans à la date d'inclusion. Vérifiez. » ou encore « Le prélèvement d'eau ne semble pas avoir eu lieu car aucun point de prélèvement n'a été renseigné sur la fiche. Vérifiez. » Comme ce dernier exemple, certains tests de structure permettaient de vérifier la présence d'un prélèvement qui devait avoir été fait pendant l'enquête. Si un test nécessitait un complément d'information de l'enquêteur, une requête lui était envoyée, généralement par courrier électronique.

2.2 Gestion des champs « Autre, précisez »

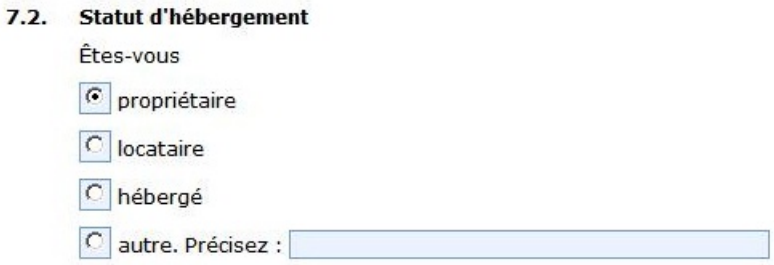
Certaines questions du questionnaire proposaient plusieurs modalités de réponses possibles et un champ pour précisez la réponse si cette dernière ne faisait pas

3. C'est-à-dire jusqu'à la veille du 7e anniversaire.

FIGURE 20 – Exemple de regroupement d'informations à partir de deux tables de données.



partie des modalités initialement proposées. Par exemple la question sur le statut d'hébergement du ménage dans le logement enquêté



possédait un champs permettant d'indiquer par exemple « Occupe un logement de fonction » comme ce fut le cas pour certaines enquêtes. Afin d'éviter un champs textuel correspondant en fait à une modalité de réponse possible, des procédures d'analyse textuelle de ces champs ont été appliquées à tous les champs « Autre, précisez » du questionnaire.

Au total 9 procédures ou tests d'analyse textuelle ont été nécessaires. En cas de correspondance avec l'une des modalités proposées la donnée était corrigée.

2.3 Tests de contraintes

Par définition, les *tests de contraintes* permettent de vérifier qu'un champs ne possède pas une valeur non autorisée ou peu plausible. Pour reprendre l'exemple du statut d'hébergement de la section 2.2, les valeurs possibles pour cet item étaient 1, 2, 3 ou 4 en base de donnée ; ces chiffres codant chacune des quatre modalités proposées respectivement. En fait, puisque la saisie des données s'est faite via l'application client de chaque enquêteur, aucune autre valeur que 1, 2, 3 et 4 ne pouvait être générée par l'application. Cependant un test de contrainte a tout de même été

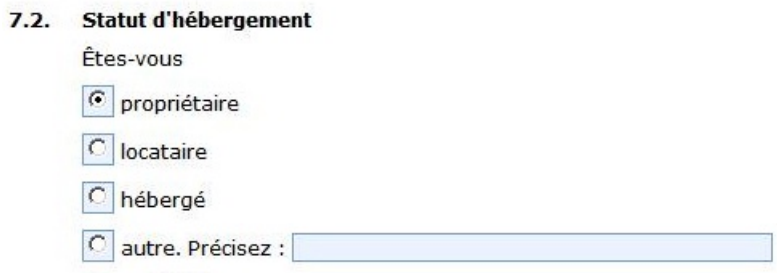
appliqué à chaque variable puisque, comme les données ont été corrigées pendant leur validation, il était nécessaire de s'assurer qu'une erreur de correction n'ait pas eu lieu. Par exemple, une erreur de frappe dans le code de correction aurait pu induire, qu'au lieu de la valeur 3, la valeur 33 ait été mise.

L'application d'un test de contrainte apparaissait plus importante pour les champs collectant une grandeur physique. Par exemple la longueur d'une pièce ne peut rationnellement pas être inférieure à une borne ni supérieure à une autre borne dans un logement. Pour ce test, une borne inférieure égale à 100 cm, et une borne supérieure à 1500 cm avaient été utilisées. Il se peut que dans certains logements considérés comme rares, une pièce puisse avoir une longueur supérieure à 15 mètres. Dans ce cas une vérification était effectuée auprès de l'enquêteur. Pour les grandeurs physiques, le test de contrainte étaient surtout utile pour repérer les erreurs d'unités. Pour la longueur d'une pièce, demandée en centimètres, le test sur cette variable a repéré les valeurs données majoritairement en mètres.

Au total, 319 tests de contraintes ont été codés sous SAS et appliqués sur l'ensemble des données. Les identifiants et les champs texte n'ont pas subi de test de contrainte de part leur nature.

2.4 Tests sur les enchainements logiques

Certains items du questionnaire ne devaient être remplis que si un item précédent avait reçu une réponse particulière. Par exemple, en reprenant de nouveau la question sur le statut d'hébergement,



le champs « Précisez » ne devait avoir une donnée que si « autre » avait été préalablement choisi. Le champs « autre » a été appelé « question mère », et le champs « Précisez » a été appelé « question fille ». À noter qu'un champs pouvait être à la fois « mère » et « fille ».

Les tests sur les enchainements logiques ont été de deux natures d'implication. La première est une implication directe, si « A alors B = SO » où A est la question mère, B la question fille et SO signifie « Sans Objet ». Dans l'exemple du statut d'hébergement, le test direct s'est appliqué aux situations où propriétaire, locataire ou hébergé avait été choisi ET le champs « Précisez » ne valant pas SO [A ET Non(B = SO)]. Le *query text* associé était alors « La question B a une réponse alors que

l'enchainement des questions ne le permet pas. Vérifiez. »
L'autre nature d'implication est inverse : si « B = SO alors A ». Le test a détecté les situations où la question fille valait SO alors que cela n'était pas justifié [(B = SO) ET Non(A)]. Par exemple, pour la question sur l'hébergement, cela se produit si le champs « Précisez » était vide alors que la réponse « autre » avait été préalablement choisie. Le *query text* associé étant alors « La question B vaut Sans Objet (777) alors que le ppsi n'est pas vérifié. Vérifiez. » où « ppsi » était la condition pour valoir SO et signifiait « n'est pas posée si ». En fait, le test inverse s'apparente à détecter une donnée manquante : si « autre » est choisi alors si le champs « Précisez » est vide, ce dernier a une donnée manquante. Mais le test inverse n'a pas été mis en place pour détecter les données manquantes. En effet, à la construction des tables plates, les champs vides ont été recodés en une valeur 777 indiquant SO et une valeur -999 indiquant une non-réponse. Le test inverse a alors détecté les situations où un item fille valant initialement 777, n'était plus justifié après que son item mère relatif ait été corrigé lors de la validation des données.

Au total 166 tests directs et inverses sur les enchainements logiques ont été codés et appliqués au données.

2.5 Tests de cohérence

À travers les questions posées, certaines informations pouvaient être redondantes, partiellement redondantes ou bien reliées par une implication. Afin d'assurer une cohérence entre ces informations, des tests dit de *cohérence* ont été codés et appliqués aux données.

Il existait certaines redondances sur lesquelles un test a du être effectué. Par exemple un test produisait le *query text* suivant : « La question 13.3 indique une salle de bain alors que la question 12.7 indique qu'il n'y en a pas. Corrigez. » Concernant les implications, par exemple un test produisait un *query text* indiquant « La question 14.2.2 indique un ventilateur dans la salle de bain alors que la question 12.7 indique qu'il n'y a pas de salle de bain. Corrigez. », ou encore « Le loisir est pratiqué mais la fréquence de pratique est nulle. Corrigez. »

Au total, 54 tests de cohérence concernant les données de Plomb-Habitat ont été réalisés. Certaines informations entre les bases de données de l'enquête Saturn-Inf et de Plomb-Habitat étant communes ou proches, 16 tests de cohérences ont de plus été appliqués pour vérifier la cohérence de ces informations.

2.6 Tests de détections des données manquantes

Afin de limiter le nombre de données manquantes, une procédure automatique détectant les non-réponses a été codée. Ces données manquantes ont pu parfois être complétées de façon déterministe c'est-à-dire à partir d'une autre information

donnée, soit figurant dans le questionnaire, soit fournie de façon annexe. Ces informations annexes ont été majoritairement fournies après avoir recontacté l'enquêteur.

Au total, 261 tests de détection des données manquantes ont été réalisés sur les données.

2.7 Points particuliers

Certaines informations n'ont pas pas être vérifiées ; la présence de canalisations en plomb par exemple. En effet il est parfois impossible sur le terrain de savoir si des canalisations en plomb sont présentes dans un logement car encastrées dans les murs. En logement collectif, les accès aux caves permettent plus facilement de repérer ces canalisations. Cette difficulté de repérage explique la présence d'une modalité « Ne sait pas » dans la table 10 présentée en résultat en section 2.1 du chapitre 2. On peut faire l'hypothèse que parmi les logements relevés comme n'ayant pas de canalisations en plomb, certains d'entre eux en possédaient pourtant.

3 Synthèse

Les procédures utilisées dans le cadre de ce travail afin de rendre les données de l'enquête Plomb-Habitat aussi propres que possible ont été exposées : la vérification de l'appartenance de la valeur d'un item à son ensemble de définition, la cohérence entre l'information de plusieurs variables, l'enchainement des items ainsi que la complétion déterministe des données manquantes.

À l'issu de ce chapitre, le travail effectué permet la fourniture des données aux différents partenaires de l'enquête Plomb-Habitant exploitant les données. Le travail permet surtout de rendre possible les analyses statistiques réalisées dans le chapitre 2 qui suit ainsi que celles des chapitres 3 et 4.

Chapitre 2

Estimation des niveaux en plomb dans les compartiments environnementaux en milieu résidentiel

À l'échelle métropolitaine, les distributions et la prévalence de dépassement de seuils, des teneurs en plomb dans les différents compartiments environnementaux des logements français ont été estimées pour la première fois. Les compartiments environnementaux ciblés par l'enquête Plomb-Habitat ont été l'eau du robinet, la poussière déposée au sol à l'intérieur des logements, la poussière déposée au sol en parties communes, les revêtements intérieurs et le sol de l'aire de jeu extérieure de l'enfant. La population décrite était constituée des résidences principales en France métropolitaine, abritant au moins un enfant âgé de 6 mois à 6 ans en 2008.

Les outils de la théorie des sondages (section 5 de la partie « Spécificités des données d'enquête ») ont été utilisés afin d'obtenir les estimations souhaitées. Afin d'améliorer les estimations, les poids de sondage des logements de l'enquête Plomb-Habitat ont été tout d'abord redressés par post-stratification sur des critères relatifs aux logements. Dès lors, une stratégie de description des niveaux en plomb a été mise en place, en particulier dans des sous-populations (domaines), afin de fournir un état de la contamination le plus informatif possible pour les pouvoirs publics.

Un article scientifique intitulé *« Lead Contamination in French Children's Homes and Environment »*, relatif aux résultats de ce chapitre a été publié en 2012 [Lucas et al., 2012] (cf. annexe 12).

1 Redressement des poids de sondage des logements et estimation des niveaux en plomb

1.1 Post-stratification

Afin d'améliorer les estimations, un redressement des poids de sondage des logements, w_j^b, fournis par l'InVS et décrits en section 8.2 de la partie « Spécificités des données d'enquête » a été réalisé. Le redressement a été fait dans le présent travail par post-stratification. Les informations auxiliaires disponibles au niveau de la population de logements décrites, via le recensement 2006 [1] réalisé par l'INSEE [INSEE, 2008], et en lien supposé avec les niveaux en plomb dans les logements étaient :

- la période de construction du logement (avant 1949, à partir de 1949) ;
- la région administrative (22 régions métropolitaines) dans laquelle le logement se situe et ;
- le type de logement (individuel/collectif).

Concernant la période de construction, ajouter un découpage par rapport à l'année 1915 aurait fourni une information plus riche au regard de la réglementation sur l'usage de la céruse et des dérivés du plomb dans la peinture en France [Lucas, 2011]. Cependant cette information auxiliaire n'était pas disponible à l'échelle nationale dans la base INSEE du rencensement 2006 (la première plage disponible est « avant 1949 »).
Quatre-vingt huit post-strates ont été constituées ($2 \times 22 \times 2$). Certaines de ces post-strates ne possédaient aucun logement parmi l'échantillon de 484 logements. Certaines d'entre elles ne possédaient qu'un nombre de logements jugé trop faible pour les calculs de variance comme le montre la table 3.

Algébriquement, pour calculer les variances, chaque post-strate doit contenir au moins 2 logements. Il est donc nécessaire de regrouper des post-strates. Puisque la variance diminue lorsque le nombre d'entités augmente, il est néanmoins préférable de faire en sorte que ces regroupements comportent plus de 2 logements. Il ne semble pas exister de recommandation sur le nombre d'entités que doit comporter chaque regroupement. Dès lors, il a été choisi arbitrairement un nombre minimal de 10 logements.

Afin de conserver l'information sur la période de construction jugée la plus importante (importante au regard des niveaux en plomb) parmi le type de logement (collectif ou individuel), la région et cette période de construction, les logements construits avant 1949 et les logements construits à partir de 1949 n'ont pas été regroupés. Le premier critère de regroupement a été le type de logement jugé *a priori* le moins important parmi les trois. Des regroupements de régions voisines ont

1. Ces informations résultent du cumul des enquêtes annuelles de recensement, réalisées de 2004 à 2008 en partenariat avec les communes et apportent un éclairage sur de nombreux thèmes, notamment sur les caractéristiques de la population, l'emploi, la formation, les déplacements, les logements.

TABLE 3 – Répartition des 484 logements dans les 88 post-strates.

Post-strate	n	Post-strate	n	Post-strate	n	Post-strate	n
11-COL-AP49	28	24-COL-AP49	1	41-COL-AP49	4	53-COL-AP49	1
11-COL-AV49	12	24-COL-AV49	0	41-COL-AV49	2	53-COL-AV49	1
11-IND-AP49	16	24-IND-AP49	14	41-IND-AP49	10	53-IND-AP49	15
11-IND-AV49	13	24-IND-AV49	6	41-IND-AV49	9	53-IND-AV49	7
21-COL-AP49	6	25-COL-AP49	0	42-COL-AP49	6	54-COL-AP49	1
21-COL-AV49	3	25-COL-AV49	0	42-COL-AV49	0	54-COL-AV49	0
21-IND-AP49	7	25-IND-AP49	13	42-IND-AP49	7	54-IND-AP49	12
21-IND-AV49	7	25-IND-AV49	4	42-IND-AV49	1	54-IND-AV49	5
22-COL-AP49	2	26-COL-AP49	0	43-COL-AP49	4	72-COL-AP49	1
22-COL-AV49	1	26-COL-AV49	1	43-COL-AV49	1	72-COL-AV49	0
22-IND-AP49	5	26-IND-AP49	2	43-IND-AP49	7	72-IND-AP49	14
22-IND-AV49	8	26-IND-AV49	15	43-IND-AV49	2	72-IND-AV49	6
23-COL-AP49	4	31-COL-AP49	11	52-COL-AP49	3	73-COL-AP49	4
23-COL-AV49	2	31-COL-AV49	0	52-COL-AV49	0	73-COL-AV49	1
23-IND-AP49	7	31-IND-AP49	18	52-IND-AP49	13	73-IND-AP49	6
23-IND-AV49	4	31-IND-AV49	18	52-IND-AV49	5	73-IND-AV49	7
74-COL-AP49	0	93-COL-AP49	13	82-COL-AP49	9	94-COL-AP49	0
74-COL-AV49	1	93-COL-AV49	0	82-COL-AV49	2	94-COL-AV49	2
74-IND-AP49	7	93-IND-AP49	10	82-IND-AP49	11	94-IND-AP49	3
74-IND-AV49	2	93-IND-AV49	10	82-IND-AV49	6	94-IND-AV49	2
91-COL-AP49	2	83-COL-AP49	4				
91-COL-AV49	1	83-COL-AV49	1				
91-IND-AP49	6	83-IND-AP49	4				
91-IND-AV49	4	83-IND-AV49	1				

Légende. COL : Logement collectif, IND : Logement individuel, AV49 : Logement construit avant 1949, AP49 : Logement construit à partir de 1949. Les chiffres correspondent à la codification INSEE des régions (voir annexe 1).

été effectués dès lors que le regroupement par type de logement n'avait pas permis d'atteindre 10 logements au minimum (figure 21). *In fine*, les post-strates utilisées pour le redressement sont celles affichées dans la table 4. Pour la région Île-de-France, les logements enquêtés étaient en assez grand nombre pour ne procéder à aucun regroupement. Ces regroupements sont dépendants de l'exploitant des données bien que basés sur des critères objectifs ; une autre personne aurait alors pu aboutir à différentes post-strates. Les regroupements ont abouti à 24 post-strates finales.

Les données du recensement 2006 de l'INSEE indiquent que le nombre de résidences principales en France métropolitaine abritant au moins un enfant âgé de 6 ans au plus est de 3 857 529. Mais notre population décrite ne comporte pas les résidences principales abritant uniquement au moins un enfant de moins de 6 mois. Afin de corriger ce nombre, l'hypothèse faite a été la suivante : on suppose que les enfants âgés de 6 ans au plus se répartissent de manière uniforme dans les 14 semestres composant la plage d'âges [0 mois ; 84 mois[. Cette hypothèse ne semblait pas être mise en défaut après avoir établi la répartition des français métropolitains par classe d'âge à partir de données mises à disposition par l'INSEE [2], montrée en

2. Population par sexe et âge en 2007 fournie par le fichier BTX_TD_POP1B_2007.xls téléchargé le 4 avril 2011. Les populations légales et les résultats statistiques 2007 sont obtenus à partir du cumul des informations collectées lors des cinq enquêtes de recensement de 2005 à 2009.

FIGURE 21 – Regroupements de régions, avec leur code INSEE, symbolisés par une même couleur afin d'obtenir 24 post-strates.

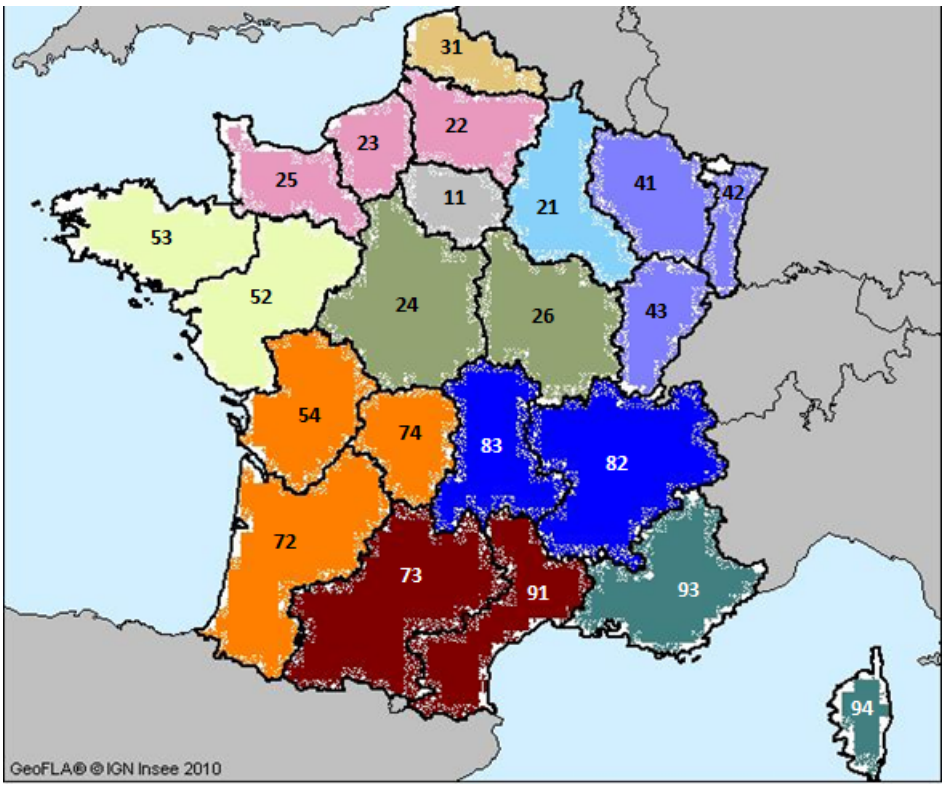


figure 22. Dès lors l'hypothèse sous-jacente faite est que la répartition à travers les 14 semestres doit être approximativement uniforme. La décision collégiale prise en COPIL Plomb-Habitat a été d'appliquer un coefficient égal à 13/14 aux effectifs INSEE de chaque post-strate afin d'obtenir leur taille dans la population de logements à décrire. Ainsi le nombre de résidences principales en France métropolitaine abritant au moins un enfant âgé de 6 mois à 6 ans était de 3 581 991.

Le poids de sondage, w_j^b, fourni par l'InVS comme explicité en fin de section 8.2 de la partie « Spécificités des données d'enquête », ont pour somme par post-strate h finale, la somme indiquée dans la colonne $\sum_{j\in h} w_j^b$ de la table 5. Le but de la post-stratification est que chaque somme devienne égale à la valeur indiquée en colonne $\sum_{j\in h} w_j^{\text{insee}}$ calculée à partir des données INSEE. Pour cela les poids finaux redressés doivent être calculés selon :

$$\widetilde{w}_j^b = w_j^b \times c_h$$

pour chaque logement j appartenant à la post-strate finale h, avec c_h le coefficient de redressement pour la strate h indiqué dans la table 5.

À partir de ces poids finaux $\widetilde{w}_j^b$ les paramètres d'intérêt de la section 1.2 qui suit peuvent dès lors être estimés.

TABLE 4 – Répartition des 484 logements dans les 24 post-strates finales.

Post-strate finale	**n**
11-COL-AP49	28
11-COL-AV49	12
11-IND-AP49	16
11-IND-AV49	13
21-AP49	13
21-AV49	10
22-23-25-AP49	31
22-23-25-AV49	19
24-26-AP49	17
24-26-AV49	22
31-AP49	29
31-AV49	18
41-42-43-AP49	38
41-42-43-AV49	15
52-53-AP49	32
52-53-AV49	13
54-72-74-AP49	35
54-72-74-AV49	14
73-91-AP49	18
73-91-AV49	13
82-83-AP49	28
82-83-AV49	10
93-94-AP49	26
93-94-AV49	14

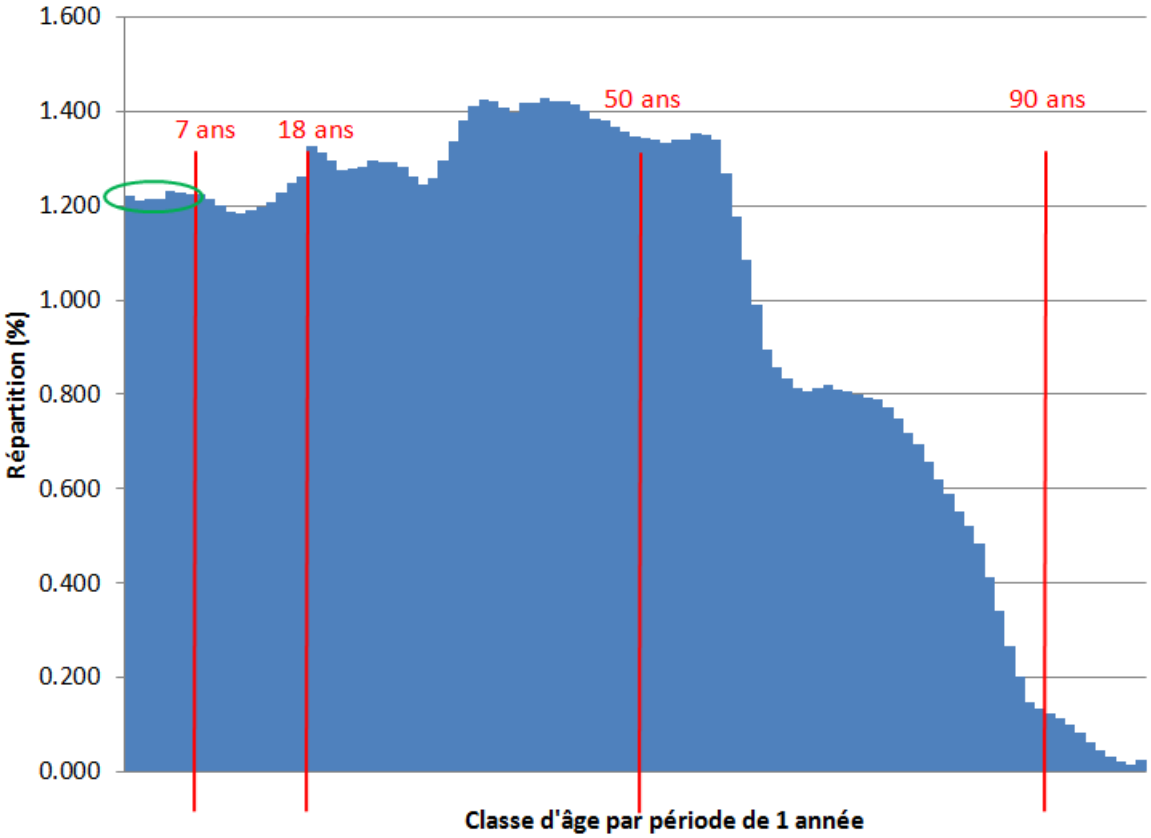
1.2 Estimation en population des niveaux en plomb et dans des sous-populations

Afin de fournir un état de la contamination par le plomb du milieu résidentiel français abritant des enfants en bas âges, les estimations des niveaux en plomb doivent concerner les compartiments environnementaux dont on sait qu'ils peuvent contenir du plomb. Le protocole de l'enquête Plomb-Habitat a été dimensionné de manière à ce que ces compartiments soient échantillonnés. Dès lors, la première échelle de description est l'échelle globale, c'est-à-dire des estimations dans la population entière à décrire.

Parmi les réglementations des différents pays, seule la réglementation française se base sur un dosage acido-soluble du plomb dans les compartiments environnementaux (à notre connaissance). Afin de comparer les niveaux en plomb total et en plomb acido-soluble, la description de la contamination par le plomb est faite pour ces deux dosages. Cette description permettra au pouvoirs publics de jauger l'utilité du dosage acido-soluble plutôt que le dosage en plomb total.

Les niveaux en plomb sont décrits dans l'eau du robinet ; pour l'eau le plomb est à la fois total et acido-soluble car il est présent sous forme ionique (Pb2+). Les niveaux en plomb dans la poussière intérieure déposée au sol sont décrits ainsi que les niveaux dans la poussière déposée au sol dans les parties communes le cas échéant. Les niveaux en plomb dans les revêtements intérieurs sont de plus décrits

FIGURE 22 – Répartition par classe d'âge de la population française en 2007.



en particulier selon les différents supports des revêtements (métalliques versus non métalliques) afin de décrire autant que possible la prévalence de la céruse appliquée habituellement sur des supports non métalliques. Enfin les niveaux en plomb de l'aire de jeu extérieure principale de l'enfant sont décrits.

De plus, afin de mettre en évidence le cas échéant l'impact de facteurs sur les niveaux en plomb, une description à partir de certaines sous-populations est réalisée en fonction du compartiment environnemental décrit. En particulier, l'influence de la période de construction du logement, eu égard à l'usage historique du plomb dans l'habitat, est d'intérêt pour les revêtements ainsi que pour la poussière intérieure. La réglementation relative aux canalisations en plomb rend pertinente la description des niveaux en plomb dans l'eau du robinet selon la période de construction des logements. En outre, les niveaux selon la présence de canalisations au plomb déclarée est naturellement d'intérêt pour l'eau du robinet. Les niveaux en plomb sont de plus décrits selon l'environnement urbain ou rural car particulièrement d'intérêt pour l'eau du robinet, la poussière intérieure et le sol extérieur des aires de jeu.

Afin de fournir des résultats en fonction, le plus possible, des dates de la réglementation de l'usage du plomb dans la peinture, les périodes de construction des logements utilisées ont été les suivantes : Avant 1949, de 1949 à 1973, de 1974 à 1993, à partir de 1994. Les dates de 1949 et 1993 correspondent respectivement aux décrets du 30 décembre 1948 n° 48-2034 et à l'arrêté du 1er février 1993 respective-

TABLE 5 – Calcul des coefficients de redressement à appliquer par post-strate finale.

Post-strate finale	$\sum_{j\in h} w_j^b$	$\sum_{j\in h} w_j^{\text{insee}}$	**Coef. de redressement**
11-COL-AP49	464 879,1600	396 131,4945	0,8521
11-COL-AV49	212 034,3640	112 537,7553	0,5308
11-IND-AP49	201 848,7780	164 522,1135	0,8151
11-IND-AV49	140 329,8130	63 088,9185	0,4496
21-AP49	62 893,3467	53 536,6922	0,8512
21-AV49	42 098,6561	22 956,0943	0,5453
22-23-25-AP49	275 001,9155	217 819,1005	0,7921
22-23-25-AV49	156 170,0755	94 048,7763	0,6022
24-26-AP49	179 064,8137	160 667,1800	0,8973
24-26-AV49	138 009,2050	69 125,7608	0,5009
31-AP49	276 065,5997	160 599,2225	0,5817
31-AV49	79 613,3204	91 094,2868	1,1442
41-42-43-AP49	309 310,1469	219 919,6624	0,7110
41-42-43-AV49	104 577,8622	82 580,6942	0,7897
52-53-AP49	421 364,5930	302 018,7463	0,7168
52-53-AV49	118 785,4198	82 500,3748	0,6945
54-72-74-AP49	327 126,0990	211 653,7780	0,6470
54-72-74-AV49	65 176,8922	80 612,7860	1,2368
73-91-AP49	287 422,5918	224 374,9033	0,7806
73-91-AV49	98 881,3942	65 625,6177	0,6637
82-83-AP49	397 820,4259	330 972,2938	0,8320
82-83-AV49	193 953,5793	97 419,1912	0,5023
93-94-AP49	298 098,7059	220 112,6255	0,7384
93-94-AV49	72 531,2979	58 073,1257	0,8007
Somme	4 923 058	3 581 991	-

ment. La date de 1974 ne correspond à rien relativement à la réglementation. Cette date était une borne disponible parmi les périodes de construction proposées dans le questionnaire sur la date de construction du logement enquêté. Elle permet de fournir 2 périodes de construction d'amplitude d'environ 20 ans. La création de 2 périodes comprises entre 1949 et 1993 permet d'étudier 4 périodes de construction au total et de mieux appréhender l'évolution des niveaux selon l'âge des logements, contrairement à ce qu'aurait pu fournir comme information l'utilisation de seulement 3 périodes (Avant 1949, de 1949 à 1993, à partir de 1994).

Le caractère urbain ou rural associé à l'environnement extérieur de chaque logement investigué a été déterminé grâce à la commune où se situait le logement. Les communes sont classées par l'INSEE dans une tranche détaillée d'unité urbaine. La variable INSEE (TDUU1999 [3]) comporte initialement 27 modalités (de la commune rurale de moins de 50 habitants à la commune appartenant à l'unité urbaine de Paris). Les modalités « rurales » et « urbaines » de cette variable TDUU1999 ont été regroupées en une seule modalité respective.

3. Données téléchargées le 13 octobre 2010 à partir de l'adresse : http://www.statistiques-locales.insee.fr/esl/baseTelechProduit.asp?strProd=1637&IdSousTheme=2&IdSource=&NomThemeOuSource=R%C3%A9gions%2C+d%C3%A9partements+et+villes+de+France

Afin de pouvoir calculer des variances dans les analyses de sous-populations, des regroupement de strates ont du être réalisés puisque certaines strates au premier degré ne possédaient qu'une seule PSU (hôpital) (cf. fin de section 3.4 de la partie « Spécificités des données d'enquête »). Dans ce cas les regroupements ont été faits sur une base de proximité géographique et des groupes à risque plomb (cf. section 8.2). Pour cela, à chaque strate a été associée une liste de strates du même groupe à risque plomb et jugées proches géographiquement ; lors des regroupements cette liste a été utilisée pour décider à quelle strate, la strate avec une seule PSU devait être combinée. Si ce premier regroupement ne réglait pas le problème de l'unique PSU, la fusion se faisait avec la seconde strate de la liste *etc.*

Afin de réaliser les calculs, le package « survey » [Lumley, 2004, Lumley, 2010b, Lumley, 2010a] du logiciel R[4] a été utilisé. Les fonctions du package qui ont été utilisées pour réaliser les estimations et les graphiques relatifs sont celles décrites dans la table 6.

TABLE 6 – Fonctions du package « survey » de R utlisées pour l'inférence descriptive.

Nom de la fonction	**Usage**
`svydesign`	Spécification du plan de sondage et des poids de sondage
`svytotal`	Estimation de totaux
`svymean`	Estimation de moyennes et de proportions
`svyquantile`	Estimation de quantiles
`svyby`	Estimation dans des sous-populations
`svyhist`	Affichage d'un histogramme
`svyboxplot`	Affichage d'une boîte à moustache
`svyplot`	Affichage d'un nuage de points

Les paramètres statistiques estimés sont les quantiles (ou percentiles) d'ordre 0,05, 0,25, 0,50, 0,75, 0,95. La moyenne géométrique est de plus estimée car les distributions des niveaux en plomb dans les médias environnementaux comme dans le sang, sont généralement proche d'une distribution Log-Normale. La moyenne géométrique est communément fournie à travers la littérature lorsqu'il s'agit, par une seule quantité, d'indiquer une idée des niveaux en plomb dans un média donné. La moyenne arithmétique est également estimée à titre indicatif et comparatif car certains articles scientifiques affichent cette statistique.

Les résultats descriptifs sont présentés sur cas complets c'est-à-dire à partir des logements ayant une valeur disponible. Autrement dit, les logements ayant une donnée manquante pour la quantité décrite, ont été supprimés. Des éléments de discussion sur ce sujet sont donnés en section 2.5 de la partie « Discussion ». Les valeurs inférieures aux limites de quantification (LQ) du laboratoire ont été remplacées par

4. R Core Team (2010). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http ://www.R-project.org/.

la valeur LQ/2. De même des éléments de discussion sur ce sujet sont donnés en section 2.5 de la partie « Discussion ».

Les résultats sont présentés en détails dans un rapport d'étude contractuel rendu à la Direction Générale de la Santé du Ministère de la Santé ainsi qu'à la Direction de le l'Habitat et de l'Urbanisme (DHUP) du Ministère du Logement. Ici seule une synthèse des principaux résultats est faite. L'ensemble des résultats ainsi que les éléments de discussion relatifs sont disponibles via ce rapport sur le site internet du CSTB à l'adresse http://www.cstb.fr/uploads/media/PLOMB-HABITAT.pdf. De plus une partie des résultats de ce rapport avec leur discussion est disponible dans un article scientifique publié [Lucas et al., 2012] disponible en annexe 12.1. Il a été choisi de retenir ici les éléments jugés les plus informatifs.

2 Résultats

La table 7 fournit des éléments descriptifs sur certaines caractéristiques des logements, en particulier sur les facteurs utilisés dans le processus de redressement. À titre informatif les effectifs et pourcentages de l'INSEE sont indiqués. Le plus grand écart entre les estimations obtenues à partir des données de l'enquête Plomb-Habitat et les chiffres INSEE concerne le type de logement. Ce résultat n'est pas anormal dans la mesure où les 2 modalités, logement individuel et logement collectif, ont été regroupées pour la construction des 22 post-strates finales (hormis pour la région Île-de-France). En revanche le regroupement de régions administratives lors de la post-stratification a eu peu d'impact ; les écarts entre chiffres INSEE et les estimations sont faibles.
L'année 1993 n'est pas une borne de classes de la variable INSEE (nommée ACHL) renseignant sur la période d'achèvement de la maison ou de l'immeuble. C'est l'année 1990 qui est utilisée et c'est la raison pour laquelle dans la table 7, 1993 n'apparaît pas dans la description de la période de construction. La précision sur la date de construction demandée dans le questionnaire de Plomb-Habitat a permis d'utiliser 1993 comme borne dans les estimations. Bien que les périodes de construction détaillées après 1949 n'aient pas été utilisées dans la post-stratification, les poids de sondage finaux relatifs aux logements permettent d'obtenir des effectifs estimés proches des effectifs INSEE.

2.1 Eau du robinet

Distribution des concentrations en plomb

La table 8 et la figure 23 associée montrent que la distribution du plomb dans l'eau du robinet dans la population de logements décrite comporte principalement des valeurs inférieures à 4 μg/L. 58% ($IC_{95\%}$ = 50-66) des logements ont une concentration en plomb dans l'eau du robinet inférieure strictement à la LQ. Pour mémoire la limite (LQ) du laboratoire à partir de laquelle il était capable de quantifier le plomb dans l'eau du robinet était de 1 μg/L.

TABLE 7 – Description des caractéristiques de la population de résidences principales.

	Estimation Plomb-Habitat				INSEE	
Caractéristique	***N***	**% estimé**	**$IC_{95\%}$**	***n***	**Effectif**	**%**
Nb. total de logements	3 581 991	100	-	484	3 581 991	100
Période de construction						
$<$ 1949	919 663	26	19-32	174	919 663	26
$\geq$ 1949	2 662 328	74	68-81	310	2 662 328	74
(Détail $\geq$ 1949 :)						
[1949 ; 1974[	964 895	27	19-35		965 270	27
[1974 ; 1990[	642 428	18	10-26		662 485	18
$\geq$ 1990	1 055 005	29	22-37		1 034 573	29
Type de logement						
Individuel	2 472 968	69	60-78	349	2 123 036	59
Collectif	1 109 023[a]	31	22-40	135	1 458 955	41
Régions administratives						
Alsace	102 123	2,9	0-5,9	14	105 119	2,9
Aquitaine	152 563	4,3	0-10,5	21	165 822	4,6
Auvergne	75 541	2,1	0-6	10	68 427	1,9
Basse-Normandie	88 885	2,5	0-5,6	17	82 953	2,3
Bourgogne	70 255	2	0,8-3,1	18	86 097	2,4
Bretagne	176 129	4,9	1,2-8,7	24	176 530	4,9
Centre	159 538	4,5	0,6-8,4	21	143 696	4
Champagne-Ardenne	76 493	2,1	0,1-4,2	23	76 493	2,1
Corse	14 471	0,4	0-1,2	7	14 812	0,4
Franche-Comté	67 096	1,9	0-4,1	14	67 204	1,9
Haute-Normandie	109 420	3,1	0-7,2	17	110 780	3,1
Île-de-France	736 280	20,6	10,3-30,8	69	736 280	20,6
Languedoc-Roussillon	143 980	4	0-9,6	13	139 769	3,9
Limousin	40 503	1,1	0,6-1,7	10	34 887	1
Lorraine	133 281	3,7	1,8-5,7	25	130 177	3,6
Midi-Pyrénées	146 021	4,1	2,9-5,3	18	150 232	4,2
Nord-Pas-de-Calais	251 694	7	0-15,2	47	251 694	7
Pays de la Loire	208 390	5,8	3,6-8,1	21	207 989	5,8
Picardie	113 563	3,2	1,3-5	16	118 135	3,3
Poitou-Charentes	99 201	2,8	1,2-4,4	18	91 557	2,6
Provence-Alpes-Côte d'Azur	263 715	7,4	4,5-10,3	33	263 374	7,4
Rhône-Alpes	352 850	9,9	3,3-16,4	28	359 964	10
Urbanisation						
Rural	933 192	26,1	17,7-34,4	164	948 573	26,5
Urbain	2 648 799	73,9	65,6-82,2	320	2 633 418	73,5
(Détail urbain :)						
Urbain $<$ 200 000 hab.	1 336 750	37,3	27,6-47,1	211	1 209 026	33,8
Urbain $\geq$ 200 000 hab.	1 132 049	36,6	26,2-47,1	109	1 424 392	39,8

Légende. [a] : dont 1 079 743 avec des parties communes (129 logements dans l'échantillon).

La table 9 et la figure 24 associée montrent que la distribution du plomb dans l'eau du robinet dans la population de logements selon la période de construction comportent des valeurs majoritairement inférieures à la LQ (médiane $<$ LQ). Sur la

TABLE 8 – Distribution des concentrations en plomb dans l'eau du robinet (μg/L).

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Estimation	472	3 461 328	< 1	< 1	< 1	1,1	5,4	1,8	< 1
$IC_{95\%}$			-	-	-	$<$ 1-1,6	3,9-9,5	1,4-2,2	-

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique.

FIGURE 23 – Distribution des concentrations en plomb dans l'eau du robinet.

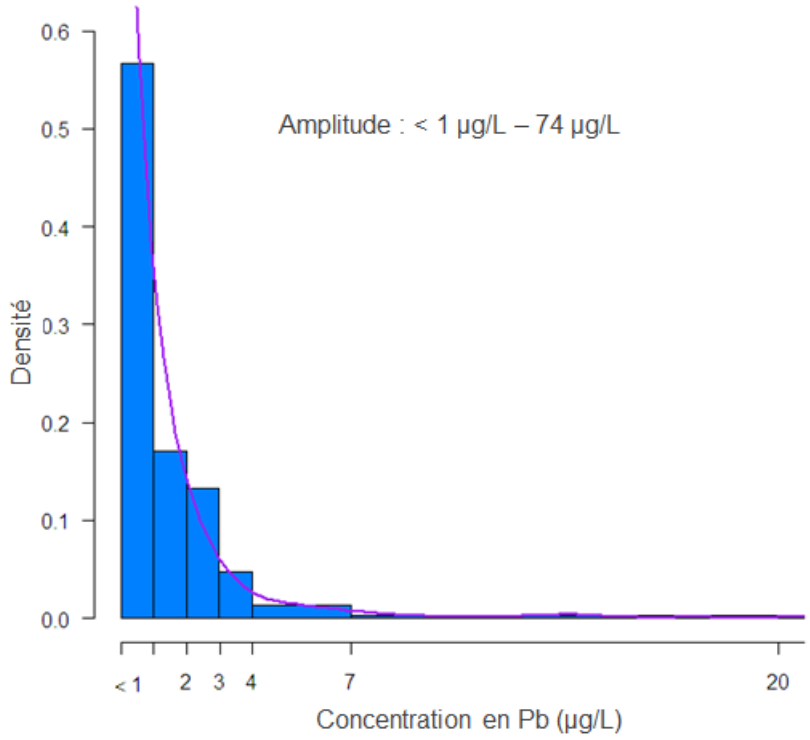


figure 24, chaque point est un logement de l'échantillon. Plus le point est gros, plus il représente de logements dans la population. La dispersion des concentrations diminue avec le caractère récent des logements. Les plus fortes concentrations croissent avec l'âge du logement.

En présence de canalisations en plomb (table 10), les concentrations en plomb dans l'eau du robinet apparaissent légèrement supérieures à celles mesurées en absence de telles canalisations. Sur l'information de la présence de canalisations en plomb, voir aussi la section « Points particuliers » en page 83.

Les concentrations en plomb dans l'eau du robinet apparaissent légèrement plus élevées en milieu urbain qu'en milieu rural (table 11). Plus de 75% des logements en milieu rural ont une concentration inférieure à la LQ.

TABLE 9 – Distribution des concentrations en plomb dans l'eau du robinet (μg/L) selon la période de construction.

Stat.	***n***	***N***	**P5**	**P25**	**P50**	**P75**	**P95**	**m.a.**	**m.g.**
Avant 1949	170	883 684	< 1	< 1	< 1	1,8	14,3	3,4	1,3
$IC_{95\%}$			-	-	-	1,2-6,3	9,5-33,3	2-4,8	1-1,7
De 1949 à 1974	121	949 133	< 1	< 1	< 1	< 1	4,7	1,3	< 1
$IC_{95\%}$			-	-	-	-	2-36	$<$ 1-1,8	-
De 1975 à 1993	64	670 275	< 1	< 1	< 1	< 1	3	1,1	< 1
$IC_{95\%}$			-	-	-	-	1,7-18	$<$ 1-1,5	-
À partir de 1994	117	958 237	< 1	< 1	< 1	1,2	2,7	1,2	< 1
$IC_{95\%}$			-	-	-	$<$ 1-2,1	1,7-5	$<$ 1-1,5	-

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique.

FIGURE 24 – Concentrations en plomb dans l'eau du robinet selon la période de construction.

Comparaison aux valeurs de référence

Pour rappel à ce sujet, voir la section 3.5 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel ».
Environ 1 % ($IC_{95\%}$ = 0-1,9) des logements ont une concentration en plomb dans l'eau du robinet dépassant le seuil de 25 μg/L et 2,9 % ($IC_{95\%}$ = 1,2-4,5) des logements ont une concentration dépassant le seuil de 10 μg/L.

TABLE 10 – Distribution des concentrations en plomb dans l'eau du robinet (μg/L) selon la présence de canalisations en plomb relevée par l'enquêteur.

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Présence	12	67 621	< 1	< 1	< 1	4,3	5,7	2,6	1,3
$IC_{95\%}$			-	-	-	$<$ 1-5,9	4,1-31	1,4-3,7	$<$ 1-2
Absence	369	2 945 849	< 1	< 1	< 1	1	4,7	1,7	< 1
$IC_{95\%}$			-	-	-	$<$ 1-1,7	3-13,6	1,2-2,2	-
Ne sait pas	91	568 522	< 1	< 1	< 1	1,2	8,6	1,9	< 1
$IC_{95\%}$			-	-	-	$<$ 1-8,9	2,3-25,1	1-2,9	-

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique.

TABLE 11 – Distribution des concentrations en plomb dans l'eau du robinet (μg/L) selon l'environnement.

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Urbain	311	2 587 384	< 1	< 1	< 1	1,3	6,1	1,9	1
$IC_{95\%}$			-	-	-	1-1,8	4,2-13,1	1,4-2,4	$<$ 1-1,1
Rural	161	873 944	< 1	< 1	< 1	< 1	2,8	1,3	< 1
$IC_{95\%}$			-	-	-	-	1,8-26	$<$ 1-1,7	-

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique.

2.2 Poussière intérieure et poussière en parties communes

Distribution des charges en plomb

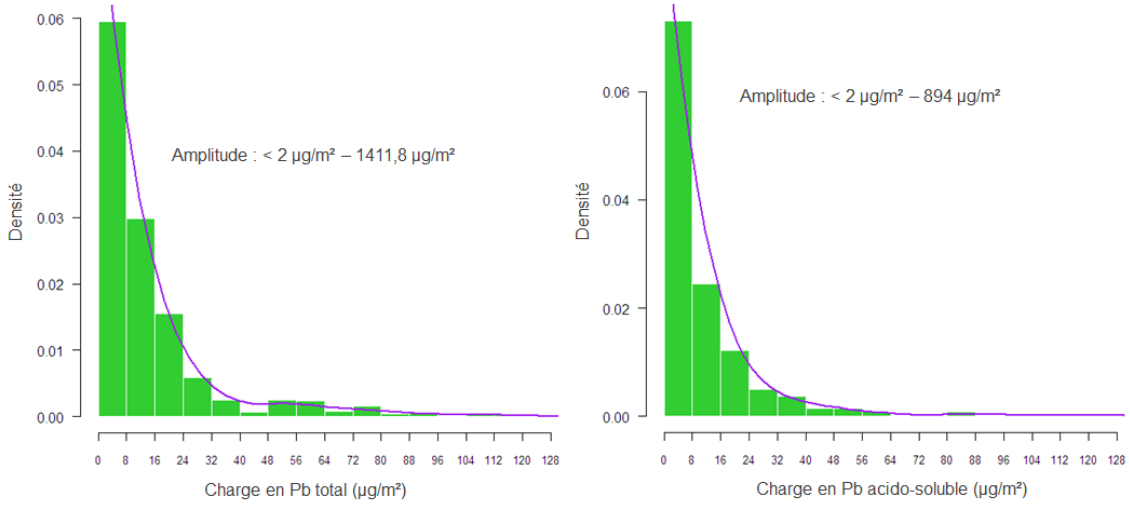
La table 12 présente la distribution des charges moyennes par logement, en plomb total et acido-soluble. La figure 25 présente les distributions sous forme d'histogramme. Cette charge moyenne a été calculée comme la moyenne arithmétique des charges disponibles à l'intérieur de chaque logement. La limite de quantification du plomb dans les poussières était de 1 μg/m^2 pour le dosage en plomb acido-soluble et de 2 μg/m^2 pour le dosage en plomb total.

TABLE 12 – Distribution des charges moyennes en plomb dans la poussière intérieure déposée au sol (μg/m^2).

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Pb total	471	3 453 789	< 2	3,7	8	17,3	62,6	18,8	8,8
$IC_{95\%}$			-	3-5,2	7-10,3	14,3-22,4	51,8-87,3	14,9-22,7	7,5-10,4
Pb a.-s.	471	3 449 152	1,4	3	7	14,3	41	13,7	6,9
$IC_{95\%}$			$<$ 1-1,8	2,4-3,5	5,7-8,3	12-17,1	34-59	10,9-16,4	5,9-8,2

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

FIGURE 25 – Charges moyennes en plomb dans la poussière intérieure déposée au sol.



Selon la période de construction, il apparaît clairement que les niveaux en plomb dans la poussière intérieure des logements les plus récents (construits à partir de 1994) sont plus faibles que dans les logements construits avant 1994. Ce constat est valable pour les charges en plomb total (table 13) comme en plomb acido-soluble (table 14). En revanche, il n'est pas clair que les niveaux en plomb croissent avec l'âge des logements. Il semble que les niveaux soient approximativement équivalent pour les logements construits avant 1994. Puisque l'hypothèse faite était que les charges en plomb dans la poussière seraient majoritairement influencées par les niveaux en plomb dans la peinture, eux-même liés à la réglementation, il était attendu que les niveaux en plomb dans la poussière diminuent lorsque les logements devenaient de plus en plus récents.

TABLE 13 – Distribution des charges moyennes en **plomb total** dans la poussière intérieure déposée au sol ($\mu g/m^2$).

Stat.	***n***	***N***	**P5**	**P25**	**P50**	**P75**	**P95**	**m.a.**	**m.g.**
Avant 1949	171	896 435	< 2	5,3	12,3	28,2	87,5	24,7	12,6
$IC_{95\%}$			-	2,5-9,2	7,9-16,1	16,3-47,8	61,6-694,8	17,3-32,1	9,1-17,6
De 1949 à 1974	120	949 097	< 2	5,1	9,6	17,9	66,3	21,9	9,7
$IC_{95\%}$			-	1,7-8,4	6,2-16,7	13,9-22,5	29,9-302,9	11,1-32,6	6,8-13,9
De 1975 à 1993	64	670 275	3,1	6,5	10,6	22,4	59	21,5	12,6
$IC_{95\%}$			2,3-3,3	3,3-9,2	7,2-20,6	11,9-56,2	34,1-1411,8	13,4-29,6	9-17,5
À partir de 1994	116	937 981	< 2	2	3,8	8,1	24,4	8,1	4,4
$IC_{95\%}$			-	$<$ 2-2,4	2,4-7,3	6-11,9	14,3-71,6	5,2-10,9	3,5-5,5

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique.

Les charges en plomb dans la poussière intérieure déposée au sol apparaissent plus élevées en milieu urbain (table 15). Le rapport des moyennes géométriques

TABLE 14 – Distribution des charges moyennes en **plomb acido-soluble** dans la poussière intérieure déposée au sol (μg/m^2).

Stat.	*n*	*N*	P5	P25	P50	P75	P95	m.a.	m.g.
Avant 1949	171	896 435	1,5	4,3	9,2	20	83,5	18,7	9,9
$IC_{95\%}$			1,5-2	2,1-6,9	6,8-11,3	12,2-38,5	42,8-408	12,6-24,9	7,1-13,6
De 1949 à 1974	121	951 064	1,4	4,3	8,7	15,4	29	16	8,2
$IC_{95\%}$			< 1-2,6	1,7-6,8	5,5-14,4	12,4-23,7	23,7-110,3	8,6-23,4	5,6-11,9
De 1975 à 1993	64	670 275	2,1	5,7	7,3	16,1	33,7	14,5	9
$IC_{95\%}$			1,9-2,2	2,5-7	41614	8,9-27,3	21,5-894	9,7-19,2	6,9-11,7
À partir de 1994	115	931 378	< 1	2	2,7	5,7	24,2	5,8	3,5
$IC_{95\%}$			-	1,6-2,2	2,2-3,9	3,4-8,9	11,2-59	3,9-7,8	2,8-4,2

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique.

indiquent que les niveaux moyens en milieu rural sont inférieurs de l'ordre de 28 % à ceux en milieu urbain.

TABLE 15 – Distribution des charges moyennes en plomb total et acido-soluble dans la poussière intérieure déposée au sol (μg/m^2) selon l'environnement.

Stat.	*n*	*N*	P5	P25	P50	P75	P95	m.a.	m.g.
Pb total									
Urbain	298	2 541 085	< 2	3,9	9,6	18,6	69,9	19,8	9,6
$IC_{95\%}$			-	2,8-6,2	7,9-12,3	15-26,2	51,3-93,6	15,6-24,1	7,9-11,6
Rural	157	841 291	< 2	3,1	6,7	10	60,2	15,6	6,9
$IC_{95\%}$			-	2,2-3,8	4,6-8,2	8,2-22	36,7-694,8	10,6-20,6	5,6-8,3
Pb a.-s.									
Urbain	302	2 554 818	1,4	3,3	7,3	15,4	47,3	14,8	7,5
$IC_{95\%}$			< 1-1,9	2,2-5,2	5,9-10,5	12,2-18,2	30,6-84,4	11,4-18,2	6,2-9,1
Rural	158	850 179	1,4	2,5	4,3	9,3	37,9	10,3	5,3
$IC_{95\%}$			< 1-1,5	1,9-3,1	3,2-5,7	6,4-16,7	37,9	7,5-13,1	4,3-6,5

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

Niveaux en plomb en parties communes

La table 16 présente la distribution des charges moyennes des parties communes pour les logements avec parties communes, en plomb total et acido-soluble. Cette charge moyenne a été calculée comme la moyenne arithmétique des charges disponibles en parties communes (2 mesures au plus : sur le palier et dans le hall de l'immeuble). La table 16 permet de constater que les niveaux en plomb dans les poussières des parties communes sont substantiellement plus élevés qu'à l'intérieur des logements. En rapport de moyennes géométriques, les niveaux sont environ 73 % plus élevés en parties communes. En plomb total, la charge mesurée la plus grande dans un prélèvement à l'intérieur était de 491 μg/m^2 contre 6271 μg/m^2 en parties

communes. En plomb acido-soluble les mesures respectives étaient de 441 μg/m^2 contre 5968 μg/m^2.

TABLE 16 – Distribution des charges moyennes en plomb total et acido-soluble (μg/m^2) dans la poussière déposée au sol pour les logements avec parties communes.

	Intérieur		**Parties communes**	
Statistique	**Estimation**	**IC$_{95\%}$**	**Estimation**	**IC$_{95\%}$**
Pb total				
n	113		114	
N	984 599		986 565	
P5	< 2	< 2-2,3	6,5	< 2-10
P25	4,8	1,9-7,8	18	10-22,5
P50	9,3	6,9-10,5	25,2	22,5-31,8
P75	14,4	10-24,3	43,7	31,5-98,8
P95	51	30,7-81,7	384,1	115,4-1574,6
m.a.	13,7	9,7-17,6	128,2	13,1-243,4
m.g	8,6	6,4-11,5	32,2	26,3-39,4
Pb a.-s.				
n	114		115	
N	986 565		991 818	
P5	1,3	< 1-2	6	2-7,9
P25	3,7	2-6,1	14,5	7,9-20,4
P50	7,3	5,8-11	23,5	20-28,3
P75	13,5	9,1-18,4	35,8	28,5-98,8
P95	28,7	23,3-252,6	362,2	102,3-1233,1
m.a.	10,7	8,2-13,2	117,2	9,2-225,2
m.g	7,1	5,3-9,6	27,5	22,4-33,8

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

La table 17 et la table 18 montrent respectivement les distributions des charges moyennes en plomb total et acido-soluble par période de construction. Contrairement à ce qui avait été observé pour l'intérieur des logements (tables 13 et 14 respectivement), les niveaux en plomb décroissent entre la période « avant 1949 » et la période 1974-1993. La moyenne géométrique pour la période « avant 1949 » a un ordre de grandeur bien plus important que celui des moyennes géométriques des 3 autres périodes de construction. En rapport de moyennes géométriques, les niveaux en plomb total dans la période de construction « avant 1949 » sont 70 % plus élevés que dans la période 1949-1974, 82 % plus élevés que dans la période 1974-1993 et 89 % plus élevés que dans la période « à partir de 1994 ». En outre la dispersion des charges est plus importante pour la période de construction « avant 1949 ».

Seuls 10 logements en immeuble collectif ont été enquêtés en milieu rural et parmi eux, seuls 7 logements ont une valeur en plomb dans la poussière des parties

TABLE 17 – Distribution des charges moyennes en plomb total dans la poussière déposée au sol en parties communes ($\mu g/m^2$).

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Avant 1949	25	173 502	10	21,7	95,4	392,3	2087,8	569,5	105
$IC_{95\%}$			10-46	10-141,7	10-448	94,7-6271	551,8-6271	-	30-366,7
De 1949 à 1974	57	489 719	8,4	22,1	31,1	46,1	115,3	40	31,7
$IC_{95\%}$			7,7-9	13,6-24	22,3-36,1	31,4-115,1	46,7-379	25,8-54,2	24,7-40,7
De 1975 à 1993	18	250 460	6,5	12	21	24,5	38	27,4	19,6
$IC_{95\%}$			6,5-10,8	6,5-18,6	7,2-29,1	15,6-335	27,4-335	15-39,9	14,5-26,6
À partir de 1994	14	72 884	< 2	4,5	9,1	24,5	38,4	17	11,8
$IC_{95\%}$			-	2,1-7,6	3-32,9	7,1-51	26,7-51	7,3-26,8	6,2-22,4

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique.

TABLE 18 – Distribution des charges moyennes en plomb acido-soluble dans la poussière déposée au sol en parties communes ($\mu g/m^2$).

Stat.	n	N	P5	P25	P50	P75	P95	m.a.	m.g.
Avant 1949	25	173 502	8,5	19,7	72,8	362,7	1930,4	531,9	92,6
$IC_{95\%}$			8,5-41,2	8,5-128,5	8,5-397,5	88,9-3929,7	508,4-5968	-	26-330
De 1949 à 1974	58	494 972	7,5	19,5	22,9	34,8	100,6	34,1	26,5
$IC_{95\%}$			7,2-7,9	9,1-20,9	20,6-28,5	28-99,6	37,8-353	21,6-46,5	20,6-34,1
De 1975 à 1993	18	250 460	6	9,6	16,6	22,5	32,5	24,2	17
$IC_{95\%}$			6-8,7	6-14,1	6-26,4	12,5-308	24,4-308	12,8-35,7	12,2-23,7
À partir de 1994	14	72 884	2,8	4,5	7,2	20,3	30,1	14,5	10,5
$IC_{95\%}$			2,1-3,4	3,1-5,9	4-28,1	5,5-47	22,3-47	6,4-22,5	5,6-19,5

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique.

communes qui leur est associée. Une analyse descriptive par type d'environnement n'aurait pas eu de sens. Les charges moyennes en parties communes sont les suivantes :

- Plomb total : 5 ; 9 ; 10,5 ; 19,5 ; 46,5 ; 52,5 et 115,5 $\mu g/m^2$
- Plomb acido-soluble : 4 ; 7,5 ; 7,5 ; 13 ; 25 ; 46,5 et 101,5 $\mu g/m^2$

Comparaison aux valeurs de référence

Pour rappel à ce sujet, voir la section 3.5 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel ».
Dans approximativement 0,21 % ($IC_{95\%}$ = 0-0,5) des logements le standard américain de 430,5 $\mu g/m^2$ est dépassé. Approximativement 4,1 % ($IC_{95\%}$ = 0-8,1) des charges moyennes en parties communes dépassent ce même standard.

2.3 Revêtement intérieur

Par logement ont été comptées les UD dont le revêtement contenait du plomb au sens de la réglementation, c'est-à-dire contenant une charge en plomb surfacique supérieure ou égale à 1 mg/cm^2. La table 19 montre la répartition de ces UD selon l'état de dégradation du revêtement. La prévalence de logements avec de la peinture au plomb dans les revêtements est égale approximativement à 24,5 % ($IC_{95\%}$ = 18,7-30,2). Cette prévalence est de 34,2 % ($IC_{95\%}$ = 16,3-52) en ce qui concerne les parties communes (table relative aux parties communes non montrée).
Les UD incriminées sont majoritairement accessibles à l'enfant, c'est-à-dire située à moins d'un mètre du sol.

D'un point de vue du risque plomb, les UD à revêtements dégradés sont les plus problématiques. Si on définit le risque plomb lié aux revêtements intérieurs comme étant le fait d'avoir au moins une UD dont le revêtement contient du plomb et est dégradé, alors 4,7 % ($IC_{95\%}$ = 2,4-6,7) des logements possèdent un tel risque. Pour les parties communes, cette prévalence est de 7,1 % ($IC_{95\%}$ = 1,8-12,4) (table relative aux parties communes non montrée). Aucun logement n'est estimé avoir strictement plus de 10 UD à revêtement dégradé ; les logements possédant de 6 à 10 UD dont le revêtement est dégradé sont très peu prévalents.

TABLE 19 – Répartition des logements selon le nombre d'unités de diagnostic (UD) par catégorie de dégradation des revêtements (tout support confondu).

Nb. UD incriminées pour un logement	**≥ 1mg/cm²**	**≥ 1mg/cm² + (EU ou D)**	**≥ 1mg/cm² + D**
0 UD	2 705 113 75,5% (69,8-81,3)	3 098 982 86,5% (82,1-90,9)	3 413 291 95,3% (93-97,5)
1 à 5 UD	606 610 16,9% (11,7-22,2)	414 477 11,6% (7,3-15,9)	168 505 4,7% (2,5-7)
et accessible(s) à l'enfant	542 109 15,1% (10-20,2)	353 048 9,9% (5,3-14,4)	118 663 3,3% (1,2-5,4)
6 à 10 UD	209 644 5,9% (2,2-9,5)	61 234 1,7% (0,1-3,4)	195 ∼0% (-)
et accessibles à l'enfant	192 961 5,4% (1,9-8,9)	53 790 1,5% (0-3,2)	0 0
> 10 UD	60 624 1,7% (0,1-3,3)	7 298 0,2% (0-0,5)	0 0
et accessibles à l'enfant	47 387 1,3% (0-2,9)	7 298 0,2% (0-0,5)	0 0

Légende. EU : état d'usage, D : dégradé, (... – ...) : intervalle de confiance à 95%.

La réglementation française visant principalement la céruse, la table 20 montre la répartition des UD dont le support était non métallique, afin de décrire autant que possible la prévalence de ce composé du plomb. Environ 3,3 % ($IC_{95\%}$ = 1,2-5,4) des logements possèdent au moins une UD sur support métallique contenant du plomb

et dont le revêtement est dégradé. Cette prévalence est de 7 % ($IC_{95\%}$ = 1,8-12,2) pour les parties communes (table relative aux parties communes non montrée).

TABLE 20 – Répartition des logements selon le nombre d'unités de diagnostic (UD) par catégorie de dégradation des revêtements (support non métallique).

Nb. UD incriminées pour un logement	**≥ 1mg/cm²**	**≥ 1mg/cm² + (EU ou D)**	**≥ 1mg/cm² + D**
0 UD	2 892 534 80,7% (75,7-86)	3 238 359 90,4% (87-93,7)	3 462 647 96,7% (94,5-98,8)
1 à 5 UD	486 337 13,6% (8,4-18,8)	281 175 7,8% (4,5-11,2)	119 149 3,3% (1,2-5,4)
et accessible(s) à l'enfant	453 415 12,6 (7,5-17,7)	250 804 7% (3,6-10,4)	105 887 2,9% (0,9-5)
6 à 10 UD	152 103 4,2% (1,3-7,1)	55 159 1,5% (0-3,2)	195 ∼0%(-)
et accessibles à l'enfant	146 109 4,1% (1,2-7)	53 790 1,5% (0-3,2)	0 0
> 10 UD	51 018 1,3% (0-2,9)	7 298 0,2% (0-0,5)	0 0
et accessibles à l'enfant	47 387 1,3% (0-2,9)	7 298 0,2% (0-0,5)	0 0

Légende. EU : état d'usage, D : dégradé, (... – ...) : intervalle de confiance à 95%.

Enfin les UD sur support métallique ont d'autre part été comptabilisées pour viser un autre composé du plomb très répandu, le minium. La table 21 affiche les résultats relatifs. Les UD sur support métallique sont moins fréquentes dans les logements que les UD sur support métallique. La table 21 n'affiche donc les résultats que pour « au moins une UD incriminée ». Les logements sont près de 11 % à posséder au moins une UD sur support métallique dont le revêtement contient du plomb. Cette prévalence est de 17,7 % ($IC_{95\%}$ = 2,2-33,3) pour les parties communes (table relative aux parties communes non montrée). En ce qui concerne le risque plomb lié à ces UD sur support métallique, la prévalence des logements avec un tel risque est de 1,4 % et pour les parties communes cette prévalence est de 0,3 % ($IC_{95\%}$ = 0-0,8) (table relative aux parties communes non montrée).

Concernant la prévalence des peintures à base de plomb et plus particulièrement de céruse, la figure 26 montre le nombre d'UD sur support non métallique en fonction de la date de construction du logement. Sur cette figure, chaque point représente un logement ; plus le point est gros plus il représente de logements dans la population. Plus l'âge du logement augmente plus la dispersion de ce nombre d'UD croît ; jusqu'à 25 UD peuvent être observées pour les logements les plus anciens. Dans une moindre mesure mais de manière non négligeable, la prévalence des revêtements sur support non métallique semble relativement importante pour la période 1949-1974 : jusqu'à 12 UD incriminées sont observées. Pour les logements plus récents, à compter de 1975, les UD sur support non métallique dont le revêtement contient du plomb sont quasi inexistantes.

TABLE 21 – Répartition des logements selon le nombre d'unités de diagnostic (UD) par catégorie de dégradation des revêtements (support métallique).

Nb. UD incriminées pour un logement	**≥ 1mg/cm²**	**≥ 1mg/cm² + (EU ou D)**	**≥ 1mg/cm² + D**
0 UD	3 208 740 89,6% (84,8-94,3)	3 382 190 94,4% (91,3-97,5)	3 532 636 98,6% (97,5-99,7)
≥ 1 UD	373 251 10,4% (5,6-15,2)	199 801 5,6% (2,4-8,7)	49 356 1,4% (0,2-2,5)
et accessible(s) à l'enfant	269 806 7,5% (3,8-11,2)	135 767 3,8% (1,2-6,3)	12 776 0,3% (0-0,7)

Légende. EU : état d'usage, D : dégradé, (... – ...) : intervalle de confiance à 95%.

FIGURE 26 – Distribution du nombre d'unités de diagnostic (UD) ≥ 1 mg/cm² sur support non métallique, selon la période de construction.

Afin de savoir si les UD sur support non métallique contenant du plomb sont fortement prévalentes ou non à l'intérieur de chaque logement, la table 22 montre la prévalence des logements contenant au moins p UD dont le revêtement contient au moins 1 mg/cm², où p varie de 1 à 10. La table se lit de la manière suivante : 50,2 % des les logements construits avant 1949 possède au moins une UD dont le revêtement contient au moins 1 mg/cm². Pour les logements construits avant 1949 la prévalence passe donc de plus de 50 % à 35 % lorsque l'on considère au moins une UD et au moins deux UD incriminées respectivement. Pour les logements de la période 1949-1974 la prévalence diminue de plus de moitié, passant de 22 % à moins de 10 %. Cette prévalence est presque constante jusqu'au critère « au moins 5 UD ».

En revanche, elle est proche de zéro au delà de 10 UD pour cette même période de construction.

TABLE 22 – Prévalence (%) de logements possédant un nombre d'unité de diagnostic (UD) à charge en plomb $\geq$ 1 mg/cm^2 selon la période de construction. Support non métallique.

	≥1UD	**IC$_{95\%}$**	**≥2UD**	**IC$_{95\%}$**	**≥3UD**	**IC$_{95\%}$**	**≥4UD**	**IC$_{95\%}$**	**≥5UD**	**IC$_{95\%}$**	**≥10UD**	**IC$_{95\%}$**
Avant 1949	50,2	38,2-62,3	34,9	22,4-47,4	31,8	19,5-44,1	22,7	11,6-33,8	16,5	7,5-25,5	6,6	0-13,5
De 1949 à 1974	22,1	8,5-35,7	9,2	0,7-17,8	8,7	0,1-17,3	8,5	0-17,1	8,2	0-16,7	0,4	0-1,1
De 1975 à 1993	1,8	0-4,6	0	-	0	-	0	-	0	-	0	-
À partir de 1994	0,1	0-0,3	0	-	0	-	0	-	0	-	0	-

Afin de principalement détecter de la céruse, la table 23 montre les prévalences estimées pour un seuil de 2 mg/cm^2. Les charges autour de 1 mg/cm^2 peuvent indiquer d'autres composés du plomb, comme des siccatifs par exemple. Ce nouveau seuil impacte principalement et substantiellement les prévalences des logements construits entre 1949 et 1974, en abaissant fortement ces prévalences.

TABLE 23 – Prévalence (%) de logements possédant un nombre d'unité de diagnostic (UD) à charge en plomb $\geq$ 2 mg/cm^2 selon la période de construction. Support non métallique.

	≥1UD	**IC$_{95\%}$**	**≥2UD**	**IC$_{95\%}$**	**≥3UD**	**IC$_{95\%}$**	**≥4UD**	**IC$_{95\%}$**	**≥5UD**	**IC$_{95\%}$**	**≥10UD**	**IC$_{95\%}$**
Avant 1949	37,9	25-50,8	23,1	13,5-32,7	20,8	11,4-30,2	17	8,3-25,6	11,6	5,6-17,5	3,2	0-6,4
De 1949 à 1974	13,2	0,1-26,4	3,7	0-8,1	3,3	0-7,6	1,8	0-4,7	1,8	0-4,7	0	-
De 1975 à 1993	1,2	0-3,7	0	-	0	-	0	-	0	-	0	-
À partir de 1994	0,1	0-0,3	0	-	0	-	0	-	0	-	0	-

2.4 Aire de jeu extérieure de l'enfant

Distribution des charges en plomb

Le prélèvement sur l'aire de jeu extérieure principale de l'enfant n'a été fait que si l'enfant jouait à l'extérieur. Les enfants sans aire de jeu extérieure sont par exemple des enfants âgés de 6 mois à 6 ans, ne marchant pas encore. On rappelle qu'une seule aire de jeu a été échantillonnée le cas échéant : si l'enfant jouait sur un sol dur comme un balcon, une charge en plomb (μg/m^2) est alors disponible ; si l'enfant jouait sur un sol meuble comme une pelouse, une concentration en plomb (mg/kg) est disponible. Dans 84,8 % (IC$_{95\%}$ = 80,5-89,1) des logements, l'enfant joue à l'extérieur, ce qui correspond à 3 038 155 logements dans la population. 73 sols durs et 319 sols meubles auraient du être prélevés. Le non respect du protocole, les prélèvements non réalisables (par exemple sol gelé) et les pertes des certains colis avec les échantillons de sols, ont réduit les mesures disponibles à 53 sols durs et 315

sols meubles.

La table 24 montre les distributions des concentrations et charges en plomb respectivement, en plomb total et en plomb acido-soluble. Les histogrammes relatifs sont affichés en figures 27 et 28. Les charges en plomb sur sols durs sont approximativement 3,2 fois plus élevées à l'extérieur que les charges les plus élevées[5] de l'intérieur des logements (en rapport de moyennes géométriques).

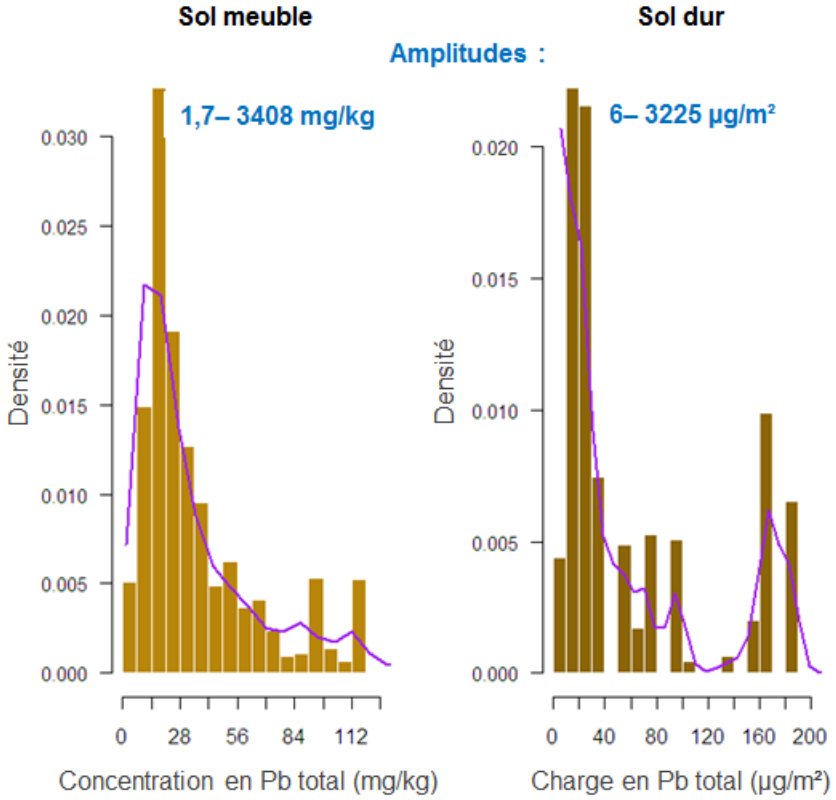
TABLE 24 – Distribution des concentrations et charges en plomb des aires de jeu extérieures des enfants.

	Sol meuble (mg/kg)		**Sol dur (μg/m²)**	
Statistique	**Estimation**	**IC$_{95\%}$**	**Estimation**	**IC$_{95\%}$**
Pb total				
n	315		53	
N	2 518 808		325 646	
P5	9,8	4,2-11,8	8,7	7,2-11
P25	17,3	15-18,9	17	11-23,6
P50	27,2	21,6-39,4	32,2	19-91
P75	60,2	42,7-93,2	99	39-373,1
P95	253,8	117,4-2174,5	393,2	187,1-3225
m.a.	73,6	38-109,3	96	48,2-143,7
m.g	33,9	27-42,6	44,4	28,3-69,7
Pb a.-s.				
n	315		53	
N	2 518 808		325 646	
P5	4,8	1,7-6,7	7,6	5,5-9
P25	10	8,1-12,3	12	9-21
P50	16,7	14,5-26	21	17,7-86,8
P75	42,3	29,5-65,6	94,1	29,2-369,7
P95	243,2	98,4-2029,4	352,4	141,2-3172
m.a.	58,2	26,2-90,1	78,5	43,1-113,8
m.g	21,7	16,9-27,9	36,9	23,9-56,6

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne artihmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

5. La distribution des charges maximales en plomb par logement dans les poussières intérieures déposées au sol n'a pas été montrée.

FIGURE 27 – Distribution des niveaux en plomb total de l'aire de jeu extérieure de l'enfant.



Les tables 25 et 26 montrent respectivement les distributions des concentrations et des charges en plomb des aires de jeu extérieure selon l'environnement urbain ou rural des logements.

FIGURE 28 – Distribution des niveaux en plomb acido-soluble de l'aire de jeu extérieure de l'enfant.

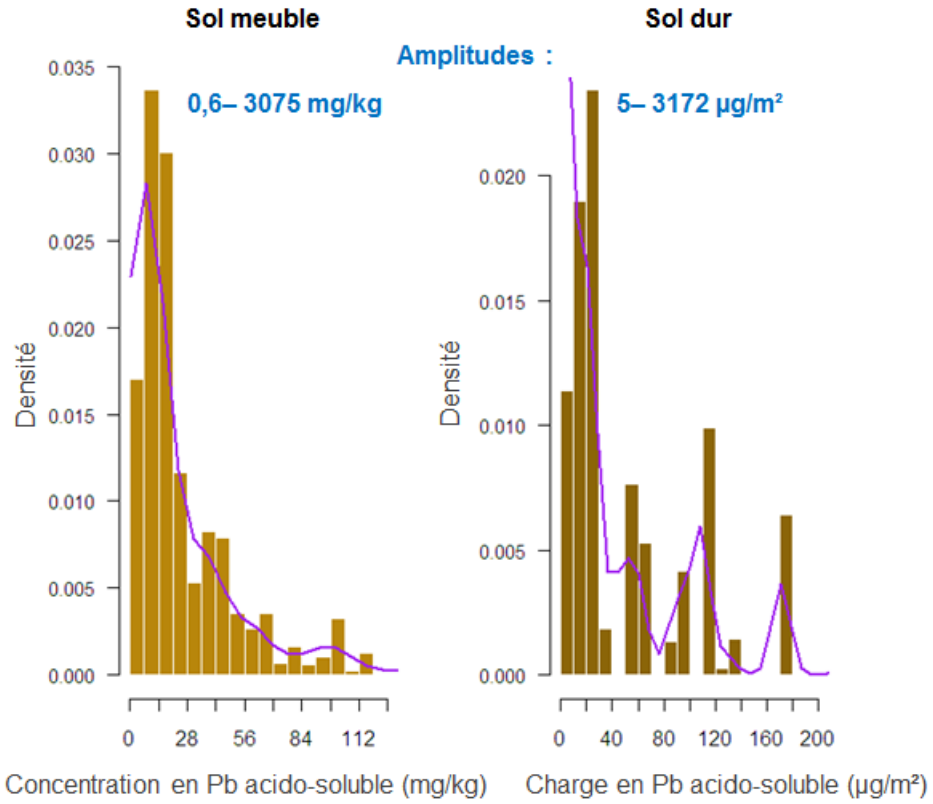


TABLE 25 – Distribution des concentrations en plomb (mg/kg) des aires de jeu extérieures sur sol meuble selon l'environnement extérieur.

	Urbain		**Rural**	
Statistique	**Estimation**	**$IC_{95\%}$**	**Estimation**	**$IC_{95\%}$**
Pb total				
n	202		113	
N	1 865 437		653 371	
P5	10,2	4,3-11,8	7,5	2,4-11,6
P25	18,2	15-23,8	14,6	10,6-17,1
P50	35,9	23,6-54,1	18,8	16,2-23,3
P75	82,8	53,3-117,6	30,3	22,5-41,7
P95	261,7	118,4-3408	53,4	42,7-192,5
m.a.	90,1	42-138,2	26,7	21,2-32,1
m.g	40	29,9-53,5	21,2	18-24,9
Pb a.-s.				
n	202		113	
N	1 865 437		653 371	
P5	4,1	0,9-6,8	4,7	1,7-6,3
P25	12,4	7,2-16	8,6	4,9-9,8
P50	22	15,8-40,5	11,6	8,9-12,9
P75	53,2	42,1-79,5	16,4	12,4-29,2
P95	243,5	100,6-3075	39	31,2-88
m.a.	72,8	29,7-115,8	16,4	12,5-20,3
m.g	26,4	19,1-36,4	12,4	10,4-14,7

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

TABLE 26 – Distribution des charges en plomb ($\mu g/m^2$) des aires de jeu extérieures sur sol dur selon l'environnement extérieur.

	Urbain		Rural	
Statistique	**Estimation**	**$IC_{95\%}$**	**Estimation**	**$IC_{95\%}$**
Pb total				
n	34		19	
N	226 383		99 264	
P5	8,4	7-9,4	14,7	6-15,7
P25	12,2	9,7-29	22,8	6-23,6
P50	52,6	12,7-137,2	23,5	20,6-43,7
P75	158,2	53-2057,2	26,3	23,4-484
P95	561,1	171,4-3225	271,9	59,1-484
m.a.	111,1	54,1-168	61,5	10,9-112
m.g	50,5	29,1-87,5	33,2	21,2-52,1
Pb a.-s.				
n	34		19	
N	226 383		99 264	
P5	7,4	6-8,5	8,2	5-9,8
P25	10,8	8,8-19,9	18,8	5-20,5
P50	43,2	11,2-99,1	20,4	17,4-46,2
P75	104	44,5-418,9	21	20,3-488
P95	362,3	138,4-3172	227,4	52,5-488
m.a.	90	48,4-131,6	52,2	7,1-97,4
m.g	42,1	24,6-72,1	27,2	17,8-41,5

Légende. Px : Percentile d'ordre x%, n : nombre de logements dans l'échantillon avec une donnée disponible, N : nombre de logements représentés, m.a. : moyenne arithmétique, m.g. : moyenne géométrique, a.-s. : acido-soluble.

Pour les sols meubles, toutes les statistiques estimées sont plus élevées en milieu urbain ; en rapport de moyennes géométriques les niveaux sont plus élevés de 47 % en plomb total et de 53 % en plomb acido-soluble.
Pour les sols durs, les estimations se basent sur des effectifs nettement plus faibles que ceux des sols meubles. Les niveaux en milieu urbain semblent être supérieurs mais les percentiles d'ordre 5 % et 25 % sont par contre plus élevés en milieu rural.

Comparaison aux valeurs de référence

Pour rappel à ce sujet, voir la section 3.5 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel ».
Approximativement 1,4 % ($IC_{95\%}$ = 0-2,9) des aires de jeu sur sol meuble ont une concentration supérieure ou égale au seuil américain de 400 mg/kg.

3 Synthèse

Pour la première fois un état de la contamination par le plomb en milieu résidentiel en France a été réalisé. Il concerne l'eau du robinet, la poussière

déposée au sol à l'intérieur des logements et en parties communes, les revêtements intérieurs ainsi que les aires de jeu extérieures. Cet état concerne les 3,6 millions résidences principales en France métropolitaine abritant au moins un enfant âgé de 6 mois à 6 ans.

Les niveaux en plomb exprimés en moyenne géométrique sont de :

- **moins de 1 µg/L** pour l'eau du robinet ;
- **8,8 µg/m²** en plomb total et **6,9 µg/m²** en plomb acido-soluble pour la poussière intérieure déposée au sol ;
- **32,1 µg/m²** en plomb total et **27,5 µg/m²** en plomb acido-soluble pour la poussière déposée au sol en parties communes ;
- **33,9 mg/kg** en plomb total et **21,7 mg/kg** en plomb acido-soluble pour les sols meubles de l'aire de jeu extérieure ;
- **44,4 µg/m²** en plomb total et **36,9 µg/m²** en plomb acido-soluble pour les sols durs de l'aire de jeu extérieure.

Les niveaux en plomb ainsi que les prévalences de dépassement de seuils réglementaires européens ou américains, définissant à ce jour un risque plomb, peuvent paraître faibles. Néanmoins, traduits en termes de nombres de logements incriminés, ils indiquent que :

- Approximativement **105 000 logements (2,9 %)** de la population décrite possèdent un niveau en plomb dans l'eau du robinet supérieur ou égal au seuil réglementaire européen égal à 10 μg/L qui sera en vigueur le 25 décembre 2013.
- Dans environ **7500 logements (0,21 %)** et **45 000 parties communes d'immeubles collectifs (4,1 %)**, la charge moyenne en plomb dans les poussières déposées au sol dépasse le seuil réglementaire américain actuellement en vigueur et égal à 430,5 μg/m^2 en plomb total.
- Approximativement **37 000 aires de jeu extérieures sur sol meuble (1,4 %)** ont une concentration en plomb total supérieure au seuil américain de 400 mg/kg.
- Environ **878 000 logements (24,5 %)** possèdent encore au moins une unité de diagnostic dont le revêtement contient au minimum 1 mg/cm^2 de plomb.

D'autre part les résultats permettent de souligner que :

- Les niveaux en plomb dans la poussière intérieure déposée au sol sont approximativement de même ordre de grandeur dans les logements construits avant 1949, ceux construits entre 1949 et 1974 et les logements construits entre 1975 et 1993. Les niveaux sont par contre plus faibles dans les logements construits après 1993.
- **En parties communes**, les charges en plomb dans la poussière déposée au sol sont 73 % plus élevées que les charges à l'intérieur des logements. Dans les parties communes les niveaux croissent avec l'âge du bâtiment.

- Environ **34,2 % des parties communes** possèdent des revêtements à base de plomb.
- Approximativement **4,7 % des logements** possèdent des revêtements dégradés à base de plomb. Cette prévalence est de **7,1 % pour les parties communes**.
- Environ **22,1 % des logements** construits entre 1949 et 1974 possèdent au moins une unité de diagnostic dont le revêtement possède une charge en plomb d'au moins 1 mg/cm^2 sur support non métallique (céruse visée). Cette prévalence devient égale à **3,7 %** lorsque au moins deux unités de diagnostic avec une charge d'au minimum 2 mg/cm^2 sont considérées.
- Les niveaux en plomb dans la poussière extérieure sont d'environ 3,2 fois plus élevés que les niveaux maximaux dans la poussière intérieure.

Cet état descriptif de la contamination par le plomb peut permettra de mettre à jour l'analyse coût-bénéfice réalisée en France en 2011, concernant les moyens de réduction des expositions et l'impact sur les niveaux des plombémies chez l'enfant [Pichery et al., 2011].

Cet état de la contamination servira dorénavant de référence pour les futures études évaluant les niveaux de contamination par le plomb en France en milieu résidentiel.

À ce stade, les niveaux en plomb ont été décrits dans 5 compartiments environnementaux du milieu résidentiel occupé par les enfants. Dans la suite un focus est réalisé sur les poussières intérieures déposées au sol puisque :

- d'une part l'effet du plomb sur l'organisme semble être sans seuil ;
- d'autre part ces poussières ont été incriminées à travers la littérature comme étant la principale voie d'exposition au plomb chez l'enfant.

Dans le chapitre suivant, la provenance du plomb dans ces poussières intérieures déposées au sol est étudiée.

Chapitre 3

Estimation de la contribution des sources en plomb à contaminer la poussière intérieure déposée au sol

Après s'être intéressé dans le chapitre précédent à la description des niveaux en plomb en particulier dans la poussière intérieure déposée au sol, dans ce chapitre on s'intéresse aux sources de la contamination par le plomb de ces poussières.

Des études ont montré que la poussière intérieure était le principal prédicteur des niveaux de plombémies chez l'enfant [Lanphear et al., 1996, Lanphear et al., 1998, Lanphear, 2002]. Dès lors il paraît important d'évaluer le rôle des sources potentielles en plomb pouvant contaminer la poussière intérieure. La connaissance de la part attribuable à chacune des sources contaminant la poussière permettra aux pouvoirs publics de mettre en place des actions de réduction de ces sources. Ainsi en jouant sur ces leviers, l'exposition au plomb de l'enfant via les poussières intérieures contaminées par le plomb pourra être réduite puisque le niveau en plomb dans ces poussières aura été abaissé.

Pour évaluer la contribution respective des sources en plomb à contaminer la poussière, un modèle explicatif des niveaux en plomb dans la poussière a été construit. Sachant que la variable réponse du modèle est construite à partir des charges en plomb dans la poussière des pièces d'un même logement, les composantes de la variable réponse ne sont pas indépendantes. Afin de prendre en compte cette corrélation, une modélisation multi-niveaux autrement appelée modélisation mixte ou hiérarchique en statistiques classiques, a été utilisée (cf. section 6 de la partie « Spécificités des données d'enquête »). Les niveaux ont été constitués des pièces (niveau 1) et des logements (niveau 2) ; le modèle dans ce cas est un modèle dit à 2 niveaux. Cette modélisation permet de plus d'estimer la corrélation entre 2 mesures du plomb dans la poussière d'un même logement, corrélation qui était d'intérêt due au type de prélèvements de poussière réalisés (cf. section 3.2 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel »). La question est de savoir si un seul prélèvement de poussière par lingette peut être suffisant pour

évaluer le niveau en plomb des poussières d'un logement.

Afin de construire le modèle, certaines informations disponibles dans la base de données ont été sélectionnées à partir de la littérature notamment. Des variables ont été agrégées afin de construire des variables qui ont été *in fine* incorporées dans le modèle. D'autres informations ont été de plus collectées à partir d'autres bases de données publiques lorsque ces informations n'étaient pas disponibles dans la base de données de l'enquête Plomb-Habitat.

Afin de prendre compte le fait que les données utilisées dans le modèle étaient des données d'enquête, un modèle avec différentes pondérations pour le niveau des logements (niveau 2) a été ajustée (cf. section 3 du présent chapitre) pour étudier l'impact de ces pondérations sur les estimations. La problématique du poids des pièces (niveau 1) n'existait pas à partir de nos données dans la mesure où les pièces enquêtées au sein d'un logement n'ont pas été tirées aléatoirement mais automatiquement investiguées.

Puisque les données manquantes présentes à travers les variables utilisées dans le modèle induisaient une perte substantielle ($> 10\%$) d'observations (pièces) dans la table de données, et induisaient donc potentiellement des estimateurs biaisés, un traitement des données manquantes par imputation multiple a été réalisé en section 4 du présent chapitre.

Dès lors, un modèle à 2 niveaux non pondéré a été ajusté pour répondre à l'objectif de ce chapitre. Un choix de présentation des résultats relatifs aux contributions de chaque variable jugée comme source en plomb contaminant la poussière intérieure a été fait afin de valoriser les résultats auprès des pouvoirs publics comme de la communauté scientifique.

Un article scientifique relatif au travail de ce chapitre, intitulé *« Source Contributions of Lead in Residential Floor Dust and Within-Home Variability of Dust Lead Loading »*, a été publié dans le journal *« Science of the Total Environment »* [Lucas et al., 2014].

1 Choix du type de modélisation

Afin d'expliquer la variation des charges en plomb dans la poussière déposée au sol, un modèle à but d'estimer les effets (*Models for Effect Estimation* [Harrell, 2001, page 82]) est ici construit. À partir des effets estimés de chacune des sources en plomb, la contribution de chacune de ces sources à contaminer la poussière intérieure déposée au sol peut être calculée.

Le caractère d'intérêt, Y, est ici la charge en plomb ($\mu g/m^2$) de la poussière intérieure déposée au sol. La charge en plomb de la pièce i au sein du logement j s'identifie par y_{ij}. On verra plus tard que ce sera une transformation de Y qui sera

modélisée et non Y directement (cf. section 2.2 de ce chapitre).

Puisque des modélisations classiques comme la régression linéaire multiple requièrent une hypothèse d'indépendance des observations (les pièces ici), une telle modélisation n'est *a priori* pas adaptée à nos données. En effet les charges en plomb de plusieurs pièces d'un même logement sont *a priori* corrélées. L'hypothèse sous-jacente faite est que, s'il y a contamination par le plomb des poussières d'une pièce, contamination provenant de l'intérieur du logement ou de l'extérieur, il est probable qu'une contamination se produise en parallèle pour une autre pièce du même logement. Dès lors, les pièces ne sont pas indépendantes sur le caractère Y.

D'autre part une modélisation du type régression linéaire, par définition ne concernant qu'un seul niveau d'information et ne permettant d'estimer que des effets fixes, n'était pas adaptée pour estimer la corrélation entre deux charges en plomb dans la poussière dans un même logement (appelée coefficient de corrélation intra-classe noté ρ. cf. la section 6.3 de la partie « Spécificités des données d'enquête »). Pour estimer cette corrélation, une modélisation à minima à 2 niveaux est requise avec l'introduction d'un effet aléatoire sur le niveau 2 (logement) (cf. sur ce sujet la section 6.3 de la partie « Spécificité des données d'enquête »).

Aucun effet aléatoire n'a été associée aux covariables car seul un effet (fixe) moyen était étudié pour chacune des sources en plomb. Dans la mesure où les 484 logements (le niveau 2) constituant l'échantillon ne nous intéressent pas en particulier et que, par nature même de l'échantillon issu d'un plan de sondage (aléatoire), ils doivent représenter la population de logements étudiée, la variabilité entre les logements est modélisée par un effet aléatoire ou, autrement dit, par un *« intercept »* aléatoire au niveau logement. C'est-à-dire qu'au lieu qu'un seul *« intercept »* soit estimé et vaut pour tous les logements, on suppose qu'un décalage en l'*« intercept »* (*shift*) existe entre les logements. Ceci est alors pris en compte par l'expression $\beta_0+\zeta_j$ de l'équation 6.6 page 51 où on rappelle que β_0 est une moyenne globale et ζ_j l'effet aléatoire au niveau 2 associé au logement j (cf. section 6.3 de la partie « Spécificités des données d'enquête »).

Un modèle a 2 niveaux, pièce en tant que niveau 1 et logement en tant que niveau 2, a été planifié dans la mesure où, d'une part seules des covariables donnant une information sur la pièce ou sur le logement ont été sélectionnées (voir section suivante,) et d'autre part il n'y avait pas d'intérêt particulier porté sur la part de variabilité expliquée par un niveau supérieur, en l'occurrence le niveau hôpital[1].

1. Il est néanmoins possible d'indiquer dans le logiciel (Stata V12) que des entités supérieures, les hôpitaux donc, englobent les éléments de niveau 2, les logements, sans que les hôpitaux soient considérés comme constituant un niveau. Cette option `vce(cluster id_des_hôpitaux)` permet d'améliorer l'estimateur de la variance des coefficients. Voir aussi la discussion sur les estimateurs robustes en page 199.

Pour plus de clarté on rappelle la définition du modèle à ajuster composé de 2 niveaux avec *« intercept »* aléatoire :

$$\text{Niveau 1} : y_{ij} = \beta_{0j} + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)} + \epsilon_{ij} \tag{1.1}$$

$$\text{Niveau 2} : \beta_{0j} = \beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \zeta_j \tag{1.2}$$

avec les perturbations notées $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_1^2)$, l'effet aléatoire au niveau 2 $\zeta_j \sim \mathcal{N}(0, \sigma_2^2)$ et avec $j = 1, \ldots, n^{(2)}$ et $i = 1, \ldots, n_j^{(1)}$. Les effets aléatoires ϵ_{ij} et ζ_j sont supposés indépendants.

2 Sélection des covariables et choix de la forme de la relation

2.1 Choix des sources en plomb et des variables de confusion

Les variables explicatives ou covariables entrant dans le modèle sont notées d'une manière générale par X. Les variables à introduire dans le modèle ont été classées en 2 catégories : les *sources* potentielles en plomb contaminant la poussière intérieure et les *variables de confusion*. On entend par variables de confusion, diverses raisons dont l'effet est sans intérêt pour la variable réponse Y, à travers desquelles X et Y seraient reliées [Lumley, 2010a].

La sélection des variables sources à introduire dans le modèle a été réalisée sur la base de la littérature du domaine : les variables pouvant contaminer en plomb la poussière intérieure, tout comme les variables jugées en lien avec la plombémie ont été sélectionnées dès lors qu'elles étaient disponibles dans la base de données Plomb-Habitat. Les études trouvées figurent en annexe 2. Le choix du lien avec la plombémie est justifié par le fait que les poussières contaminées contribuent au niveau en plomb dans le sang de l'enfant. Toutes les sources potentielles présentes à travers les données, devaient être testées et donc introduites dans le modèle puisque l'évaluation de leur effet respectif était l'objectif du modèle.
L'introduction dans le modèle de plusieurs sources de contamination dont l'effet est à estimer, ne doit néanmoins pas évincer la problématique de la multi-collinéarité d'une part et d'autre part de la réduction de la dimensionnalité si nécessaire.

Concernant la multi-collinéarité on pourra se reporter en annexe 3 pour le choix fait pour les variables « Basol », « Basias » et « Bdrep ». De plus pour les logements collectifs, seule la charge en plomb dans les poussières du palier d'appartement a été introduite dans le modèle, alors que celle du hall de l'immeuble était aussi disponible. L'hypothèse faite est que cette charge du palier devait être corrélée à la charge en plomb des poussières du hall de l'immeuble. Notons que pour certains logements,

une seule mesure en plomb des poussières des parties communes était disponible car le palier et le hall étaient confondus.

La sélection des variables de confusion a été faite selon le raisonnement suivant : Toutes les variables pouvant influencer la mesure en plomb renseignée à travers certaines variables sources doivent être introduites dans le modèle ; en particulier les influences dues à la méthode de prélèvement des poussières.
Toute information pouvant produire un « delta » sur Y sans que ces informations soient des sources, doivent être introduites dans le modèle. Par exemple, la saison d'enquête (donc de prélèvement) a été évaluée dans la littérature comme produisant des concentrations ou des charges en plomb supérieures en période estivale. Si on décidait que le 1er mai constituait un jour de référence de prélèvement, alors deux valeurs x_1^1 et x_1^2 d'une covariable X_1, relevées respectivement le 30 juin et le 31 décembre, seraient ramenées à des valeurs qui pourraient être égales à quelque chose de la forme $x_1^1 + g_1$ et $x_1^2 + g_2$, respectivement, où g_1 et $g_2 \in \mathbb{R}$. La prise en compte de la saison dans le modèle permet de ramener la mesure d'une source relative à une situation de référence.

Les sources potentielles ainsi que les variables de confusion retenues figurent en annexe 3. Les détails de construction de certaines variables agrégées ou issues d'autres bases de données sont fournis dans cette même annexe.
Sur ce sujet, la construction de certaines variables est à souligner. La construction des variables « Basol », « Basias » et « Bdrep » est basée sur l'hypothèse que, plus les sols pollués, les sites polluants et les usines émissives de plomb respectivement sont proches du logement, plus la contamination du logement doit être importante. C'est la raison pour laquelle une division par $1/d_q$ où d_q est la distance (en kilomètres) entre le logement et le q-ème site est utilisée dans la construction de chacune de ces 3 variables (voir aussi les éléments de discussion à ce sujet en section 3.2 de la partie « Discussion »).
Concernant la variable relative au tabagisme, bien que collectée sous forme catégorielle ordinale, elle est utilisée comme continue dans le modèle afin de réduire le nombre de coefficients à estimer (voir aussi sur ce sujet la section 3.2 de la partie « Discussion »).
Concernant les covariables relatives au niveaux en plomb de l'aire de jeu extérieure de l'enfant, sachant que pour chaque logement un seul type de prélèvement sur l'aire de jeu de l'enfant a été réalisé, par lingette (mesure en μg/m^2) ou par carottage (mesure en mg/kg), une de ces deux mesures est forcément vide, sans que ce soit délibérément une donnée manquante. Ces valeurs vides doivent être remplacées par une valeur non vide afin d'éviter que le logiciel comprenne ces valeurs comme données manquantes. Il a été choisi de les remplacer par zéro. L'hypothèse sous-jacente faite est que, jouant principalement sur un des deux types de sols, durs ou bien meubles, la contamination des poussières intérieures par le type de sol non prélevé est négligeable (voir à ce sujet la section 3.2 de la partie « Discussion »). Lorsque l'enfant n'avait pas d'aire de jeu extérieure, sur laquelle il passe assez de temps pour qu'elle soit considérée comme telle, la valeur zéro a été mise aux 2 covariables quantifiant le niveau en plomb pour les 2 types de sols. De part l'information portée par les covariables relatives introduites dans le modèle via une intéraction avec la fréquence de fréquentation de l'aire de jeu, seul le *« track-in »*, c'est-à-dire la

contamination par le fait d'apporter du sol ou de la poussière extérieurs contaminés à l'intérieur du logement, a été évalué.

2.2 Transformation de variables

La charge en plomb dans la poussière déposée au sol a une distribution dissymétrique à droite (*right-skewed*). Elle s'apparente dès lors plutôt à une distribution Log-normale (voir la section 2.2 du chapitre 2, même si dans cette section il s'agit de la moyenne des charges en plomb par logement). Une transformation logarithmique (népérienne notée Log) a été appliquée à la charge en plomb dans la poussière dans les pièces afin de respecter autant que possible l'hypothèse de normalité des erreurs (cf. section 6.3 de la partie « Spécificités des données d'enquête »).

Les covariables continues et en particulier les covariables renseignant sur le niveau en plomb dans les différents compartiments environnementaux ont aussi été transformées par un logarithme népérien. En effet une telle transformation s'est révélée efficace pour modéliser des données environnementales relatives au plomb dans le cadre d'une relation du type $\text{Log}(Y) = \sum \beta \text{Log}(X)$ [Jiang & Succop, 1996, Rust et al., 1997]. On pourra se reporter à la partie « Discussion », section 3.2, sur ce sujet.

Si une covariable n'avait pas de valeur nulle, la transformation $x \longmapsto \text{Log}(x)$ a été appliquée. Dans le cas contraire, la transformation $x \longmapsto \text{Log}(x+1)$ a été utilisée.

2.3 Comparaison de modèles

Certaines covariables, essentiellement des variables de confusion, ont été initialement sélectionnées pour entrer dans le modèle mais sans justification existante à partir de la littérature. L'hypothèse, quant à leur utilité dans le modèle, ne reposait donc que sur un *a priori*. Afin de vérifier leur utilité, le meilleur modèle a été sélectionné comme celui minimisant le critère BIC (*Bayesian information criterion*). La comparaison des modèles pré-spécifiés sur ce le critère BIC s'est faite sur cas complets.

Trois variables de confusion étaient concernées : la surface du palier (m^2), le nombre de problèmes d'humidité ou de nuisibles dans chaque pièce et la surface de la pièce (m^2). La sélection s'est faite à partir de modèles emboîtés pré-spécifiés : le modèle complet c'est-à-dire avec ces 3 variables, le modèle où la surface du palier a été enlevé, le modèle où les problèmes d'humidité ont été de plus enlevés et enfin le modèle où la surface de la pièce a de plus été enlevée. Les résultats du critère BIC sont montrés en table 27. À la vue des valeurs du BIC, les 3 variables n'ont pas été gardées.

TABLE 27 – Évolution du critère BIC selon les modèles emboîtés.

Critère	Modèle complet	Modèle complet \{surface du palier}	Modèle complet \{surface du palier ; problèmes d'humidité}	Modèle complet \{surface du palier ; problèmes d'humidité ; surface de la pièce}
Valeur du BIC	4489,949	4482,695	4475,315	4469,255

Au final, 23 covariables (13 quantitatives et 10 catégorielles) ont donc été conservées impliquant l'estimation de 33 paramètres dans un modèle à 2 niveau : 30 coefficients relatifs aux covariables, 1 *« intercept »* global et 2 paramètres de variance. Les 10 variables catégorielles induisent l'estimation de 17 coefficients de régression.

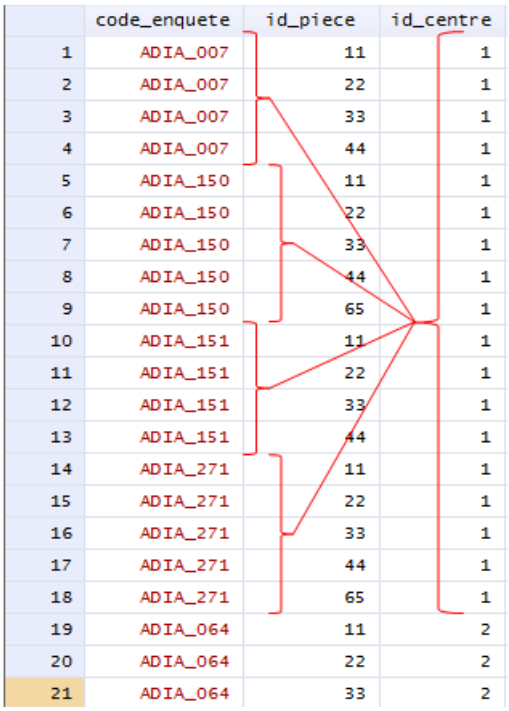
3 Application numérique : étude de l'impact des poids de niveau 2

Une fois la modélisation multi-niveaux choisie et les covariables sélectionnées, l'étape suivante consiste à procéder à l'estimation des paramètres du modèle. Puisque les données utilisées sont des données d'enquête, la question de la pondération à utiliser se pose pour ce type de modélisation (cf. section 6.3 de la partie « Spécificité des données d'enquête »).

Les unités du niveau 1, c'est-à-dire les pièces, ayant été investiguées sans tirage aléatoire préalable, leur probabilité de sélection conditionnelle (notée $\pi_{i|j}$) est simplement égale à 1 (rappel : voir section 8.2 de la partie « Spécificités des données d'enquête » à ce sujet). Les probabilités de sélection conditionnelles des unités de niveau 1 sont donc égales. La problématique des probabilités de sélection conditionnelles pour le niveau 1 souligné dans la littérature, n'a donc pas lieu d'être sur nos données (cf. section 6.3 de la partie « Spécificité des données d'enquête » sur ce sujet).

En revanche, étant donné que les logements, i.e. les unités du niveau 2, utilisés comme le plus haut niveau dans le modèle, ne correspondent pas aux entités du plus haut niveau possible (les hôpitaux en tant qu'unités hiérarchiquement supérieures), la question est de savoir quelle pondération il faut associer aux logements dans le modèle. En effet, à chaque identifiant d'hôpital (appelée « id_centre » dans la base) est associé plusieurs codes logement (appelée « code_enquete » dans la base) dans la table de données, support du modèle (voir figure 29). Si les logements avaient été échantillonnés en premier, leur poids de sondage aurait été leur poids conditionnel $w_j = 1/\pi_j$ où π_j est la probabilité de sélection du j^{e} logement. Ces w_j auraient donc été utilisés dans l'expression de la log-pseudo vraisemblance comme indiquée par l'équation 6.8 en page 54. Puisque les logements n'ont pas été échantillonnés en tout premier dans le plan de sondage, plusieurs type de poids peuvent être associés aux logement dans le modèle à 2 niveaux. Afin d'évaluer l'impact des différents poids candidats sur les estimations du modèle, une comparaison des estimations des coefficients de régression et des paramètres de variance est faite ici.

FIGURE 29 – Illustration de la hiérarchie des données.



	code_enquete	id_piece	id_centre
1	ADIA_007	11	1
2	ADIA_007	22	1
3	ADIA_007	33	1
4	ADIA_007	44	1
5	ADIA_150	11	1
6	ADIA_150	22	1
7	ADIA_150	33	1
8	ADIA_150	44	1
9	ADIA_150	65	1
10	ADIA_151	11	1
11	ADIA_151	22	1
12	ADIA_151	33	1
13	ADIA_151	44	1
14	ADIA_271	11	1
15	ADIA_271	22	1
16	ADIA_271	33	1
17	ADIA_271	44	1
18	ADIA_271	65	1
19	ADIA_064	11	2
20	ADIA_064	22	2
21	ADIA_064	33	2

Les différents type de poids identifiés pour le niveau 2 sont listés ci-après. Afin d'être clair, quelques rappels sur les notations utilisées pour décrire le plan de sondage (cf. section 8.2 de la partie « Spécificités des données d'enquête ») sont faits tout d'abord :

Les pièces sont indexées par la lettre i. Les logements (ou les enfants de la seconde phase) sont indexés par la lettre j et ont b comme exposant pour indiquer la seconde phase. Les enfants à l'hôpital sont indexés par le lettre j et ont a comme exposant pour indiquer la première phase. Les hôpitaux sont indexés par la lettre k.

La probabilité de sélection de l'hôpital k est notée π_k et son poids de sondage $w_k = 1/\pi_k$. La probabilité conditionnelle est la probabilité de sélection puisque les hôpitaux sont les éléments du premier degré de la phase 1.

La probabilité de sélection conditionnelle de l'enfant j dans l'hôpital k est notée $\pi_{j|k}$ et son poids de sondage conditionnel $w_{j|k} = 1/\pi_{j|k}$. Le poids de sondage de l'enfant j de l'hôpital k est par définition $w_j^a = 1/(\pi_k \times \pi_{j|k}) = 1/\pi_j^a$. Le poids de sondage post-stratifié est noté $\widetilde{w}_j^a$.

La probabilité pour qu'un logement j (ou de manière équivalente l'enfant qui l'occupe) figure dans l'enquête environnementale Plomb-Habitat est notée π_j^b.

Le poids de sondage du logement j est $w_j^b = \widetilde{w}_j^a/\pi_j^b = 1/\pi_j^b$ et le poids post-stratifié est noté $\widetilde{w}_j^b$.

La probabilité de sélection conditionnelle de la pièce i du logement j est notée $\pi_{i|j}$ et vaut ici 1 quelque soit i et j. Le poids de sondage conditionnel associé est noté $w_{i|j}$ et est égal par définition à $1/\pi_{i|j} = 1$. Le poids de sondage de la pièce i du logement j est $w_i = \widetilde{w}_j^b/\pi_{i|j} = \widetilde{w}_j^b$.

Pour un modèle à 2 niveaux décrits par les équations 1.1 et 1.2 page 116, les candidats pour être les poids de niveau 2, notés $w_j^{(2)}$ dans l'expression de la log-pseudo vraisemblance donnée par l'équation 6.8 en page 54, du fait de notre plan de sondage sont :

1. $\mathbf{w_1} : 1/\pi_j^b$
2. $\mathbf{w_2} : w_j^b$
3. $\mathbf{w_3} : \widetilde{w}_j^b$
4. $\mathbf{w_4} : 1/(\pi_j^a \times \pi_j^b)$
5. $\mathbf{w_5} : 1$
6. $\mathbf{w_6} : 1/(\pi_{j|k} \times \pi_j^b)$

où la notation « $\mathbf{w_1} : 1/\pi_j^b$ » indique le vecteur $\mathbf{w_1}$ dont les composantes sont $1/\pi_j^b$. On parlera dans la suite simplement des poids $\mathbf{w_1}$ pour indiquer les composantes des poids $\mathbf{w_1}$.

Les poids $\mathbf{w_1}$ peuvent être vus comme une sorte de poids conditionnels dans la mesure où π_j^b n'est pas la probabilité d'inclusion (finale) du logement j, bien que π_j^b ne soit pas une véritable probabilité conditionnelle puisqu'elle est située entre deux phases d'un plan de sondage et non entre deux degrés d'un plan de sondage. Les poids $\mathbf{w_2}$ sont les poids de sondage des logements c'est-à-dire les poids non post-stratifiés sur critères logements (période de construction, région et type de logement). Les poids $\mathbf{w_3}$ sont les poids de sondage finaux des logements ce que l'on entend par les poids post-stratifiés sur critères logements. Les poids $\mathbf{w_4}$ sont des poids de sondage sans aucune post-stratification, ni à l'étape de l'enfant, ni à l'étape du logement. Les poids $\mathbf{w_5}$ induisent une modélisation non pondérée. Les poids $\mathbf{w_6}$ peuvent être considérés comme des intermédiaires entre les poids $\mathbf{w_1}$ et $\mathbf{w_4}$ et leur présence est justifiée par le sous-échantillonnage entre l'enquête Saturn-Inf et l'enquête Plomb-Habitat ; les poids $\mathbf{w_6}$ peuvent être vus comme basés sur un produit de probabilités conditionnelles.

Bien qu'un modèle à seulement 2 niveaux ait été planifié, un modèle à 3 niveaux a aussi été testé. Ceci afin d'évaluer l'impact que pourrait avoir la déclaration des hôpitaux dans la pseudo vraisemblance comme unités hiérarchiquement supérieures aux logements. Dans ce cas les 2 phases du plan sont ignorées et le plan est alors traité comme un plan à 3 degrés (hôpitaux, logements, pièces) ; les enfants de la phase 1 ne peuvent pas être traités comme un niveau dans le modèle car à un enfant

de la phase 1 ne peut correspondre qu'un seul logement de la phase 2 à cause du sous-échantillonnage. Il n'y a donc pas de *« clusterisation »* [2] possible entre enfants et logements.

Pour la pondération à associer au niveau 3 (cf. équations 6.2, 6.3 et 6.4 page 50), c'est-à-dire à chaque hôpital, dans un modèle à 3 niveaux la seule possibilité est $1/\pi_k$ puisque qu'il n'y a pas de niveau supérieur à celui des hôpitaux. Concernant les poids du niveau 2, c'est-à-dire des logements, les candidats identifiés pour le modèle à 3 niveaux sont les suivants :

7. $\mathbf{w_7} : 1/\pi_j^b$
8. $\mathbf{w_8} : 1$
9. $\mathbf{w_9} : 1/(\pi_{j|k} \times \pi_j^b)$

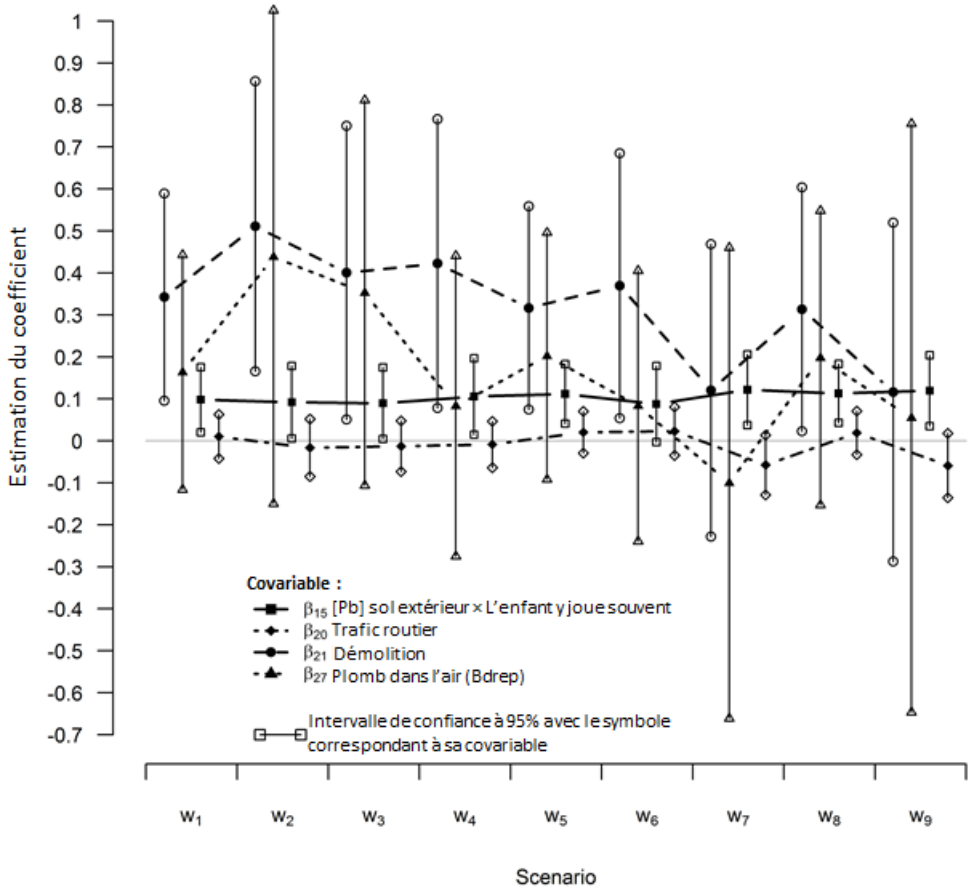
Pour le cas $\mathbf{w_8}$, au lieu d'associer $1/\pi_k$ pour les unités du niveau 3, on utilise un poids égal à 1 afin d'estimer un modèle à 3 niveaux non pondéré, de manière analogue à ce qui est fait dans la modélisation à 2 niveaux. Les poids $\mathbf{w_7}$ et $\mathbf{w_9}$ sont identiques respectivement à $\mathbf{w_1}$ et $\mathbf{w_6}$ mais la notation $\mathbf{w_7}$ et $\mathbf{w_9}$ permet alors de différencier la situation où le modèle est à 3 niveaux de la situation où le modèle est à 2 niveaux. Les candidats $\mathbf{w_2}$, $\mathbf{w_3}$ et $\mathbf{w_4}$ présents pour le modèle à 2 niveaux sont sans objet dans la situation d'un modèle à 3 niveaux car ils ne sont pas basés sur des probabilités de sélection conditionnelles qui seules doivent apparaître pour les niveaux 1 et 2.

Au total, 9 scénarios sont testés et mis en comparaison : 6 concernant une modélisation à 2 niveaux et 3 concernant une modélisation à 3 niveaux. À ce stade, l'estimation a été réalisée sur cas complets (1605 pièces réparties dans 429 logements). De plus, pour l'application numérique les données en plomb acido-soluble ont été utilisées dans la mesure où le dosage acido-soluble concerne la réglementation française. *In fine*, la considération du plomb acido-soluble plutôt que le plomb total a permis de travailler avec plus d'observations (1595 cas complets en plomb total).

Les résultats sont montrés par la figure 30 qui illustre à travers 4 covariables sources, les variations pour les coefficients estimés des 30 covariables. Il peut donc être observé plusieurs comportements selon les covariables sources : le coefficient noté β_{15} est estimé de manière stable quelque soit le modèle, à 2 niveaux ou à 3 niveaux, et quelque soit la pondération utilisée pour les unités de niveaux 2. β_{15} est estimé par une valeur située autour de $0,1$.
Le coefficient β_{20} est estimé d'une manière stable pour les modèles à 2 niveaux mais avec changement de signe. Cette stabilité semble être mise en défaut pour les modèles à 3 niveaux, avec de nouveau un changement de signe. Cependant, quelque soit le modèle, à partir de nos données nous ne sommes pas en mesure de montrer que β_{20} ne soit pas nul.
Le coefficient β_{21} possèdent des estimations avec une forte dispersion en fonction de

2. Par cette francisation on entend regroupement ; il y aurait eu *« clusterisation »* si un enfant de la phase 1 habitait dans plusieurs résidences principales ce qui n'est pas possible par définition d'un ménage.

FIGURE 30 – Illustration de l'impact des différents scénarios sur les estimations des coefficients.



la pondération utilisée. L'amplitude est d'environ 0,4 ce qui est conséquent dans la mesure où l'on travaille avec des logarithmes. Les estimations à partir du modèle à 2 niveaux sont plus élevées que celles du modèle à 3 niveaux. De plus l'effet de la variable relative, « Démolition », est mis en évidence à partir de nos données pour le modèle à 2 niveaux alors que pour le modèle à 3 niveaux cela n'est pas le cas pour 2 de ses scénarios ($\mathbf{w_7}$ et $\mathbf{w_9}$).
Enfin, une observation similaire à celle du coefficient β_{21} peut être faite quant à la variance des estimations de β_{27} en ajoutant que, les estimations ne sont pas du même signe. Néanmoins l'effet du niveau en plomb dans l'air ne semble pas mis en évidence à partir de nos données.

Il faut noter de plus qu'aucun résultat contradictoire n'a été obtenu à travers les 31 coefficients estimés [3] : si un effet a été mis en évidence avec un signe donné, il n'a jamais été mis en évidence avec un signe opposé par un autre scénario (résultats non montrés).

Il apparait donc que la pondération pour le niveau 2 du modèle introduite dans la pseudo vraisemblance, doit être adéquate pour estimer les

3. 30 covariables, source et confusion, plus l'*« intercept »*.

paramètres du modèle afin de répondre à l'objectif d'évaluer la contribution respective des sources en plomb à contaminer la poussière.

À ce stade, nous faisons l'hypothèse qu'une modélisation à 2 niveaux non pondérée (scénario w_5) est adéquate pour estimer les coefficients associés à chaque covariable source. Une étude de simulation Monte Carlo sera réalisée au chapitre 3 afin de valider cette hypothèse.

4 Données censurées et imputation des données manquantes

Afin de pouvoir comparer les modèles (section 2.3 du présent chapitre) et réaliser l'application numérique précédente, les valeurs inférieures à la LQ pour le dosage du plomb dans les poussières intérieures, ont été substituées par la valeur LQ/2. Une telle substitution avait déjà été utilisée dans le chapitre 2 et des éléments de discussion ont été donnés à ce sujet (cf. section 2.5 de la partie « Discussion »). Par contre, pour les poussières extérieures et le sol extérieur de l'aire de jeu de l'enfant, aucune valeur n'était inférieure à leur LQ.

La table de données comportant les covariables du modèle à ajuster contient 1834 observations correspondant à l'ensemble des pièces des 484 logements. Cette table contient environ 0,6 % de données manquantes ce qui apparaît comme un pourcentage faible. Cependant vu la répartition de ces données manquantes dans la table, 12,5 % des observations sont perdues (supprimées par le logiciel lors de l'estimation des paramètres du modèle). Ceci a été jugé trop important car induisant possiblement un biais pour des estimateurs (cf. section 7.3 de la partie « Spécificités des données d'enquête »). Il a été alors décidé d'imputer les données par imputation multiple.

Afin de procéder à l'imputation des données manquantes, le raisonnement a d'abord été pragmatique : qu'est-il possible de faire en termes d'imputation multiple sur le type de données sur lequel on travaille à partir de Stata V12 ?
En effet 2 caractéristiques du type de données sont importantes : le premier est le fait de travailler sur des données d'enquête dont les observations possèdent des poids de sondage ; le second est le fait que les données soient *« clusterisées »* c'est-à-dire que plusieurs pièces se trouvent dans un même logement et qu'il faut tenir compte de cela lorsqu'une pièce reçoit une valeur d'imputation (corrélation avec les autres pièces du logement). Le raisonnement s'est donc d'abord basé sur les possibilités offertes par Stata V12 pour traiter nos données et des recommandations faites [4] sur l'imputation de données *« clusterisées »*. Ces recommandations se basent sur des résultats issus de la littérature que l'on ne citera pas car non explorée.

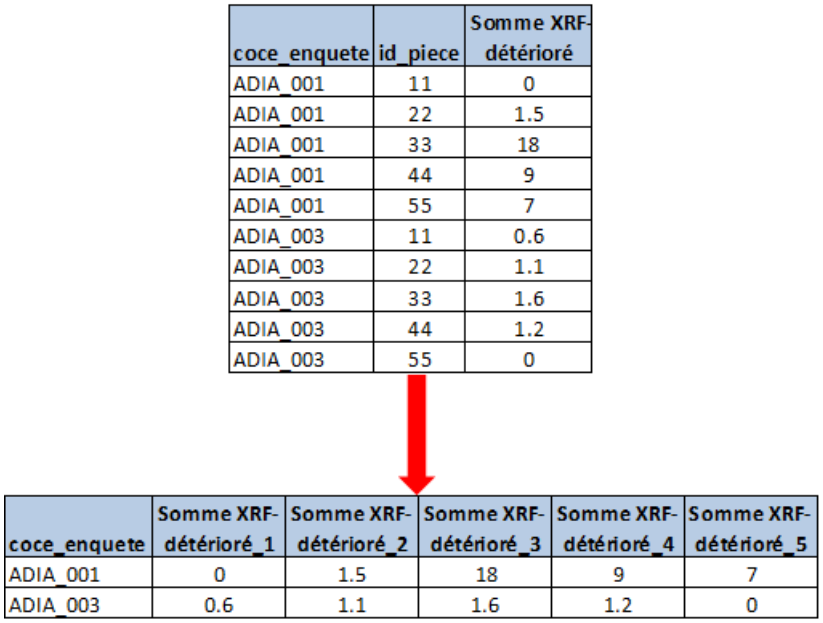
4. http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/ encore accessible le 11 juillet 2013.

Il faut souligner que cette problématique des données *« clusterisées »* en cas d'imputation concerne essentiellement les variables à imputer relatives aux pièces (niveau 1) pour ce qui concerne nos données, puisque ce sont les pièces qui sont regroupées par logement. Trois stratégies sont proposées d'après les recommandations faites sous Stata V12.
Les deux premières ont été écartées. En effet la première stratégie est recommandée lorsqu'il y a peu de groupes (*clusters*) avec beaucoup d'observations dans chaque groupe. Or nous disposons de 484 *« clusters »* (logements) avec de 2 à 5 pièces enquêtées pour chacun d'eaux ; ce que nous avons jugé comme un nombre élevé de *« clusters »* contenant très peu d'observations.
La seconde stratégie demande de même un nombre suffisant d'observations par *« clusters »* car elle permet aux données manquantes d'avoir une distribution différente par *« cluster »*. La seconde stratégie a donc été écartée en raison d'un trop faible nombre de pièces par logement.
Le principe de la 3^{e} stratégie a été retenu. Cette stratégie consiste à utiliser un modèle gaussien d'imputation multivariée pour imputer tous les *« clusters »* à la fois. Cette stratégie fonctionne bien lorsqu'il y a très peu d'observations par *« cluster »* ce qui est le cas pour nos données. Techniquement l'astuce repose sur le fait de transformer la table de données, en créant une colonne (variable) associée à chaque entité des *« clusters »* afin de disposer *in fine* d'une ligne par *« cluster »*. Autrement dit, si on considère une covariable X donnant une information pour chaque pièce i dans chaque logement, $1 \leq i \leq 5$, 5 colonnes seront construites : la première $\mathbf{X}^1$ dont les composantes indiquent la valeur prise par X pour les pièces indexées en $i = 1$, la seconde $\mathbf{X}^2$ dont les composantes indiquent la valeur prise par X pour les pièces indexées en $i = 2$ *etc.* Ceci est illustré sur la figure 31 où $X =$ « Somme XRF-détérioré ». Par ce moyen il est alors possible de prendre en compte les dépendances à l'intérieur de chaque logement dans le modèle d'imputation. Les observations au sein de chaque logement sont vues comme un échantillon issu d'une distribution normale avec une structure de covariance libre [5]. Cependant cette stratégie possède un désavantage par rapport à nos données : elle fonctionne au mieux pour des données équilibrées, autrement dit pour un nombre égal d'observations pour chacun des *« clusters »*. Or il y a de 2 à 5 pièces dans nos 484 logements et non un nombre de pièces fixe.

Cette dernière stratégie et notamment son reformatage des données, ont été retenus mais en remplaçant le modèle gaussien multivarié par une imputation ICE (cf. section 7.4 de la partie 5). Ceci afin de pouvoir considérer une imputation multivariée mais sur différents types de variables, continues et catégorielles. De plus il était nécessaire de pouvoir utiliser une méthode d'imputation adaptée à chacune des variables à imputer, et en particulier des imputations tronquées ; ICE permet de procéder ainsi. Le problème des données déséquilibrées (nombre de pièces variant selon le logement) a été indiqué dans la commande du logiciel codant le modèle d'imputation. Ce problème ne concerne que la variable réponse, Y, la charge en plomb dans les poussières, car c'est la seule variable « niveau pièce » entrant dans le modèle et ayant des données manquantes (figure 32). Ce problème a eu pour conséquence de ne pas pouvoir imputer Y_{i_1} à partir des prédicteurs Y_{i_2} lorsque $i_1 < i_2$, où Y est indexé suite au reformatage des données. C'est ce point qui a été indiqué dans

5. $Cov(X_{i_1 j}, X_{i_2 j})$ peut être différente de $Cov(X_{i_3 j}, X_{i_4 j})$ si $(i_1, i_2) \neq (i_3, i_4)$.

FIGURE 31 – Reformatage des données *« clusterisées »* pour l'imputation multivariée.

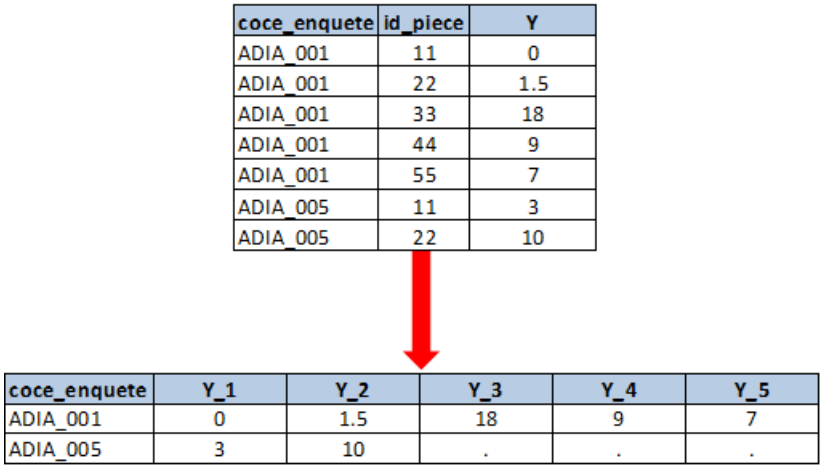


coce_enquete	id_piece	Somme XRF-détérioré
ADIA_001	11	0
ADIA_001	22	1.5
ADIA_001	33	18
ADIA_001	44	9
ADIA_001	55	7
ADIA_003	11	0.6
ADIA_003	22	1.1
ADIA_003	33	1.6
ADIA_003	44	1.2
ADIA_003	55	0

coce_enquete	Somme XRF-détérioré_1	Somme XRF-détérioré_2	Somme XRF-détérioré_3	Somme XRF-détérioré_4	Somme XRF-détérioré_5
ADIA_001	0	1.5	18	9	7
ADIA_003	0.6	1.1	1.6	1.2	0

la commande du logiciel. En revanche il est possible d'utiliser les Y_{i_2} pour imputer Y_{i_1} si $i_1 < i_2$ et ainsi d'introduire l'information de dépendance entre les charges en plomb des pièces d'un même logement.

FIGURE 32 – Problématique lors du reformatage des données dues au déséquilibre des données *« clusterisées »*.



coce_enquete	id_piece	Y
ADIA_001	11	0
ADIA_001	22	1.5
ADIA_001	33	18
ADIA_001	44	9
ADIA_001	55	7
ADIA_005	11	3
ADIA_005	22	10

coce_enquete	Y_1	Y_2	Y_3	Y_4	Y_5
ADIA_001	0	1.5	18	9	7
ADIA_005	3	10	.	.	.

Huit variables (7 covariables + la variable réponse) du modèle d'analyse ont du être imputées. Une neuvième variable supplémentaire à imputer a été ajoutée dans l'imputation ICE (la mesure XRF maximale du palier) dans la mesure où elle a été considérée *a priori* comme prédictrice des charges en plomb des poussières. Les variables continues à imputer introduites sous forme transformée (Log) dans le modèle d'analyse que l'on cherche à ajuster, ont été imputées sous leur forme logarithmique par le modèle d'imputation.

Puisque les deux dosages du plomb, total et acido-soluble, ont été réalisés dans les prélèvements environnementaux, l'un a été introduit pour imputer l'autre dosage. En effet les 2 dosages sont très corrélés. Ainsi dans le modèle d'imputation d'une variable indiquant un niveau en plomb (par exemple la variable Y) selon le dosage acido-soluble par exemple, la variable homologue en plomb total a été introduite comme prédicteur. Par contre pour les autres prédicteurs indiquant un niveau en plomb (par exemple dans le sol extérieur), seule la variable du même dosage que la variable à imputer à été introduite. Par exemple si la charge en plomb acido-soluble dans les poussières était à imputer, la concentration en plomb total du sol n'a pas été utilisée comme prédicteur. Ceci afin d'éviter la multi-collinéarité des prédicteurs.

Les variables du plan de sondage de Saturn-Inf/Plomb-Habitat (cf. section 8.2 de la partie « Spécificités des données d'enquête ») ont été introduites dans le modèle d'imputation. En particulier la stratification du degré hôpital a été introduite via les identifiants des strates. En revanche il n'a pas été possible d'introduire les identifiants des hôpitaux en tant que PSU car cela produisait une non convergence de l'algorithme de ICE. Cependant suite aux recommandations de [Reiter et al., 2006] la taille des PSU dans la population i.e. la somme des poids de sondage des logements par PSU, a été introduite dans le modèle d'imputation en substitution des identifiants des PSU. En outre, l'utilisation des poids de sondage des logements a conduit à une non convergence de l'algorithme ICE et n'ont donc pas été retenus dans le modèle d'imputation : la matrice $\mathbf{T}$ de variance-covariance des estimateurs IM, définie par l'équation 7.5 page 62 étant non définie positive[6]. Ceci est généralement du à la matrice de variance-covariance *inter*, $\mathbf{B}_M$, qui n'est pas de plein rang ou bien dont l'estimation est peu fiable (quand M est petit).

Le modèle d'imputation *in fine* retenu pour chacune des variables à imputer figure en annexe 5. Il a été nécessaire de réviser plusieurs fois le modèle d'imputation (suppression de prédicteurs pour certaines variables à imputer car ils induisaient une non convergence du modèle ; passage d'une méthode à une autre, par exemple de la régression à la méthode *predictive mean matching* ; *etc.*). La comparaison des distributions observées et celles des données imputées été utilisée pour réviser le modèle d'imputation.

$M = 100$ imputations ont été réalisées.

5 Calcul des contributions des sources à contaminer la poussière et interprétation des résultats

Afin d'évaluer la contribution d'une covariable continue, X, dans son unité d'origine (i.e. covariable non transformée), sur la charge en plomb originale, Y, il est judicieux d'estimer le changement en Y provoqué par un changement en X qui a du sens [Harrell, 2001, section 5.3]. Passer du percentile d'ordre 25 % (P25) de X

6. Une matrice réelle M, $p \times p$, est définie positive si $z^\top M z \geq 0$ pour tout $z \in \mathbb{R}^n \backslash \{0\}$.

à son percentile d'order 75 % (P75) est un changement utile dans la mesure où ce changement recouvre la moitié des valeurs de X observées. Puisque les distributions des variables continues utilisées dans le modèle sont généralement dissymétriques à droite (*right-skewed*), on choisit de plus de s'intéresser à un changement en X allant du P50 au P90 ainsi qu'au changement P50-P95 et P50-P97,5. Il a été décidé de ne pas aller au delà du P97,5 dans la mesure où plus l'ordre du percentile est élevé plus l'incertitude autour de l'estimation du percentile est importante.

La contribution d'une source potentielle, X, à contaminer la poussière intérieure déposée au sol, exprimée en pourcentage d'augmentation en Y, a été calculée par :

$$100 \times \left(\frac{x_1}{x_0}\hat{\beta} - 1\right) \% \text{ pour les covariables} X \text{transformées par Log}(X)$$
$$100 \times \left(\frac{x_1 + 1}{x_0 + 1}\hat{\beta} - 1\right) \% \text{ pour les covariables} X \text{transformées par Log}(X + 1)$$
$$100 \times (\exp(\hat{\beta}) - 1)\% \text{ pour les covariables} X \text{binaires passant d'une valeur 0 à 1,}$$

où x_0 est le percentile d'ordre le plus faible dans le changement en X, par exemple le P25, et x_1 est le percentile d'ordre le plus élevé dans le changement en X, par exemple le P75 ; $\hat{\beta}$ l'estimation du coefficient de régression β associé à X dans le modèle mutli-niveaux où β représente indistinctement φ ou ψ dans les équations 1.1 et 1.2 page 116.

Afin de calculer ces contributions, la distribution des sources en plomb est nécessaire ainsi que les effets estimés ($\hat{\beta}$) de chacune de ces sources. La distribution de chacune des sources sera donc présentée dans un premier temps puis l'estimation des effets de chaque source sera ensuite affichée. La distribution des covariables de confusion ainsi que leur effet respectif seront de plus présentés dans un souci de comparaison avec d'autres études le cas échéant, bien que leur contribution respective ne sera pas présentée dans la mesure où elle est sans intérêt pour l'objectif.

Dans le tableau indiquant l'estimation de l'effet de chaque covariable, X, seront associées un intervalle de confiance à 95% et une *p-value* afin d'indiquer :

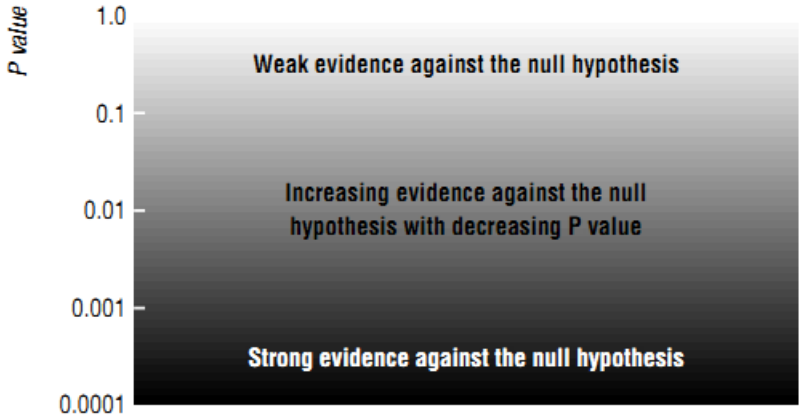
- pour le premier, une gamme plausible de valeurs pour l'effet de X dans la population (et par la même pour la contribution correspondante) et,
- pour la seconde, l'intensité de la présomption contre l'hypothèse nulle c'est-à-dire l'hypothèse que l'effet soit nul (et par la même que la contribution correspondante est nulle).

L'interprétation des résultats est faite sur la base des résultats de l'inférence à partir de nos données avec confrontation aux résultats de précédentes études disponibles dans la littérature, concernant les sources de contamination des poussières. L'intensité de la présomption ou autrement dit la mise en évidence de l'effet, est lue à partir de la *p-value* selon la figure 33. Une source potentielle a été caractérisée dans notre étude comme une source contaminant en plomb la poussière par un signe positif de

l'estimation de son effet, accompagné d'une mise en évidence de son effet.

Les résultats sur données imputées sont pris en considération pour discuter et conclure. Cependant les résultats sur cas complets sont de même affichés comme il l'est recommandé par [Sterne et al., 2009]. Si les résultats entre l'analyse sur cas complets et celle sur données imputées diffèrent de manière trop importante, dans le sens où ils mènent à des décisions inférentielles différentes, cela doit être indiqué et on se doit de comprendre pourquoi.

FIGURE 33 – Suggestion d'interprétation de la *p-value* proposée dans l'article [Sterne & Davey Smith, 2001].



6 Résultats

6.1 Données manquantes

Les tables 28 et 29 montrent la répartition (*pattern*) des données manquantes à travers les variables du modèle comportant des données manquantes. On rappelle que cette table contient 1834 lignes correspondant au 1834 pièces réparties dans les 484 logements. Un exemple de lecture des tables est le suivant (à partir de la table 28) : « Il y a 1605 observations (lignes ↔ pièces) qui ne contiennent aucune donnée manquante à travers les variables utilisées dans le modèle ; il y a 3 observations qui ont à la fois une donnée manquante pour V1, V2 et V6 uniquement » (cf. ligne 13 de la table 28). La table 30 affiche la quantité de données manquantes par variable utilisée dans le modèle multi-niveaux.

Les tables 31 et 32 illustrent, avec 2 covariables imputées, les comparaisons de distributions qui ont été utilisées afin de vérifier l'imputation : il ne semble pas d'avoir d'écart substantiel entre la distribution des valeurs disponibles et celle des valeurs imputées. On s'est limité dans les faits aux 2 premiers jeux de données [7] pour réaliser ce genre de comparaisons pour l'ensemble des variables imputées.

TABLE 28 – *Pattern* des données manquantes (variables Pb acido-soluble).

V1	V2	V3	V4	V5	V6	V7	V8	Nb. de DM	Effectif
+	+	+	+	+	+	+	+	0	1605
+	+	+	+	+	.	+	+	1	49
.	+	+	+	+	+	+	+	1	39
+	+	.	+	+	+	+	+	1	38
+	+	+	+	+	+	+	.	1	34
+	+	+	.	+	+	+	+	1	24
.	+	+	+	+	.	+	+	2	12
.	+	+	+	.	+	+	+	1	11
.	+	.	+	+	+	+	+	2	6
.	+	+	+	.	+	+	+	2	4
.	+	.	+	+	.	+	+	3	4
.	+	+	.	+	+	+	+	2	3
.	.	+	+	+	.	+	+	3	3
+	+	+	+	+	+	.	+	1	2

Légende. DM : données manquantes, V1 = Y (variable réponse) ; V2 = Emplacement du logement ; V3 = Charge en Pb-poussière extérieure (L'enfant y joue souvent) ; V4 = Charge en Pb-poussière extérieure (L'enfant y joue tout le temps) ; V5 = Concentration en Pb-sol extérieur (L'enfant y joue souvent) ; V6 = Charge en Pb-palier ; V7 = Trafic routier ; V8 = Tabagisme journalier. « + » = pas de DM ; « . » = DM.

TABLE 29 – *Pattern* des données manquantes (variables Pb total).

V1	V2	V3	V4	V5	V6	V7	V8	Nb. de DM	Effectif
+	+	+	+	+	+	+	+	0	1595
+	+	+	+	+	.	+	+	1	49
.	+	+	+	+	+	+	+	1	49
+	+	.	+	+	+	+	+	1	38
+	+	+	+	+	+	+	.	1	34
+	+	+	.	+	+	+	+	1	25
.	+	+	+	+	.	+	+	2	12
+	+	+	+	.	+	+	+	1	10
.	+	.	+	+	+	+	+	2	6
.	+	+	+	.	+	+	+	2	5
.	+	.	+	+	.	+	+	3	4
.	.	+	+	+	.	+	+	3	3
+	+	+	+	+	+	.	+	1	2
.	+	+	.	+	+	+	+	2	2

Légende. DM : données manquantes, V1 = Y (variable réponse) ; V2 = Emplacement du logement ; V3 = Charge en Pb-poussière extérieure (L'enfant y joue souvent) ; V4 = Charge en Pb-poussière extérieure (L'enfant y joue tout le temps) ; V5 = Concentration en Pb-sol extérieur (L'enfant y joue souvent) ; V6 = Charge en Pb-palier ; V7 = Trafic routier ; V8 = Tabagisme journalier. « + » = pas de DM ; « . » = DM.

7. Sur les 100 jeux de données produits par l'imputation multiple.

TABLE 30 – Données manquantes par variable entrant dans le modèle multi-niveaux.

Variable		Plomb acido-soluble		Plomb total	
		Nb. de DM	% de DM	Nb. de DM	% de DM
Y		71	3,87	81	4,42
Emplacement du logement		3	0,16	3	0,16
Saison		0	0	0	0
Lavage humide du palier		0	0	0	0
Type de pièce		0	0	0	0
Fréq. Lavage humide-pièce		0	0	0	0
Fréq. Lavage sec-pièce		0	0	0	0
Endroit du prélèvement poussière		0	0	0	0
Nombre d'activités à risque		0	0	0	0
XRF garde-corps		0	0	0	0
Charge en Pb-poussière extérieure	L'enfant y joue souvent	48	2,62	48	2,62
	L'enfant y joue tout le temps	27	1,47	27	1,47
Concentration en Pb-sol extérieur	L'enfant y joue souvent	15	0,82	15	0,82
	L'enfant y joue tout le temps	0	0	0	0
Charge en Pb-palier		68	3,71	68	3,71
Trafic routier		2	0,11	2	0,11
Démolition		0	0	0	0
Fréquence de loisir		0	0	0	0
Travaux extérieurs		0	0	0	0
Travaux intérieurs		0	0	0	0
Sites polluants (Basias)		0	0	0	0
Sols pollués (Basol)		0	0	0	0
Pb dans l'air (Bdrep)		0	0	0	0
Tabagisme journalier		34	1,85	34	1,85
Somme XRF-détérioré		0	0	0	0
Somme XRF-état d'usage		0	0	0	0

Légende. DM : données manquantes.

TABLE 31 – Illustration d'une comparaison de distributions après l'imputation multiple pour le premier jeu de données fourni par l'imputation. Variable Log(Charge en plomb acido-soluble des poussières (μg/m^2)) pour la chambre de l'enfant.

	Nb. Obs.	Moyenne	Écart-type	Min	Max
Valeurs disponibles	462	1,685176	1,229753	-2,302585	5,953243
Valeurs imputées	22	1,489981	1,692356	-0,3566749	5,814131
Ensemble	484	1,676304	1,252833	-2,302585	5,953243

6.2 Distribution des variables du modèle

La table 33 présente la distribution de chacune des variables utilisées dans le modèle multi-niveaux. Le niveau en plomb dans chaque compartiment environnemental est donné en plomb acido-soluble. Seuls les percentiles utilisés dans le calcul des contributions sont présentés. La table 34 est l'homologue de la table 33 pour le

TABLE 32 – Illustration d'une comparaison de distributions après l' imputation multiple pour le premier jeu de données fourni par l'imputation. Variable Log (Concentration en plomb total du sol extérieur (mg/kg)) lorsque l'enfant y joue souvent.

	Nb. Obs.	**Moyenne**	**Écart-type**	**Min**	**Max**
Valeurs disponibles	198	3,644543	1,046053	1,064711	8,134174
Valeurs imputées	4	4,181882	1,47365	2,33328	5,940501
Ensemble	202	3,655184	1,053801	1,064711	8,134174

plomb total ; évidemment seules les distributions des covariables renseignant sur les niveaux en plomb changent entre les 2 tables.

En note de bas de table sont données les distributions en cas complets des variables qui ont subi l'imputation.

TABLE 33 – Distributions (pondérées) des covariables utilisées dans le modèle multi-niveaux (**Pb acido-soluble**).

Covariable		**Percentile (covariables continues) ou répartition (covariables catégorielles)**					
		P25	**P50**	**P75**	**P90**	**P95**	**P97,5**
Emplacement du logement∗[a]		Semi-enterré (5,8%) ; Rdc (61%) ; En étage (33,2%)					
Saison		Automne/hiver (30,4%) ; Printemps/été (69,6%)					
Lavage humide du palier		Pas de palier (72,4%) ; Oui (23,4%) ; Non (4,2%)					
Type de pièce		Chambre (34,6%) ; Entrée (12,9%) ; Séjour (26,1%) ; Cuisine (19,8%) ; Salle de jeux (6,5%)					
Fréq. Lavage humide-pièce (Nb. fois/semaine)		1	2	3	7	7	7
Fréq. Lavage sec-pièce (Nb. fois/semaine)		2	3	7	7	10	15
Endroit du prélèvement poussière		Endroit de jeu préféré de l'enfant (40,2%) ; Au centre de la pièce (59,8%)					
Nombre d'activités à risque		0	0	0	2	4	5
XRF garde-corps (mg/cm^2)		0	0	0	0,2	2,6	5,1
Charge en Pb-poussière extérieure (μg/m^2)	L'enfant y joue souvent∗[b]	0	0	0	0	16	48
	L'enfant y joue tout le temps∗[c]	0	0	0	0	0	82
Concentration en Pb-sol extérieur (mg/kg)	L'enfant y joue souvent∗[d]	0	0	14,5	42,4	82,9	243,7
	L'enfant y joue tout le temps	0	0	5	39,1	65,5	100,1
Charge en Pb-palier (μg/m^2)∗[e]		0	0	6	30	44	148
Trafic routier (milliers/an/km)∗[f]		18	44	135	627	1210	2222
Démolition		Oui (11,1%) ; Non (88,9%)					
Fréquence de loisir		0	0	0	0	1	1
Travaux extérieurs		Oui (7,3%) ; Non (92,7%)					
Travaux intérieurs		Oui (29%) ; Non (71%)					
Sites polluants (Basias) (1/km)		0	3,5	29,6	100,2	252	314,3
Sols pollués (Basol) (1/km)		0	0	0	1,2	2,2	3,1
Pb dans l'air (Bdrep) (kg/an/km)		0	0	0	0,6	1	1,3
Tabagisme journalier (h/jour)∗[g]		0	0	0	1,5	3,5	5
Somme XRF-détérioré (mg/cm^2)		0	0	0	0	0	0
Somme XRF-état d'usage (mg/cm^2)		0	0	0	0	1,1	5,9

Légende. À cause des arrondis la somme des pourcentages peut ne pas être égale à 100. ∗ : covariables pour lesquelles les statistiques ont été estimées sur données imputées. Leurs mêmes statistiques sur cas complets étaient les suivantes :
[a] : 5, 8% 61% 33, 2% ; [b] : 0 0 0 0 0 11 21 ; [c] : 0 0 0 0 0 0 32 ; [d] : 0 0 23, 0 64, 6 97, 6 267, 5 ; [e] : 0 0 5 36 46 82 ; [f] : 18 44 135 627 1210 2222 ; [g] : 0 0 0 1, 5 3, 5 5.

TABLE 34 – Distributions (pondérées) des covariables utilisées dans le modèle multi-niveaux (**Pb total**).

Covariable		**Percentile (covariables continues) ou répartition (covariables catégorielles)**					
		P25	**P50**	**P75**	**P90**	**P95**	**P97,5**
Emplacement du logement∗[a]		Semi-enterré (5,8%) ; Rdc (61%) ; En étage (33,2%)					
Saison		Automne/hiver (30,4%) ; Printemps/été (69,6%)					
Lavage humide du palier		Pas de palier (72,4%) ; Oui (23,4%) ; Non (4,2%)					
Type de pièce		Chambre (34,6%) ; Entrée (12,9%) ; Séjour (26,1%) ; Cuisine (19,8%) ; Salle de jeux (6,5%)					
Fréq. Lavage humide-pièce (Nb. fois/semaine)		1	2	3	7	7	7
Fréq. Lavage sec-pièce (Nb. fois/semaine)		2	3	7	7	10	15
Endroit du prélèvement poussière		Endroit de jeu préféré de l'enfant (40,2%) ; Au centre de la pièce (59,8%)					
Nombre d'activités à risque		0	0	0	2	4	5
XRF garde-corps (mg/cm²)		0	0	0	0,2	2,6	5,1
Charge en Pb-poussière extérieure (μg/m²)	L'enfant y joue souvent∗[b]	0	0	0	0	20	53
	L'enfant y joue tout le temps∗[c]	0	0	0	0	0	91
Concentration en Pb-sol extérieur (mg/kg)	L'enfant y joue souvent∗[d]	0	0	23	64,6	97,6	267,5
	L'enfant y joue tout le temps	0	0	10,3	53,4	102,1	118,5
Charge en Pb-palier (μg/m²)∗[e]		0	0	7,7	41,1	51,7	172,5
Trafic routier (milliers/an/km)∗[f]		18	44	135	627	1210	2222
Démolition		Oui (11,1%) ; Non (88,9%)					
Fréquence de loisir		0	0	0	0	1	1
Travaux extérieurs		Oui (7,3%) ; Non (92,7%)					
Travaux intérieurs		Oui (29%) ; Non (71%)					
Sites polluants (Basias) (1/km)		0	3,5	29,6	100,2	252	314,3
Sols pollués (Basol) (1/km)		0	0	0	1,2	2,2	3,1
Pb dans l'air (Bdrep) (kg/an/km)		0	0	0	0,6	1	1,3
Tabagisme journalier (h/jour)∗[g]		0	0	0	1,5	3,5	5
Somme XRF-détérioré (mg/cm²)		0	0	0	0	0	0
Somme XRF-état d'usage (mg/cm²)		0	0	0	0	1,1	5,9

Légende. À cause des arrondis la somme des pourcentages peut ne pas être égale à 100. ∗ : covariables pour lesquelles les statistiques ont été estimées sur données imputées. Leurs mêmes statistiques sur cas complets étaient les suivantes :
[a] : 5, 8% 61% 33, 2% ; [b] : 0 0 0 0 13 24 ; [c] : 0 0 0 0 0 32 ; [d] : 0 0 23, 0 64, 3 97, 4 267, 5 ; [e] : 0 0 5 36 46 82 ; [f] : 18 44 135 627 1210 2222 ; [g] : 0 0 0 1, 5 3, 5 5.

6.3 Estimation des coefficients du modèle et des contributions des sources

La table 35 affiche les estimations des coefficients du modèle à 2 niveaux, expliquant la (Log) charge en plomb des poussières des pièces en fonction des covariables sélectionnées (sources et variables de confusion). La table 35 concerne le dosage en plomb acido-soluble.

La partie « Sur données imputées » est relative à l'estimation du modèle faite à partir des $M = 100$ jeux de données obtenus par l'imputation multiple. Les estimations obtenues à partir des cas complets sont affichées dans la même table dans la partie « Sur cas complets ».

La table 36 est analogue à la table 35 pour le dosage en plomb total.

La première constatation est que toutes les covariables suspectées comme étant des sources en plomb à contaminer la poussière déposée au sol ont semblées être confirmées comme telles par un coefficient estimé positivement excepté pour deux d'entre elles : Travaux extérieurs et Log(Basol+1) [8]. Cependant ces effets protecteurs n'ont pas été mis en évidence.

La seconde constatation (36) que les covariables pour lesquelles il y a une mise en évidence très importante de leur effet respectif (à partir de nos données) sont :

- Log(XRF garde-corps+1)
- Log(Charge en Pb-palier+1)
- Log(Tabagisme journalier+1)
- Log(Somme XRF-détérioré+1)

De plus les covariables pour lesquelles il y a une mise en évidence importante de l'effet à partir de notre jeu de données sont :

- Log(Concentration en Pb-sol ext.+1) lorsque l'enfant joue souvent sur l'aire de jeu
- Log(Concentration en Pb-sol ext.+1) lorsque l'enfant joue tout le temps sur l'aire de jeu
- Log(Somme XRF-état d'usage+1)

Enfin, les covariables pour lesquelles il y a une mise en évidence moyennement importante de l'effet et qui mériteraient une nouvelle expérience pour que leur effet soit confirmé sont :

- Démolition

8. C'est-à-dire la réalisation de travaux sur l'extérieur du logement dans les 6 mois précédant l'enquête au domicile, et le score relatif (en Log) aux sols pollués dans un rayon de 2 km autour du logement (cf. annexe 3).

– Log(Basias+1)

Toutes les autres covariables, sous la forme telles qu'elles apparaissent dans le modèle, ont un effet faiblement mis en évidence à très faiblement mis en évidence à partir de nos données.

Pour le moment ces constatations seules ne permettent pas de savoir si l'effet d'une covariable est important par rapport aux autres et encore moins si sa contribution est importante, en particulier à cause des transformations logarithmiques qui ne rendent pas l'interprétation directe.

TABLE 35 – Résultats du modèle à 2 niveaux pour le logarithme de la charge en plomb **acido-soluble**.

		Sur données imputées (Nb. Obs. = 1834)			Sur cas complets (Nb. Obs. = 1605)		
Covariable introduite dans le modèle	**Modalités**	**Estimation**	***p***	**IC 95%**	**Estimation**	***p***	**IC 95%**
« intercept »		0,903	0,008	(0,241 ;1,566)	0,968	0,007	(0,265 ;1,671)
Variable de confusion							
Emplacement du lgt	Semi-enterré	réf.			réf.		
	RDC	0,093	0,715	(-0,405 ;0,591)	0,029	0,912	(-0,480 ;0,537)
	En étage	0,322	0,288	(-0,272 ;0,916)	0,327	0,284	(-0,272 ;0,926)
Saison	Automne/hiver	réf.			réf.		
	Printemps/été	0,323	0,002	(0,122 ;0,523)	0,34	0,002	(0,126 ;0,554)
Lavage humide du palier	Pas de palier	réf.			réf.		
	Oui	-1,649	0,000	(-2,233 ;-1,065)	-1,533	0,000	(-2,042 ;-1,023)
	Non	-1,87	0,000	(-2,700 ;-1,041)	-1,684	0,000	(-2,338 ;-1,029)
Type de pièce	Chambre	-0,476	0,000	(-0,615 ;-0,338)	-0,517	0,000	(-0,656 ;-0,377)
	Entrée	réf.			réf.		
	Salon	-0,234	0,000	(-0,355 ;-0,114)	-0,273	0,000	(-0,394 ;-0,153)
	Cuisine	-0,046	0,394	(-0,15 ;0,059)	-0,039	0,450	(-0,141 ;0,062)
	Salle de jeu	-0,319	0,002	(-0,524 ;-0,114)	-0,35	0,001	(-0,557 ;-0,143)
Fréq. Lavage humide-pièce		0,132	0,063	(-0,007 ;0,271)	0,074	0,317	(-0,071 ;0,219)
Fréq. Lavage sec-pièce		-0,016	0,797	(-0,135 ;0,104)	-0,001	0,986	(-0,129 ;0,127)
Endroit du prélèvement	Endroit de jeu préféré	réf.			réf.		
	Centre de la pièce	0,003	0,945	(-0,092 ;0,098)	-0,016	0,737	(-0,111 ;0,078)
Source							
Log(Nombre d'activités à risque+1)		0,052	0,572	(-0,127 ;0,230)	0,084	0,381	(-0,104 ;0,273)
Log(XRF garde-corps+1)		0,324	0,000	(0,151 ;0,498)	0,413	0,000	(0,200 ;0,626)
Log(Charge en Pb-poussière ext.+1)	Ne joue pas à l'ext.	réf.			réf.		
	Y joue souvent	0,081	0,104	(-0,017 ;0,179)	0,121	0,04	(0,005 ;0,237)
	Y joue tout le tps	0,082	0,165	(-0,034 ;0,199)	0,138	0,046	(0,003 ;0,274)
Log(Concentration en Pb-sol ext.+1)	Ne joue pas à l'ext.	réf.			réf.		
	Y joue souvent	0,102	0,002	(0,037 ;0,166)	0,112	0,002	(0,041 ;0,182)
	Y joue tout le tps	0,107	0,005	(0,033 ;0,181)	0,117	0,005	(0,036 ;0,198)
Log(Charge en Pb-palier+1)		0,412	0,000	(0,240 ;0,585)	0,317	0,000	(0,201 ;0,434)
Log(Trafic routier)		0,023	0,309	(-0,022 ;0,069)	0,021	0,419	(-0,029 ;0,07)

Suite page suivante

TABLE 35 – *Suite de la page précédente*

		Sur données imputées (Nb. Obs. = 1834)			Sur cas complets (Nb. Obs. = 1605)		
Covariable introduite dans le modèle	**Modalités**	**Estimation**	***p***	**IC 95%**	**Estimation**	***p***	**IC 95%**
Démolition	Non	réf.			réf.		
	Oui	0,25	0,029	(0,026 ;0,475)	0,317	0,010	(0,074 ;0,559)
Log(Fréquence de loisir+1)		0,077	0,509	(-0,151 ;0,306)	0,042	0,743	(-0,207 ;0,291)
Travaux extérieurs	Non	réf.			réf.		
	Oui	-0,183	0,197	(-0,460 ;0,095)	-0,09	0,549	(-0,384 ;0,204)
Travaux intérieurs	Non	réf.			réf.		
	Oui	0,159	0,142	(-0,053 ;0,372)	0,123	0,267	(-0,094 ;0,340)
Log(Basias+1)		0,078	0,025	(0,010 ;0,145)	0,09	0,016	(0,017 ;0,163)
Log(Basol+1)		-0,14	0,431	(-0,489 ;0,209)	-0,095	0,678	(-0,541 ;0,352)
Log(Bdrep+1)		0,057	0,607	(-0,160 ;0,274)	0,202	0,178	(-0,092 ;0,496)
Log(Tabagisme journalier+1)		0,308	0,001	(0,120 ;0,497)	0,284	0,002	(0,104 ;0,464)
Log(Somme XRF-détériorié+1)		0,162	0,001	(0,063 ;0,261)	0,157	0,000	(0,070 ;0,243)
Log(Somme XRF-état d'usage+1)		0,090	0,011	(0,021 ;0,160)	0,088	0,033	(0,007 ;0,169)

TABLE 36 – Résultats du modèle à 2 niveaux pour le logarithme de la charge en **plomb total**.

		Sur données imputées (Nb. Obs. = 1834)			Sur cas complets (Nb. Obs. = 1595)		
Covariable introduite dans le modèle	**Modalités**	**Estimation**	***p***	**IC 95%**	**Estimation**	***p***	**IC 95%**
« intercept »		1,046	0,009	(0,262 ;1,83)	1,166	0,006	(0,327 ;2,005)
Variable de confusion							
Emplacement du lgt	Semi-enterré	réf.			réf.		
	RDC	0,262	0,417	(-0,371 ;0,896)	0,196	0,555	(-0,455 ;0,847)
	En étage	0,52	0,157	(-0,201 ;1,242)	0,547	0,144	(-0,187 ;1,282)
Saison	Automne/hiver	réf.			réf.		
	Printemps/été	0,274	0,016	(0,052 ;0,497)	0,286	0,019	(0,047 ;0,526)
Lavage humide du palier	Pas de palier	réf.			réf.		
	Oui	-1,675	0,000	(-2,328 ;-1,022)	-1,466	0,000	(-2,05 ;-0,882)
	Non	-1,853	0,000	(-2,777 ;-0,929)	-1,557	0,000	(-2,289 ;-0,826)
Type de pièce	Chambre	-0,48	0,000	(-0,628 ;-0,331)	-0,522	0,000	(-0,680 ;-0,363)
	Entrée	réf.			réf.		
	Salon	-0,201	0,002	(-0,330 ;-0,072)	-0,243	0,000	(-0,375 ;-0,111)
	Cuisine	-0,012	0,824	(-0,122 ;0,097)	-0,002	0,968	(-0,116 ;0,111)
	Salle de jeu	-0,345	0,007	(-0,595 ;-0,095)	-0,384	0,002	(-0,632 ;-0,136)
Fréq. Lavage humide-pièce		0,178	0,017	(0,032 ;0,325)	0,142	0,065	(-0,009 ;0,292)
Fréq. Lavage sec-pièce		-0,011	0,866	(-0,141 ;0,119)	-0,011	0,878	(-0,147 ;0,126)
Endroit du prélèvement	Endroit de jeu préféré	réf.			réf.		
	Centre de la pièce	0,036	0,505	(-0,070 ;0,141)	0,001	0,982	(-0,111 ;0,113)
Source							
Log(Nombre d'activités à risque+1)		0,096	0,334	(-0,099 ;0,291)	0,141	0,174	(-0,063 ;0,345)
Log(XRF garde-corps+1)		0,306	0,001	(0,125 ;0,486)	0,407	0,000	(0,19 ;0,625)
Log(Charge en Pb-poussière ext.+1)	Ne joue pas à l'ext.	réf.			réf.		
	Y joue souvent	0,043	0,402	(-0,058 ;0,144)	0,064	0,309	(-0,059 ;0,186)
	Y joue tout le tps	0,053	0,354	(-0,059 ;0,165)	0,093	0,146	(-0,033 ;0,219)
Log(Concentration en Pb-sol ext.+1)	Ne joue pas à l'ext.	réf.			réf.		
	Y joue souvent	0,074	0,031	(0,007 ;0,141)	0,075	0,039	(0,004 ;0,146)
	Y joue tout le tps	0,073	0,065	(-0,005 ;0,150)	0,073	0,081	(-0,009 ;0,154)
Log(Charge en Pb-palier+1)		0,369	0,000	(0,189 ;0,549)	0,255	0,000	(0,129 ;0,38)
Log(Trafic routier)		0,012	0,611	(-0,035 ;0,059)	0,007	0,787	(-0,046 ;0,061)

Suite page suivante

TABLE 36 – *Suite de la page précédente*

		Sur données imputées (Nb. Obs. = 1834)			Sur cas complets (Nb. Obs. = 1605)		
Covariable introduite dans le modèle	**Modalités**	**Estimation**	***p***	**IC 95%**	**Estimation**	***p***	**IC 95%**
Démolition	Non	réf.			réf.		
	Oui	0,268	0,029	(0,028 ;0,508)	0,313	0,02	(0,049 ;0,578)
Log(Fréquence de loisir+1)		0,097	0,376	(-0,118 ;0,313)	0,092	0,417	(-0,13 ;0,314)
Travaux extérieurs	Non	réf.			réf.		
	Oui	-0,133	0,383	(-0,431 ;0,165)	-0,067	0,672	(-0,378 ;0,244)
Travaux intérieurs	Non	réf.			réf.		
	Oui	0,149	0,194	(-0,076 ;0,374)	0,105	0,36	(-0,119 ;0,328)
Log(Basias+1)		0,098	0,008	(0,026 ;0,17)	0,109	0,006	(0,032 ;0,186)
Log(Basol+1)		-0,165	0,364	(-0,520 ;0,191)	-0,104	0,67	(-0,58 ;0,373)
Log(Bdrep+1)		0,040	0,731	(-0,188 ;0,269)	0,169	0,304	(-0,154 ;0,492)
Log(Tabagisme journalier+1)		0,359	0,000	(0,170 ;0,547)	0,339	0,000	(0,157 ;0,521)
Log(Somme XRF-détériorié+1)		0,166	0,003	(0,057 ;0,276)	0,171	0,000	(0,086 ;0,255)
Log(Somme XRF-état d'usage+1)		0,093	0,011	(0,022 ;0,164)	0,097	0,037	(0,006 ;0,189)

À partir des estimations et des intervalles de confiance figurant dans les tables 35 et 36, les contributions de chacune des sources en plomb à contaminer la poussière intérieure sont calculées et affichées sur les figures 34 et 35 respectivement. Les données utilisées relatives à ces 2 figures sont consultables en annexe 7. Ces mêmes données pour les cas complets sont aussi disponibles en annexe 8.

À partir de la figure 34, il est dorénavant possible de voir que la contribution des poussières contaminées du palier est sans commune mesure la contribution la plus importante parmi toutes les contributions des sources suspectées. Quelque soit le passage d'un percentile à un autre, parmi ceux considérés, les poussières contaminées du palier ont la plus forte contribution : de 128 % à presque 700 % d'augmentation de la charge en plomb dans la poussière intérieure. Cependant ces importantes contributions peuvent en fait varier dans une gamme de valeurs assez importante, à la vue de leurs intervalles de confiance, par exemple entre 232 % à 1764 % pour la contribution estimée proche de 700 %. Ces contributions n'en restent pas moins de toute manière substantielles.

Après les poussières du palier, le sol extérieur de l'aire de jeu de l'enfant semble être le plus contributeur à contaminer la poussière intérieure par le phénomène de *« track-in*[9] *»*. Cette constatation se fait à partir d'un passage du P25 au P75 pour les 2 variables « sol ». Ces 2 covariables avaient pourtant un effet mis en évidence de manière moins importante que d'autres covariables ; mais ce sont ces 2 covariables qui apparaissent les plus intéressantes, en particulier lorsque l'on s'intéresse à la gamme de valeurs centrales (P25-P75) des distributions des sources, gamme qui recouvre la moitié de leurs valeurs totales.

Si on s'intéresse à une gamme de valeurs « plus à droite » dans les distributions, P50-P95 ou P50-P97,5, les peintures au plomb des garde-corps à l'extérieur jouent un rôle dépassant même celui des sols de l'aire de jeu extérieure. Il en est de même du temps où l'on fume à l'intérieur dont la contribution est du même ordre de grandeur que celle des sols.

Les covariables relatives aux sites polluant, à la charge XRF des peintures en état d'usage et au démolition autour du logement, dont l'effet respectif semblait moyennement prouvé, n'ont qu'un intérêt limité en termes de contributions à contaminer en plomb la poussière intérieure, bien qu'elles peuvent néanmoins jouer un rôle.

Enfin, la somme des mesures XRF des peintures détériorées, bien qu'elles aient un effet dont la preuve est très importante, n'ont pas d'intérêt en terme de contribution à la contamination des poussières, dans la gamme de valeurs allant jusqu'à leur P97,5. Il est nécessaire d'aller au-delà du P97.5 pour observer une certaine contribution mais qui restera de toute manière bien inférieure à celles des autres sources

9. Fait d'introduire dans le logement du sol contaminé par ses vêtements et chaussures.

décrites précédemment.

En revanche les poussières extérieures contaminées, situées sur l'aire de jeu extérieure de l'enfant, ont une contribution non négligeable pour les 2 covariables relatives, dès lors que des fortes charges en plomb, de l'ordre du P97.5 sont atteintes. Ce constat se fait bien que leur effet ait été que faiblement mis en évidence à partir de nos données.

Les résultats sur cas complets ne diffèrent que peu par rapport à ceux sur données imputées. Les différences ne remettent pas en cause les conclusions en termes décisionnels quant à l'identification des sources sur lesquelles il semble judicieux d'agir pour réduire la charge en plomb des poussières intérieures. Certaines différences importantes existent néanmoins comme par exemple pour la contribution de la covariable « Charge en Pb-palier ». Mais là aussi il n'y a pas de conséquence en termes décisionnels ; la contribution du palier restant supérieure à celles des autres sources.

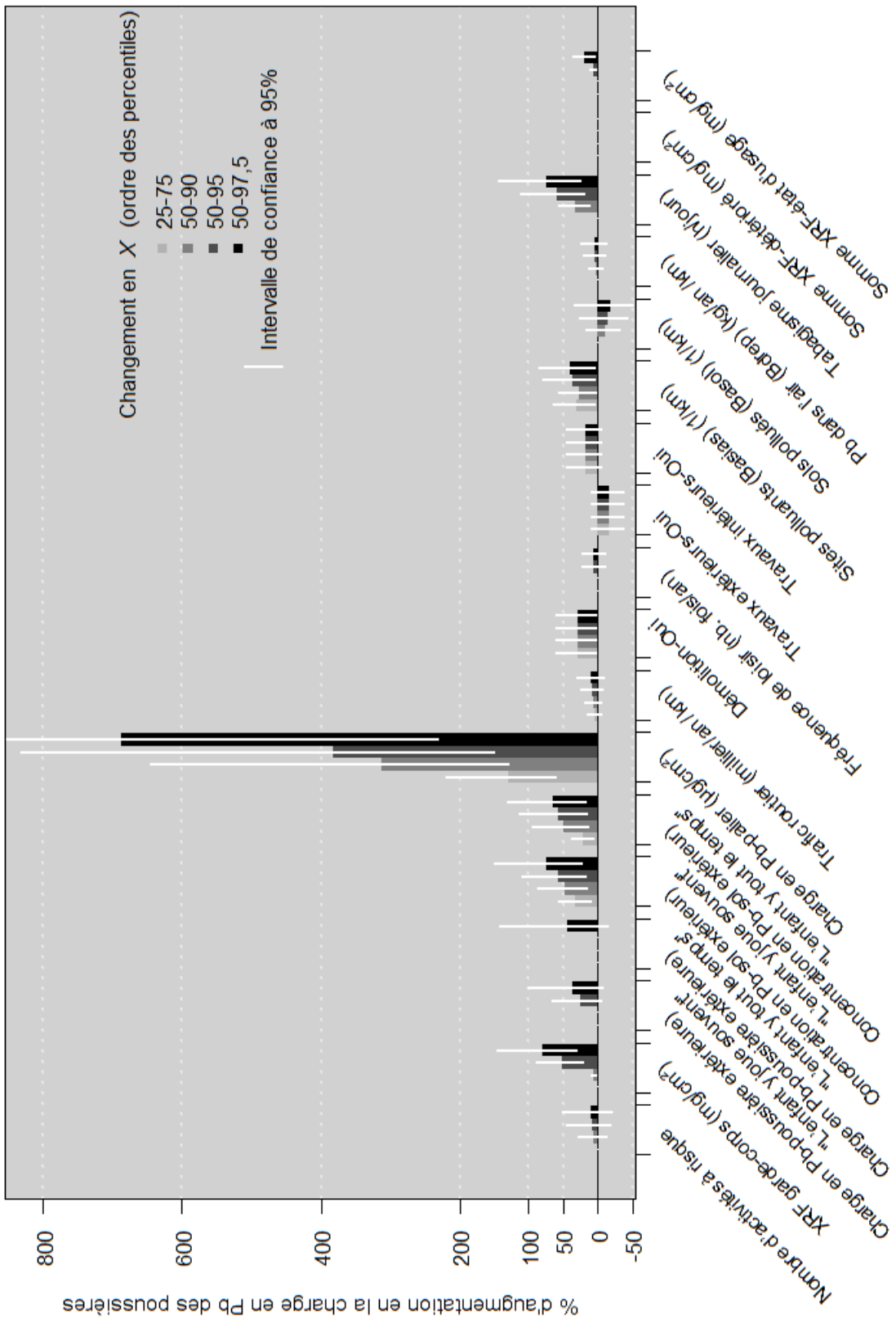


FIGURE 34 – Contribution de chaque source X exprimée en % d'augmentation en la charge en plomb des poussières ($\mu g/m^2$), calculée à partir des données imputées. **Plomb acido-soluble.**

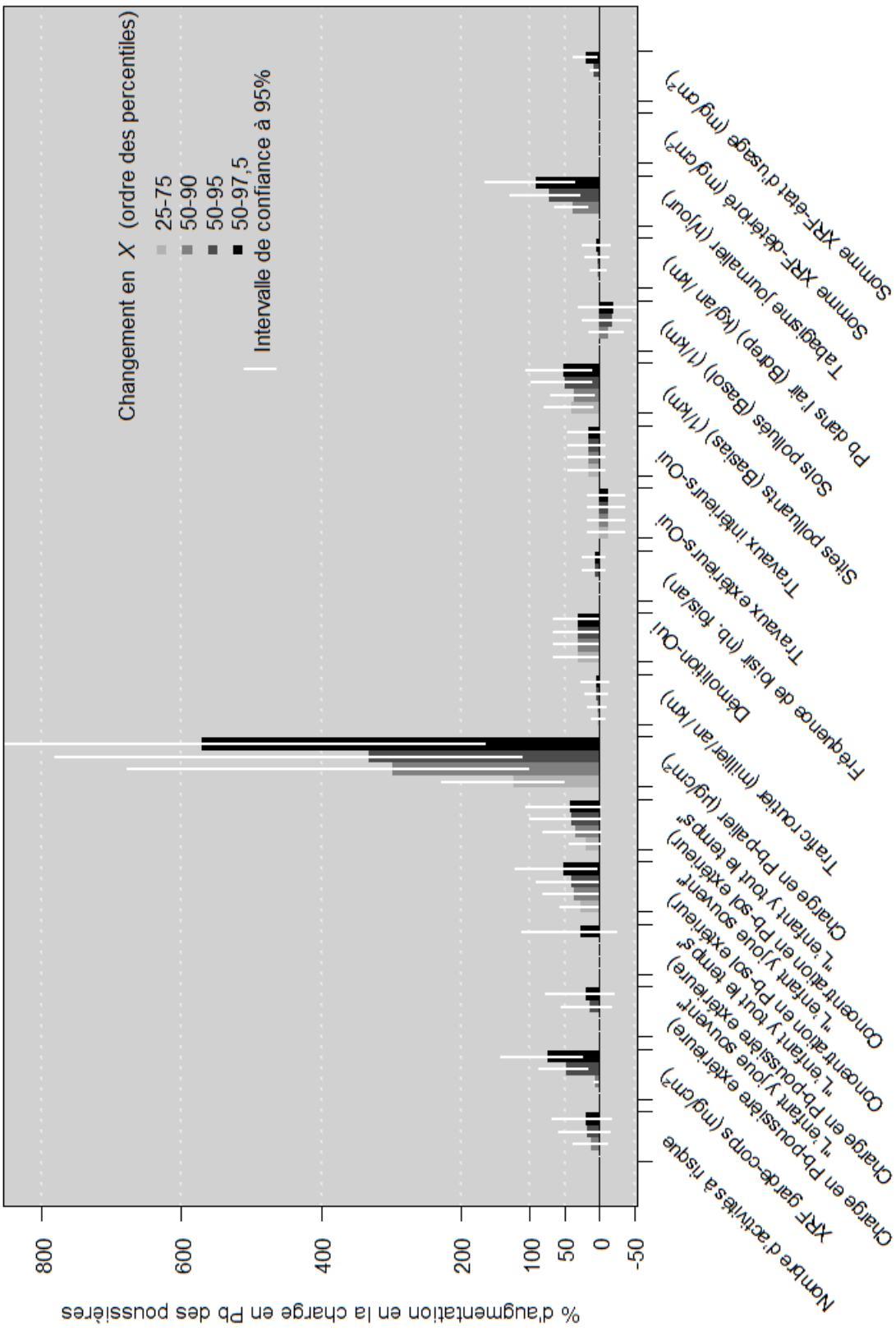


FIGURE 35 – Contribution de chaque source X exprimée en % d'augmentation en la charge en plomb des poussières ($\mu g/m^2$), calculée à partir des données imputées. **Plomb total**

6.4 Corrélation entre les charges en plomb dans la poussière

La variance σ_2^2 de l'effet aléatoire ζ_j sur l'*« intercept »* lié au niveau logement (niveau 2), et la variance σ_1^2 des perturbations ϵ_{ij} ont été estimés selon les valeurs de la table 37.

Ainsi la variance expliquée par le niveau logement (variance *inter* i.e. due à la variabilité entre les logements) est estimée entre 0,6 et 0,65. Autrement dit la corrélation [10] entre 2 (Log) charges en plomb dans les poussières d'un même logement est approximativement de 0,62.

TABLE 37 – Variances inter et intra logement estimées.

Modèle	σ_1^2	**IC 95 %**	σ_2^2	**IC 95 %**	ρ ***(ICC)***
Plomb acido-soluble					
Modèle sur données imputées	0,499	(0,431 ; 0,579)	0,896	(0,746 ; 1,075)	0,642
Modèle sur cas complets	0,468	(0,411 ; 0,533)	0,883	(0,727 ; 1,072)	0,653
Plomb total					
Modèle sur données imputées	0,643	(0,566 ; 0,731)	1,039	(0,879 ; 1,227)	0,618
Modèle sur cas complets	0,615	(0,540 ; 0,701)	1,020	(0,857 ; 1,213)	0,624

7 Synthèse

Dans ce chapitre les **contributions des sources** potentielles jouant un rôle dans la contamination en plomb des poussières déposées au sol à l'intérieur des logements **ont été estimées**.

Pour la première fois un grand nombre de sources pouvant contaminer en plomb la poussière intérieure au sol ont été étudiées **conjointement**.

Les sources en plomb contribuant à contaminer la poussière intérieure sont :

- **les poussières contaminées** du palier d'appartement qui, sans commune mesure avec les autres sources, **contribuent le plus majoritairement**. Si on passait d'une contamination nulle à une contamination d'environ 150 μg/m^2 de plomb (P97,5) dans les poussières du palier, la charge en plomb des poussières à l'intérieur des logements subirait une augmentation de près de 700 %.
- **le sol extérieur** contaminé de l'aire de jeu qui, transporté à l'intérieur du logement, contamine les poussières intérieures. Cette contamination des poussières intérieures est observable à des concentrations en plomb dans les sols

10. cf. page 51 pour sa définition.

de l'ordre d'une dizaine de milligrammes de plomb par kilogramme de sol (≈ P75).
- **la peinture au plomb des gardes-corps** extérieurs, situés sur une terrasse ou un balcon, contribue à augmenter de plus de 50 % la charge en plomb des poussières lorsque la peinture des garde-corps contient au moins 2,6 mg/cm^2 de plomb.
- **le tabagisme** à l'intérieur du logement fait augmenter la charge en plomb des poussières intérieures de près de 60 % dès lors que l'on fume à l'intérieur au moins 1,5 h/jour.
- les sites polluant (Basias) autour du logement, les démolitions autour du logement et la charge en plomb surfacique des revêtements intérieurs en état d'usage contribuent mais de manière limitée à la contamination des poussières intérieures.

Les **peintures intérieures détériorées** à base de plomb ne sont pas associées de manière substantielle à la contamination des poussières intérieures, relativement aux autres sources précédemment listées.

L'identification des sources de la contamination de ces poussières et l'évaluation de leur part d'importance relative peut permettre aux pouvoirs publics de prendre des décisions sur la réduction des niveaux en plomb de ces sources.

Une nouvelle étude à partir de covariables dont l'information aura été collectée plus finement pour certaines d'entre elle, devrait permettre de préciser certaines contributions telles que celle du tabagisme par exemple.

La corrélation dite intra-classe, entre 2 charges en plomb de la poussière à l'intérieur d'un même logement, a été estimée approximativement **égale à 0,60**.

Les résultats de ce chapitre ont été obtenus à partir d'une **modélisation à 2 niveaux** sur données d'enquête. Nous avons montré que **le choix de la pondération** à utiliser pour le niveau 2 (les logements ici, le niveau 1 étant constitué des pièces) **n'est pas évident**. Le type de pondération pour le niveau 2 d'un modèle multi-niveau ne semble pas avoir été précédemment étudié.

Dans le chapitre suivant cette pondération à associer aux logements a été étudiée afin de valider le choix fait dans le présent chapitre d'ajuster un modèle non pondéré ($\mathbf{w_5}$) pour estimer les effets des sources en plomb à contaminer la poussière.

Chapitre 4

Évaluation par simulation de l'impact des poids de niveaux 2 introduits dans la pseudo-vraisemblance dans un modèle à 2 et à 3 niveaux

Dans le chapitre précédent, un modèle multi-niveau, précisément à 2 niveaux à *« intercept »* aléatoire, a été estimé afin d'évaluer la contribution de sources en plomb à contaminer la poussière du sol de l'intérieur des logements. Les logements constituaient les unités du niveau 2 et les pièces de ces logements étaient les unités de niveau 1 (cf. le cadre en page 116).

Le modèle a été ajusté sur des données issues d'un plan de sondage. Il a été vu que le choix de la pondération à introduire dans la pseudo vraisemblance pour chaque niveau était important. Ce choix est important en particulier pour le niveau 2 (logement) car il impacte sur les estimateurs des paramètres (coefficients de régression et variances des effets aléatoires) ; les estimations pouvant varier de manière importante pour certains paramètres. À la fin de la section 3 du chapitre précédent le choix a été de ne pas introduire de pondération pour le niveau 2 (les logements) et donc d'ajuster un modèle non pondéré puisque aucun poids n'était à associer au niveau 1 (les pièces).

Dans le présent chapitre une étude par simulation est réalisée afin de valider le choix d'ajuster un modèle à 2 niveaux non pondéré fait au chapitre 3 à l'issue de la section 3.

Cette simulation consiste à créer artificiellement plusieurs populations de logements/pièces caractérisées par les covariables du modèle multi-niveaux à ajuster. La base du recensement INSEE 2006 permet de disposer du socle de la population de logements d'intérêt. Cette base comporte de plus des informations auxiliaires en commun avec la base de données Plomb-Habitat.

Afin de générer les covariables entrant dans le modèle multi-niveaux, les distributions des covariables estimées à partir des données de Plomb-Habitat ont été utilisées. Pour construire la variable réponse i.e. la charge en plomb des poussières intérieures, les coefficients de régression ainsi que les paramètres de variance des effets aléatoires apparaissant dans le modèle ont été fixés à l'avance (cf. fin de section 1.2 du présent chapitre). Ces valeurs fixées feront office de vraies valeurs de paramètres dans la population.
Pour évaluer la pondération, $w_j^{(2)}$, introduite dans le modèle au niveau 2 il est nécessaire de disposer d'un échantillon. Pour cela, les éléments (logements/pièces) qui figureront dans l'échantillon sont tirés aléatoirement à partir de chacune des populations générées, par un plan de sondage analogue à celui mis en œuvre dans l'enquête Plomb-Habitat. Chacun des scénarios définissant les types de pondération possibles pour le niveau 2 (logements), est appliqué à l'échantillon produit. Ces scénarios correspondent aux 9 scénarios comparés dans le chapitre 3 en section 3. Les coefficients de régression et les paramètres de variance du modèle sont ainsi estimés.
Afin de déterminer le meilleur scénario, en particulier quelle pondération pour les logements était la meilleure, les estimations des coefficients de régression et des paramètres de variance ont été comparés à leur vraies valeurs respectives fixées auparavant. Les critères de comparaison sont explicités en section 3.

Cinq cents populations ont été générées.

Ainsi, le plan de simulation qui sera détaillé dans les sections qui suivent, se résume de la manière suivante :

1. Construire un fichier dont les N [1] lignes sont les logements de la population d'intérêt et créer un fichier associé dont les lignes sont les pièces des logements ;
2. Créer dans ces fichiers les colonnes correspondant à chaque covariable du modèle, en simulant les valeurs de ces covariables ;
3. Générer Y par le modèle multi-niveaux à partir des covariables simulées en ayant fixé la valeur des coefficients de régression et des paramètres de variance ;
4. Tirer un échantillon aléatoire [2] à partir du plan de sondage défini ;
5. Estimer le modèle multi-niveaux pour chacun des 9 scénarios ;
   Retourner à l'étape 1. et recommencer 500 fois.
6. Comparer les estimations des coefficients de régression et des paramètres de variance avec leur vraie valeur.

Un article relatif aux résultats de ce chapitre et intitulé *« Multilevel Modelling of Survey Data : Impact of the 2-level Weights used in the Pseudolikelihood »* a été accepté pour publication dans *Journal of Applied Statistics* [Lucas et al., 2013] (cf. annexe 12).

1. $N = 3\,581\,991$.
2. On verra que seul un échantillon i.e. un seul vecteur d'indicatrice $\mathbf{I} = (I_1, I_2, \ldots, I_{484})$ pour les logements est tiré et utilisé pour les 500 répétitions.

1 Génération de populations

L'objectif est ici de construite un fichier dont les lignes représentent les $N = 3\,581\,991$ logements de notre population. Puisque l'on s'intéresse à la charge en plomb des poussières de leurs pièces, techniquement un second fichier dont les lignes représentent leurs pièces doit aussi être créé.
Une fois ces 2 fichiers de lignes construits, la simulation consiste ensuite à remplir les colonnes. Ces colonnes sont les covariables sélectionnées pour figurer dans le modèle multi-niveaux (cf. chapitre précédent).

Le fichier de logements est disponible via la base du recensement INSEE 2006. On rappelle que certaines informations de cette base avait été utilisées pour procéder au redressement par post-stratification des poids de sondage des logements (cf. section 1.1 du chapitre 2).

1.1 Base du recensement INSEE 2006

Le fichier de logements, représentant les 3 581 991 logements de notre population d'intérêt, qui est extrait de la base INSEE ne contient pas 3 581 991 lignes mais 2 767 857 lignes. Cela s'explique par le fait que l'INSEE associe à chaque ligne un poids [3] dont la somme égale 3 581 991. Ces poids ne sont pas des poids de sondage mais des poids fréquences. Une ligne avec un poids fréquence de 10 indique que la ligne doit être comptée 10 fois dans l'analyse. Il n'y a pas de plan de sondage à prendre en compte lorsque l'on travaille avec de tel poids ; cela n'a rien à voir avec la théorie des sondages.

Afin de créer les 3 581 991 lignes nécessaires dans le fichier, les lignes avec un poids de p doivent être répliquées p fois dans le fichier. La difficulté avec ces poids fréquence INSEE est qu'ils ne sont pas indiqués par des nombres entiers mais par des décimaux. Il n'est donc pas possible de répliquer directement les lignes. Pour palier à cette difficulté la stratégie suivante a été mise en place.

On crée de nouveaux poids fréquences p_{new} à partir des poids fréquence p :

1. Si $p < 1$ alors $p_{\text{new}} = 1$ car on considère que l'on ne peut pas faire « moins d'une ligne » pour un logement ; sinon $p_{\text{new}} = \lfloor p \rfloor$ (la partie entière de p) ;
2. Chaque ligne est recopiée p_{new} fois dans le fichier.

En ayant procédé ainsi, le nombre de lignes ($= \sum p_{\text{new}}$ n'est que de 3 493 511 c'est-à-dire que 88 480 logements manquent dans le fichier. Pour rétablir ces logements manquants, un tirage aléatoire répété 88 480 fois peut être réalisé parmi les logements présents. Cependant il a été décidé de considérer chaque région administrative pour rétablir le nombre de logements manquants ; l'information région étant importante dans la mesure où elle sera utilisée dans le plan de sondage de la simulation (cf.

3. Variable « IPONDL » dans la base INSEE.

section 2 du présent chapitre). Le détail des pondérations par région est donné par la table 38. Toutes les décimales sont conservées afin d'éviter les erreurs d'arrondi.

TABLE 38 – Poids fréquence par région dans le fichier INSEE.

REGION	$\sum p$	$\sum p_{\text{new}}$	$\sum p - \sum p_{\text{new}}$	**Nb. de lignes avec** $p > 1$	**Nb. de lignes avec** $p < 1$
11	736 280,281 8	662 755	73 525,281 77	197 356	199 867
21	76 492,786 51	75 518	974,786 512 7	54 233	9 923
22	118 134,513 1	118 646	−511,486 904 7	92 144	12 269
23	110 780,244 3	109 596	1 184,244 317	77 267	14 459
24	143 696,101 3	143 815	−118,898 695 9	107 099	16 397
25	82 953,119 31	84 537	−1 583,880 691	69 820	7 155
26	86 096,839 5	87 189	−1 092,160 502	69 303	8 438
31	251 693,509 4	240 369	11 324,509 35	146 577	43 250
41	130 176,903 2	131 652	−1 475,096 791	102 687	13 480
42	105 119,123 9	102 491	2 628,123 904	68 479	15 458
43	67 204,329 43	68 647	−1 442,670 575	55 052	6 362
52	207 989,490 2	210 111	−2 121,509 853	158 510	22 760
53	176 529,630 9	180 001	−3 471,369 06	143 818	17 042
54	91 557,255 7	94 317	−2 759,744 298	77 812	7 650
72	165 821,816 6	167 569	−1 747,183 418	120 646	19 967
73	150 231,782 3	151 655	−1 423,217 719	109 916	19 120
74	34 887,491 73	35 154	−266,508 273 2	27 334	3 517
82	359 964,346	357 523	2 441,345 998	245 564	49 123
83	68 427,139 04	69 685	−1 257,860 965	56 662	5 838
91	139 768,738 7	139 972	−203,261 262 6	98 205	18 079
93	263 374,209 1	247 610	15 764,209 12	106 915	60 117
94	14 811,542 05	14 699	112,542 053 1	10 209	1 978

Dans la table 38 on peut donc lire par exemple qu'il y manque dans le fichier « 73 525,281 77 » logements pour la région 11 (Île-de-France ; cf. annexe 1) ; 511,486 904 7 logements sont en trop dans la région 22. Afin de rétablir les effectifs par région la procédure suivante a été utilisée :

1. S'il y a trop de logements, on tire selon un tirage aléatoire simple $\lfloor \lvert \sum p - \sum p_{\text{new}} \rvert \rceil$ [4] logements qui sont enlevés des logements avec $p < 1$;
2. S'il manque des logements, on tire selon un tirage aléatoire simple $\lfloor \sum p - \sum p_{\text{new}} \rceil$ logements parmi les logements avec $p > 1$,

les logements avec initialement $p < 1$ sont considérés comme ayant été favorisés alors que ceux avec $p > 1$ sont considérés avoir été défavorisés lors de la recopie des lignes selon p_{new} faite précédemment.

4. $\lfloor \cdot \rceil$ désigne l'entier le plus proche.

1.2 Génération des variables

L'étape précédente a permis de construire un fichier de 3 581 991 lignes représentant les logements de la population d'intérêt. Ce fichier est donc relatif au niveaux 2 du modèle multi-niveaux. Il faudra disposer d'un second fichier, correspondant aux pièces c'est-à-dire au niveau 1. Sa construction est décrite ci-après.

La génération des covariables utilisées dans le modèle multi-niveaux à estimer correspond à remplir les colonnes des 2 fichiers. Pour le fichier « logement », les covariables donnant une information de niveau 2 seront générées et placées dans ce fichier. Pour le fichier « pièce », les covariables donnant une information de niveau 1 et utilisées dans le modèle seront générées et placées dans ce fichier.

Afin de générer les covariables, une simulation de leurs réalisations est faite à partir de leur distribution marginale. On fait l'hypothèse que les covariables sont indépendantes (comme dans le cadre du construction du modèle en chapitre 3). Leur distribution réelle n'est pas connue puisque les covariables ne figurent pas dans la base du recensement INSEE. Pour pouvoir disposer d'une distribution pour chaque covariable à partir de laquelle chacune des covariables seront simulées, leur distribution est estimée à partir des données de Plomb-Habitat.

Afin d'estimer les distributions des covariables continues, leur forme a été fixée comme étant Log-Normale. C'est cette hypothèse qui a été faite dans le chapitre 3 en utilisant les covariables dans le modèle après transformation logarithmique. Les 2 paramètres de la distribution Log-Normale ont été estimés par la méthode des moments en prenant en compte les poids de sondage ; l'incertitude autour de ces estimations n'est pas considérée car non utilisée dans la simulation. L'estimation a été faite avec le logiciel XLstat [5]. Une fois la distribution empirique déterminée, une réalisation de la covariable est tirée selon cette loi. Le tirage s'est fait sous le logiciel SAS à partir de `exp(a+b×rannor(0))` qui génère une réalisation d'une variable aléatoire $\sim \text{Log}\mathcal{N}(\mu = a, \sigma = b)$ où `rannor(0)` génère une réalisation d'une variable aléatoire $\sim \mathcal{N}(0, 1)$. On verra que la simulation de certaines covariables est fait par groupes, groupes fournis par une ou plusieurs informations auxiliaires.

Les distributions des covariables catégorielles ont été estimées à partir de la répartition empirique de leurs modalités, en tenant compte des poids de sondage. La simulation des covariables catégorielles a été faite de la manière suivante :

Soit X une covariable catégorielle dont les modalités sont $m_1, \ldots, m_u, \ldots, m_q$ dont les fréquences respectives estimées sont $f_u = Pr(X = m_u)$ avec $\sum_u f_u = 1$. L'intervalle $]0; 1]$ peut être alors partitionné selon :

$$]0; 1] = \bigcup_{u=0}^{q-1}]f_u; \sum\nolimits_{n=0}^{u+1} f_n]$$

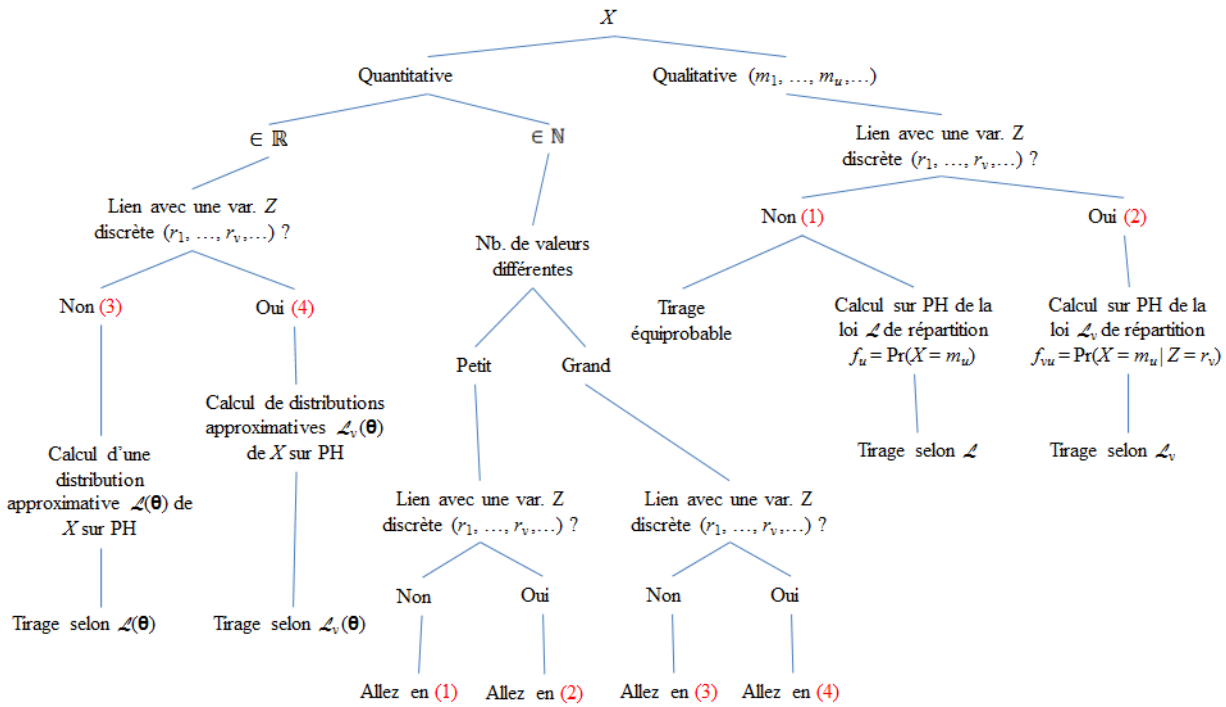
5. Addinsoft (2012). XLSTAT 2012, Logiciel d'Analyse de Données et de Statistique pour Microsoft Excel. Paris, France.

avec $f_0 = 0$. Il est alors procédé à un tirage d'une réalisation d d'une loi uniforme sur $]0;1]$. La modalité qui est alors associée pour la covariable simulée X est m_{u+1} si $d \in]f_u; \sum_{n=0}^{u+1} f_n]$.

Il a été fait l'hypothèse que certaines covariables X étaient reliées à une information auxiliaire. Lorsque cette variable auxilaire est disponible à la fois dans la base de Plomb-Habitat et dans la base INSEE il est alors possible de prendre en compte cette information auxiliaire Z pour simuler X. Pour ce faire la distribution de X est estimée à partir des données de Plomb-Habitat par modalité de Z. Puisqu'à chaque ligne du fichier issu de la base INSEE Z est connue, la valeur de X est simulée selon la distribution de X estimée pour la valeur de Z de cette ligne.

La stratégie de simulation selon le type de variable est résumé par le logigramme de décision montré en figure 36. Certaines covariables à valeurs entières figurent comme variables continues car elles ont été utilisée comme telle dans le modèle multi-niveau, comme par exemple le nombre d'activités à risque, afin d'estimer un seul coefficient de régression qui leur était associé plutôt qu'un nombre de coefficients égal au nombre de valeurs entières différentes relevées (moins une) et qui pouvait être grand. Les informations auxiliaires utilisées à travers la simulation des

FIGURE 36 – Stratégie de simulation des covariables selon leur type.



covariables ont été :

- Le type de logement (individuel/collectif) ;
- Le nombre de pièces principales ;
- L'environnement extérieur (urbain/rural) ;

- La période de construction ($<$ 1949, 1949-1974, 1975-1993, $\geq$ 1994) ;
- La statut d'occupation (propriétaire, hébergé gratuitement, locataire privé, locataire HLM) ;
- La région administrative

La région administrative est une variable importante à prendre en compte dans la simulation car une stratification par région a été utilisée dans le plan de sondage de Saturn-Inf/Plomb-Habitat (cf. section 3.3 de la partie « Spécificités des données d'enquête »).
Le détails de la simulation pour chaque covariable, en particulier en fonction des variables auxiliaires, sont indiqués en annexe 9. La prise en compte de la corrélation de la variable réponse Y entre les pièces pour simuler les valeurs y_{ij} est détaillée dans cette même annexe.

Afin de construire le fichier pièce, le nombre de pièces (lignes) à associer à chaque logement du fichier logement, a été tiré aléatoirement selon son nombre de pièces principales utilisé comme variable auxiliaire (cf. annexe 9).

Une fois les covariables générées, les valeurs y_{ij} dans la population finie de la variable réponse Y, c'est-à-dire la charge en plomb des poussières intérieures déposées au sol de la pièce i du logement j, ont été générées selon le modèle m à 2 niveaux de superpopulation :

$$y_{ij} = \beta_0 + \sum_{r=1}^{q_2} \psi_r x_j^{(r)} + \zeta_j + \sum_{m=1}^{q_1} \varphi_m x_{ij}^{(m)} + \epsilon_{ij} \tag{1.1}$$

avec $\zeta_j \sim \mathcal{N}(0, \sigma_2^2)$ et $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_1^2), Cov(\zeta_j, \epsilon_{ij}) = 0, j = 1, \ldots, N^{(2)}, i = 1, \ldots, N_j^{(1)}$.
Le nombre d'unités de niveau 2 i.e. de logements est $N^{(2)} = 3\ 581\ 991$; $N_j^{(1)}$ est au maximum égal à 5.

La forme de l'équation du modèle m est définie sur la base de la fusion en une équation des équations 6.5 et 6.6 de la partie « Spécificités des données d'enquête » définissant un modèle à 2 niveaux. Pour générer Y les covariables continues ont été considérées dans le modèle sous forme logarithmique. Les coefficients de régression et les paramètres de covariance ont été fixés selon ce qui est indiqué en annexe 10. Les valeurs ont été arbitrairement fixées en faisant un compromis (moyenne) entre les différentes estimations obtenues selon les scénarios en niveau 2 ($\mathbf{w_1}$ à $\mathbf{w_5}$ [6]) sur cas complets en considérant les variables plomb en dosage acido-soluble (c'est-à-dire à l'issue des résultats de l'application numérique de la section 3 du chapitre 3). Dans le cas où la moyenne sur les scénarios n'était pas cohérente au regard de ce qui était attendu pour un coefficient, la valeur a été arbitrairement corrigée. En particulier sur ce qui était attendu pour les sources potentielles, c'est-à-dire un coefficient à signe positif. Sur le même principe les paramètres de variance des effets aléatoires ζ_j et ϵ_{ij} ont été fixés à $\sigma_2^2 = 0,45$ et $\sigma_1^2 = 0,80$ respectivement. Pour simplifier, **les**

6. Les scénarios $\mathbf{w_6}$ à $\mathbf{w_9}$ n'avaient pas encore été envisagés à l'époque où les valeurs ont été fixées.

coefficients de régression dans l'équation 1.1 sont notés dorénavant par β qui remplace φ et ψ, avec β_0 étant toujours l'*« intercept »* et $q_1 + q_2 = 30$. La notation des 2 paramètres de variance est inchangée.

1.3 Réplications et approche modèle

L'étape précédente a donc permis de disposer d'un fichier de logements à 3 581 991 lignes représentant les logements de notre population d'intérêt. Un fichier contenant les pièces des 3 581 991 logements a de plus été créé. Les covariables du modèle multi-niveau recherché (le modèle ajusté en chapitre 3) ainsi que la variable réponse (la charge en plomb des poussières de chaque pièce) figurent à travers ces 2 fichiers. Ainsi avec ces 2 fichiers nous nous mettons dans la situation similaire à celle qui a précédé le tirage de Saturn-Inf/Plomb-Habitat (cf. section 8.2 de la partie « Spécificités des données d'enquête »).

Afin d'obtenir une distribution des valeurs de chaque coefficient de régression et des paramètres de variance que l'on cherche à estimer, on procède par réplications. Ces réplications portent sur la population de logements/pièces car l'analyse multi-niveaux se situe dans une approche modèle (cf. section 6.3 de la partie « Spécificités des données d'enquête »). Ainsi c'est la distribution dite de superpopulation des paramètres qui nous intéresse et non la distribution d'échantillonnage induite par le plan de sondage. Dans ce dernier cas, les réplications auraient porté sur l'échantillonnage c'est-à-dire que l'on aurait répété la procédure de tirage des logements un certain nombre de fois. Le processus de réplication est décrit de la manière suivante :

$$\text{Super } U \left\{ \begin{array}{cc} \xrightarrow{m} & U^{(1)} : \theta^{(1)} = f\left(\mathbf{z}_N^{(1)}\right) \\ \xrightarrow{m} & U^{(2)} : \theta^{(2)} = f\left(\mathbf{z}_N^{(2)}\right) \\ \vdots & \vdots \\ \xrightarrow{m} & U^{(500)} : \theta^{(500)} = f\left(\mathbf{z}_N^{(500)}\right) \end{array} \right\} \xrightarrow{p(\cdot)} s : \left\{ \begin{array}{c} \hat{\theta}^{(1)} = f\left(\mathbf{z}_n^{(1)}\right) \\ \hat{\theta}^{(2)} = f\left(\mathbf{z}_n^{(2)}\right) \\ \vdots \\ \hat{\theta}^{(500)} = f\left(\mathbf{z}_n^{(500)}\right) \end{array} \right. \tag{1.2}$$

où θ désigne un paramètre d'intérêt, par exemple un coefficient de régression β_k de l'équation 1.1, m est le modèle de superpopulation, les populations $U^{(\cdot)}$ sont de taille $N = 3\,581\,991$, l'échantillon s est de taille $n = 484$, $\mathbf{z}_N^{(r)} = \left(z_1^{(r)}, \ldots, z_N^{(r)}\right)^\top$ et $\mathbf{z}_n^{(r)} = \left(z_1^{(r)}, \ldots, z_n^{(r)}\right)^\top$. L'argument dans $f(\cdot)$ est composé que d'une seule caractéristique Z pour simplifiée, mais l'argument pourrait être plusieurs caractéristiques. De même au lieu de θ, un vecteur $\boldsymbol{\theta}$ de paramètres pourrait être d'intérêt. Les réplications ont été réalisées à partir d'un programme codé en SAS.

Ainsi la population de logements a été répliquée 500 fois et un seul échantillon s de logements a été tiré c'est-à-dire un unique vecteur d'indicatrices $(I_1, I_2, \ldots, I_{484})^\top$. Étant donné que pour un logement sélectionné, toutes ses pièces figurant dans le fichier de population des pièces construits sont incluses, le vecteur d'indicatrices pour les pièces est le même entre les 500 échantillons. Dès lors une valeur $z_u^{(r_1)}$, $1 \leq u \leq n$ (resp. $1 \leq u \leq n_j$), de la variable réponse ou d'une covariable pour le niveau

logement (resp. le niveau pièce), n'est différente de la valeur $z_u^{(r_2)}$ qu'à cause des réplications sur U (i.e. pas à cause de $p(\cdot)$).

L'échantillon s a été tiré selon le plan explicité dans la section suivante.

2 Plan de sondage

Afin de pouvoir déterminer la meilleure pondération à utiliser dans l'expression de la pseudo vraisemblance d'un modèle multi-niveaux appliqué à nos données, il est nécessaire de reproduire un plan de sondage $p(\cdot)$ (voir équation 1.2 ci-avant) similaire à celui utilisé dans l'enquête Plomb-Habitat (cf. section 8.2 de la partie « Spécificités des données d'enquête »).

Pour cela il est nécessaire de disposer dans le fichier logement construit précédemment, des informations utilisées dans le plan de sondage de Plomb-Habitat (cf. section 8.2 de la partie « Spécificités des données d'enquête »). Or, certaines informations étaient relatives aux hôpitaux et aux enfants dont nous ne disposons pas dans notre fichier de logements représentant la population d'intérêt. En particulier, la stratification faite au premier degré à partir des zones à risque plomb géographiques construites à partir des bassins de population auxquels appartenait chaque hôpital ; il n'y a d'ailleurs pas d'entité « hôpital » dans le fichier logement. De même la stratification par niveau de plombémie de l'enfant faite lors du sous-échantillonnage (2^e phase) est non disponible car il n'y a pas d'entité enfant dans notre fichier logement.

Dès lors, le plan de sondage utilisé dans la simulation est un plan identique (nombre d'unités échantillonnées à chaque degré, nombre d'unités échantillonnées par strate, *etc.*) à celui utilisé pour les enquêtes Saturn-Inf/Plomb-Habitat excepté pour les seules parties non reproductibles.
Les PSU (hôpitaux) à la phase 1 ont été remplacées par des Établissements Publics de Coopération Intercommunale (EPCI) définis selon l'INSEE[7]. La base de données des EPCI est disponible sur le site de l'INSEE[8]. Les enfants du degré 2 de la phase 1 ont été remplacés par des résidences principales sur la base d'une hypothèse de bijection entre l'ensemble des enfants pouvant être inclus au seconde degré et l'ensemble des résidences principales, puisque seul un enfant d'une fratrie (habitant

7. Les établissements publics de coopération intercommunale (EPCI) sont des regroupements de communes ayant pour objet l'élaboration de « projets communs de développement au sein de périmètres de solidarité ». Ils sont soumis à des règles communes, homogènes et comparables à celles de collectivités locales. Les communautés urbaines, communautés d'agglomération, communautés de communes, syndicats d'agglomération nouvelle, syndicats de communes et les syndicats mixtes sont des EPCI.

8. http://www.insee.fr/fr/ppp/bases-de-donnees/donnees-detaillees/base-cc-table-appartenance-geo-communes/base-cc_table-appartenance-geo-communes-11.zip, accès le 26/07/2012.

dans une unique résidence principale[9]) pouvait être inclus.

La comparaison du plan de sondage utilisé dans la présente simulation avec le plan de sondage de Saturn-Inf/Plomb-Habitat est montré par la table 39.

Au 1er degré de la phase 1, 135 EPCI (PSU) ont été tirés parmi 2594. De 2 à 19 EPCI ont été sélectionnés par strate (région).
Au second degré de la phase 1, 3623 logements (SSU) ont été échantillonnés au sein des 135 EPCI. Un minimum de 3 SSU et un maximum de 45 SSU ont été sélectionnés. Le redressement réalisé à la fin du second degré pour Saturn-Inf/Plomb-Habitat a été reproduit dans la simulation excepté le fait que des informations relatives à l'enfant (âge, sexe, CMUc) n'ont pu être appliquées aux entités logements de la simulation. Dans la partie 3 du redressement seule la post-stratification par ZEAT[10] a pu être conservée. Cette contrainte rend la partie 4 du redressement obsolète pour la partie simulation.

Un sous-échantillon (2e phase) de 1032 logements, stratifié par région, a été ensuite sélectionné pour représenter les 1032 ménages qui ont accepté de participer à l'enquête environnementale Plomb-Habitat. 484 logements ont été ensuite échantillonnés avec un minium de 10 logements et un maximum de 173 logements par strate (région). La post-stratification réalisée sur les poids des logements et décrite en section 1.1 du chapitre 2 a été reproduite.
Une fois les logements sélectionnés, toutes leurs pièces issues du fichier pièce construit précédemment, ont été incluses. Étant donné que le nombre de pièces par logement n'était pas contrôlé, le nombre total de pièces fournies par le tirage de la partie simulation diffère légèrement du nombre de pièces au sein des 484 logements de Plomb-Habitat (1873 pièces au lieu de 1834).

L'échantillonnage a été réalisé sous le logiciel SAS en particulier avec la procédure `surveyselect` pour les étapes de tirage. Le code n'est pas montré car incompréhensible dans la mesure où les tables SAS sur lesquelles le code se base ne sont pas disponibles hors version en fichier informatique. Le code et les tables relatives sont néanmoins à disposition.

3 Critères de jugement de la meilleure pondération

Par le principe de l'équation 1.2 ci-avant, les coefficients de régression et les paramètres de variance ($(\theta = (\beta_0, \ldots, \beta_{30}, \sigma_1^2, \sigma_2^2)^\top$) ont été estimés par pseudo

9. Le nombre de résidence principales est identique au nombre de ménages selon l'INSEE. Cependant on note que : depuis 2005, une résidence principale peut comporter plusieurs ménages si ceux-ci ont des budgets séparés.

10. Rappel : Zone d'Études et d'Aménagement du Territoire, correspondant à un regroupement de régions administratives.

TABLE 39 – Comparaison du plan de sondage de Saturn-Inf/Plomb-habitat avec celui de la simulation ; « (ou d'un ...) » est relatif à la partie simulation.

	Saturn-Inf/Plomb-Habitat	***n***	**Simulation**	***n***
Phase 1				
1er degré	Hôpital	135	EPCI	135
Stratification	Par région administrative et par zone géographique à risque plomb		Par région administrative	
	Poids de sondage (= poids conditionnel) d'un hôpital (ou d'un EPCI) : $1/\pi_k$			
2e degré	Enfant	3623	Logement	3623
	Poids conditionnel d'un enfant (ou d'un logement) : $1/\pi_{j\|k}$ Poids de sondage d'un enfant (ou d'un logement) : $w_j^a = 1/\pi_{j\|k} \times \pi_k$			
Redressement partie 1	La somme des poids des enfants de l'hôpital k est égale à l'effectif d'enfants hospitalisés dans l'hôpital k en 2008.		La somme des poids des logements de l'EPCI k est égale à l'effectif de logements dans l'EPCI k en 2008.	
Redressement partie 2	La somme des poids des enfants inclus dans une strate h est égale au nombre d'enfants réellement hospitalisés dans la strate h en 2008		La somme des poids des logements inclus dans une strate h est égale au nombre de logements présents dans la strate h en 2008	
Redressement partie 3	La somme des poids des enfants est égale au nombre réel d'enfants par ZEAT par classe d'âge, selon le sexe et selon le fait de bénéficier de la CMUc.		La somme des poids des logements est égale au nombre réel de logements par ZEAT	
Redressement partie 4	La somme des poids des enfants est égale au nombre réel d'enfants par ZEAT		Obsolète	
	Poids de sondage redressé d'un enfant (ou d'un logement) : $\widetilde{w}_j^a = w_j^a \times coef_{redressement}^1$			
Phase 2				
1er degré	Enfant ou Logement	484	Logement	484
Stratification	Région et niveau de plombémie		Région	
	Poids « conditionnel » d'un enfant (ou d'un logement) : $1/\pi_j^b$ Poids de sondage d'un enfant (ou d'un logement) : $\omega_j^b = \widetilde{w}_j^a/\pi_j^b$			
Redressement partie 5	La somme des poids des enfants inclus dans une région d'habitation donnée est égale au nombre d'enfants recensés dans cette région		La somme des poids des logements inclus dans une région d'habitation donnée est égale au nombre de logements recensés dans cette région	
	Poids de sondage d'un enfant (ou d'un logement) : $w_j^b = \omega_j^b \times coef_{redressement}^2$			
Redressement partie 6	La somme des poids des logements égale le nombre réel de logements par post-strate finale.		La somme des poids des logements égale le nombre réel de logements par post-strate finale.	
	Poids de sondage redressé d'un logement (ou d'un logement) : $\widetilde{w}_j^b = w_j^b \times coef_{redressement}^3$			
2e degré	Pièce	1834	Pièce	1873
	Poids conditionnel d'une pièce (ou d'une pièce) : $1/\pi_{i\|j} = 1$ Poids de sondage d'une pièce (ou d'une pièce) : $w_i = \widetilde{w}_j^b$			

maximum de vraisemblance (cf. équation 6.8 de la partie « Spécificités des données d'enquête »). On dispose alors d'un tableau de la forme indiquée par la table 40 et ceci pour chacun des 9 scénarios, $\mathbf{w_1}, \ldots, \mathbf{w_9}$.

TABLE 40 – Tableau disponible pour chaque scénario après estimation des paramètres du modèle.

	β_0	β_1	$\cdots$	β_{30}	σ_1^2	σ_2^2
U_1	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
U_2	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
$\vdots$	$\hat{\beta}_0^{(r)}$	$\hat{\beta}_1^{(r)}$	$\cdots$	$\hat{\beta}_{30}^{(r)}$	$\hat{\sigma}_1^{2(r)}$	$\hat{\sigma}_2^{2(r)}$
U_{500}	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$

Si on fait l'hypothèse d'une « bonne pondération » $\mathbf{W}$ parmi les 9 scénarios alors elle implique qu'une fois conditionné sur les covariables $\mathbf{X}$ et la pondération $\mathbf{W}$, le plan de sondage $p(\cdot \mid \mathbf{X}, \mathbf{W})$ est non informatif pour Y (cf. section 5 de la partie « Spécificités des données d'enquête »). Autrement dit le modèle m est valable pour les données de l'échantillon s et il est donc possible d'estimer les paramètres du modèle à partir de s.

Ainsi, pour chacun des 9 scénarios il est alors possible d'estimer le biais des estimateurs de $\beta_0, \beta_1, \cdots, \beta_{30}, \sigma_1^2, \sigma_2^2$ à partir de leur table 40 respective en calculant la moyenne de leurs 500 estimations respectives :

$$\widehat{B}(\hat{\theta}) = \frac{1}{500} \sum_{r=1}^{500} (\hat{\theta}^{(r)} - \theta_{\text{fixe}})$$

où θ est mis pour indistinctement un β_k, σ_1^2 ou σ_2^2, avec θ_{fixe} désigne sa vraie valeur que l'on a fixée dans la population.

Afin de juger les scénarios, le biais individuel des estimateurs de chacun des paramètres $\beta_0, \beta_1, \cdots, \beta_{30}, \sigma_1^2, \sigma_2^2$, est considéré. Pour pouvoir estimer la distribution des 33 valeurs de biais, le biais relatif $B_R(\hat{\theta})$ est considéré et estimé par $\widehat{B}(\hat{\theta})/\theta_{\text{fixe}}$.

Si le biais ne permet pas de départager les scénarios et donc la pondération au niveau 2, le critère d'efficacité est considéré dans un second temps, via l'estimation de la variance des estimateurs, menant alors à l'estimation de l'erreur quadratique moyenne (EQM) (racine de l'EQM *in fine* : $REQM$). La variance est estimée par :

$$\widehat{V}(\hat{\theta}) = \frac{1}{500 - 1} \sum_{r=1}^{500} \left(\hat{\theta}^{(r)} - \overline{\hat{\theta}} \right)^2$$

où $\overline{\hat{\theta}} = \sum_{r=1}^{500} \hat{\theta}^{(r)}/500$. La variance relative $V_R(\hat{\theta})$ est estimée par $\widehat{V}(\hat{\theta})/\theta_{\text{fixe}}^2$. La $REQM$ est estimée par :

$$\widehat{REQM}(\hat{\theta}) = \sqrt{\widehat{B}(\hat{\theta})^2 + \widehat{V}(\hat{\theta})}$$

et la $REQM$ relative, $REQM_R$, est estimée par $\sqrt{\widehat{B_R}(\hat{\theta})^2 + \widehat{V_R}(\hat{\theta})}$.

Après ces estimations, la table 40 devient la table 41.

TABLE 41 – Tableau disponible pour chaque scénario après estimation des paramètres du modèle et estimations des critères de jugement.

	β_0	β_1	$\cdots$	β_{30}	σ_1^2	σ_2^2
U_1	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
U_2	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
$\vdots$	$\hat{\beta}_0^{(r)}$	$\hat{\beta}_1^{(r)}$	$\cdots$	$\hat{\beta}_{30}^{(r)}$	$\hat{\sigma}_1^{2(r)}$	$\hat{\sigma}_2^{2(r)}$
U_{500}	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
B	$\widehat{B}(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\widehat{B}(\hat{\beta}_{30})$	$\widehat{B}(\hat{\sigma}_1^2)$	$\widehat{B}(\hat{\sigma}_2^2)$
V	$\widehat{V}(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
$REQM$	$\widehat{REQM}(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
B_R	$\widehat{B_R}(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
V_R	$\widehat{V_R}(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$
$REQM_R$	$\widehat{REQM}_R(\hat{\beta}_0)$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$

À partir des 9 tables 41 il est possible de comparer pour chaque estimateur individuel de $\beta_0, \beta_1, \cdots, \beta_{30}, \sigma_1^2, \sigma_2^2$ leurs 9 bias, leurs 9 variances et leurs 9 $REQM$. Afin de synthétiser l'information du biais, de la variance et du $REQM$ sur les 33 paramètres estimés, la distribution du biais relatif, de la variance relative et de la $REQM$ relative peuvent être tracées sous forme de boîtes à moustaches pour chacun des 9 scénarios. Les résultats seront présentés en ce sens.

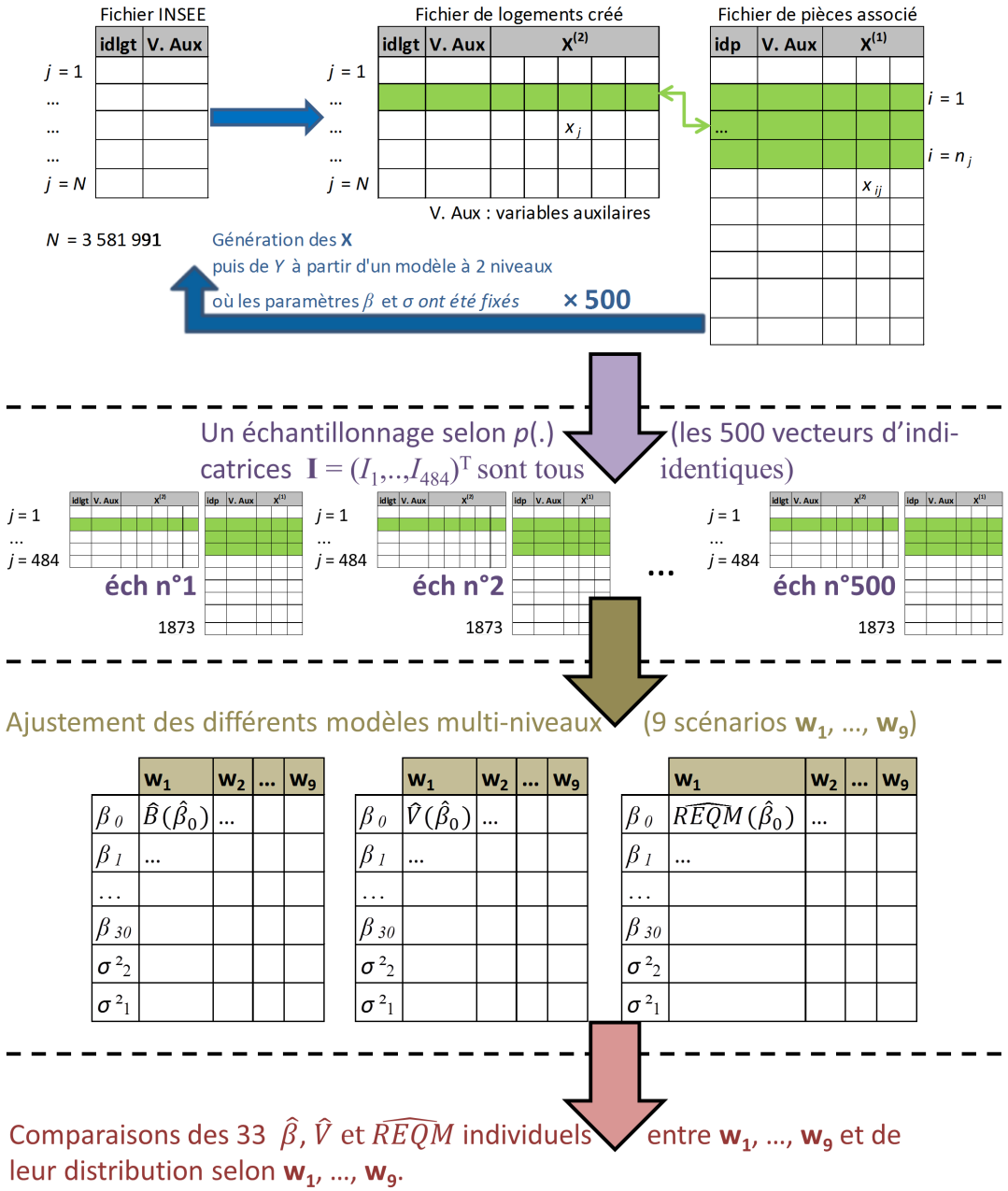
La figure 37 schématise et résume la stratégie de simulation afin de déterminer quelle pondération utiliser pour le niveau 2 du modèle multi-niveaux envisageable pour nos données.

4 Résultats

La table 42 affiche les 3 meilleurs scénarios parmi les 9 pour chaque estimateur de chacun des paramètres $\beta_0, \beta_1, \cdots, \beta_{30}, \sigma_1^2, \sigma_2^2$. On rappelle que la définition des 9 scénarios et en particulier à travers la pondération du niveau 2 (logement) est donnée en section 3 du chapitre 3.

En ce qui concerne le biais (en valeur absolue), chacun des 9 scénarios apparaît au moins une fois à travers les 33 paramètres comme le meilleur (au sens du plus faible biais en valeur absolue). Les valeurs détaillées du biais par scénario et pour chaque paramètre sont données en annexe 11. La distribution du biais relatif des 33 paramètres estimés selon le scénario est montrée avec la figure 38. Toutes les distributions du biais relatif sont autour de 0. La plupart des estimateurs, quelque soit les scénarios, sont peu biaisés. Cela est d'autant plus vrai pour les scénarios $\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_7}$ et $\mathbf{w_8}$. Par exemple pour le scénario $\mathbf{w_1}$ le percentile d'ordre 25 du biais relatif est égal à $-1,26\%$ et le percentile d'ordre 75 est égal à $0,82\%$ (valeurs des

FIGURE 37 – Stratégies de simulation et de comparaison des différentes pondérations au niveau 2.



percentiles non affichés). Des valeurs extrêmes[11] existent néanmoins pour chaque scénario : 7 valeurs pour $\mathbf{w_1}$, 5 pour $\mathbf{w_2}$, 4 pour $\mathbf{w_3}$, 7 pour $\mathbf{w_4}$, 4 pour $\mathbf{w_5}$, 7 pour $\mathbf{w_6}$, 7 pour $\mathbf{w_7}$, 5 pour $\mathbf{w_8}$, 5 pour $\mathbf{w_9}$. Le biais relatif ne permet pas d'obtenir un scénario comme le meilleur à travers les 33 valeurs individuelles. En revanche, à travers la considération globale des 33 valeurs par une distribution, 4 scénarios ($\mathbf{w_1}$,

11. Au sens où elles se situent au-delà de la moustache supérieure et en-deçà de la moustache inférieure respectivement. La moustache supérieure est égale à la valeur parmi les 33, la plus grande parmi les valeurs inférieures à P25 + 1,5× (P75 - P25). La moustache inférieure est égale à la valeur parmi les 33, la plus petite parmi les valeurs supérieures à P25 - 1,5× (P75 - P25).

$\mathbf{w_5}$, $\mathbf{w_7}$ et $\mathbf{w_8}$) semblent fournir des biais relatifs proches de 0 et moins dispersés (figure 38).

En ce qui concerne l'efficacité des estimateurs en fonction des 9 scénarios, les 3 meilleurs (au sens de la plus faible variance) scénarios à travers les 33 paramètres sont le $\mathbf{w_1}$, qui est le meilleur excepté pour 3 paramètres (β_{24}, β_{25} et β_{30}) ; $\mathbf{w_8}$ est le second et $\mathbf{w_1}$ le 3e. Si on regarde comment se classent les 9 scénarios, et non pas que les 3 meilleurs, suivent dans l'ordre $\mathbf{w_7}$, $\mathbf{w_3}$, $\mathbf{w_2}$, $\mathbf{w_4}$ puis $\mathbf{w_9}$ et $\mathbf{w_6}$ étant à tour de rôle les 2 plus mauvais scénarios selon le paramètre (résultats non montrés). La considération des 33 variances sous forme de distribution montre que, comme pour le biais relatif, les 4 scénarios ($\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_7}$ et $\mathbf{w_8}$) semblent fournir des variances relatives plus faibles et moins dispersées (figure 39).

Pour la REQM les constats sont identiques à ceux faits pour la variance à partir de la table 42 et de la figure 40. La figure 41 illustre graphiquement pour la REQM à travers quelques paramètres, comment se classe chaque scénario par rapport aux autres. Cette figure permet de voir un classement des scénarios presque toujours identique à travers les paramètres affichés. D'autre part elle permet de voir que les résultats des 4 scénarios, $\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_7}$ et $\mathbf{w_8}$, sont proches et semblent bien meilleurs que les résultats des 5 autres scénarios.

FIGURE 38 – Distribution des 33 valeurs du biais relatif pour chaque scénario.

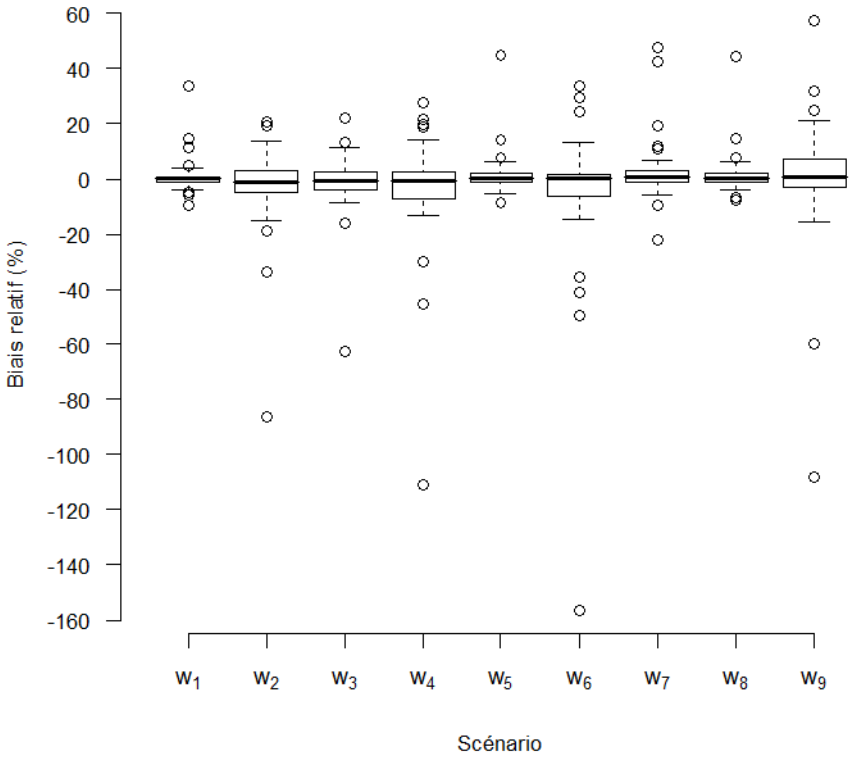


TABLE 42 – Meilleurs scénarios (*top 3*) pour le biais, la variance et la REQM de chaque estimateur des paramètres du modèle.

Coef.	**Vraie valeur**	**top3**($\lvert\widehat{B}\rvert$)	**top3**($\widehat{V}$)	**top3**($\widehat{REQM}$)
β_0	-0,451	$\mathbf{w_7}$, $\mathbf{w_3}$, $\mathbf{w_2}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_1	-0,185	$\mathbf{w_7}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_2	-0,262	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_3	0,435	$\mathbf{w_6}$, $\mathbf{w_1}$, $\mathbf{w_5}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_4	1,572	$\mathbf{w_9}$, $\mathbf{w_7}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_5	0,170	$\mathbf{w_6}$, $\mathbf{w_2}$, $\mathbf{w_9}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_6	-0,247	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_7	0,170	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_8	-0,007	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_9	0,211	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{10}	0,037	$\mathbf{w_7}$, $\mathbf{w_8}$, $\mathbf{w_5}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{11}	-0,072	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{12}	0,030	$\mathbf{w_1}$, $\mathbf{w_6}$, $\mathbf{w_5}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{13}	0,084	$\mathbf{w_7}$, $\mathbf{w_1}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{14}	0,473	$\mathbf{w_7}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{15}	0,105	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_7}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{16}	0,052	$\mathbf{w_1}$, $\mathbf{w_8}$, $\mathbf{w_5}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{17}	0,099	$\mathbf{w_9}$, $\mathbf{w_4}$, $\mathbf{w_3}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{18}	0,114	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_7}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{19}	0,297	$\mathbf{w_1}$, $\mathbf{w_9}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{20}	0,020	$\mathbf{w_6}$, $\mathbf{w_3}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{21}	0,399	$\mathbf{w_1}$, $\mathbf{w_9}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{22}	0,134	$\mathbf{w_4}$, $\mathbf{w_3}$, $\mathbf{w_2}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{23}	-0,069	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{24}	0,261	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_7}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{25}	0,125	$\mathbf{w_2}$, $\mathbf{w_6}$, $\mathbf{w_3}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{26}	0,022	$\mathbf{w_3}$, $\mathbf{w_2}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{27}	0,248	$\mathbf{w_6}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{28}	0,263	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_7}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{29}	0,133	$\mathbf{w_1}$, $\mathbf{w_4}$, $\mathbf{w_5}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{30}	0,015	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
σ_2^2	0,800	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$
σ_1^2	0,450	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$

FIGURE 39 – Distribution des 33 valeurs de la variance relative pour chaque scénario (valeurs extrêmes non tracées).

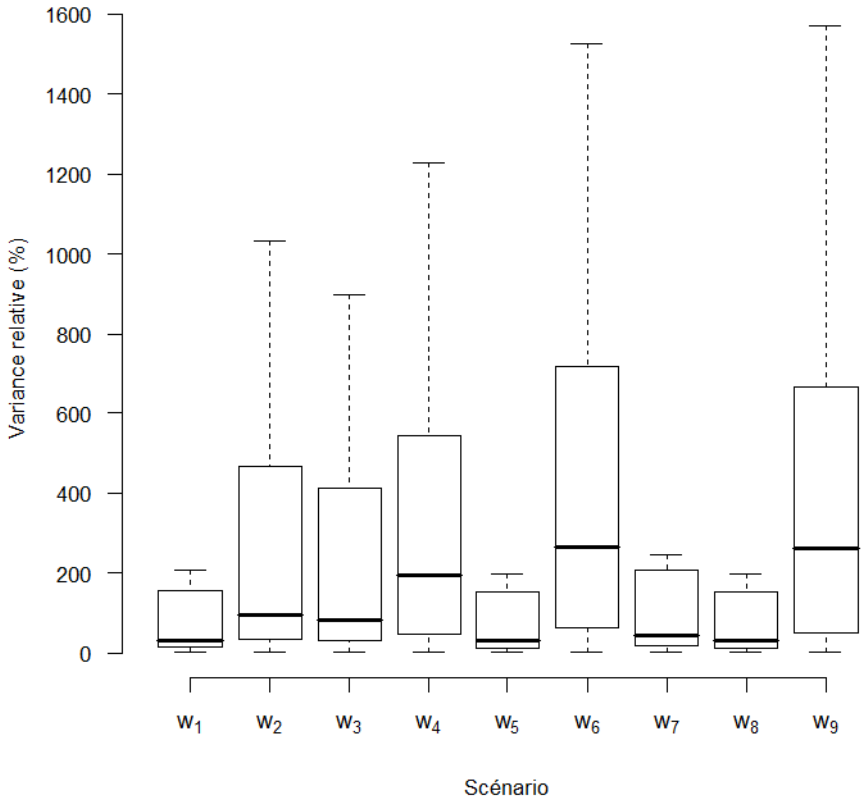
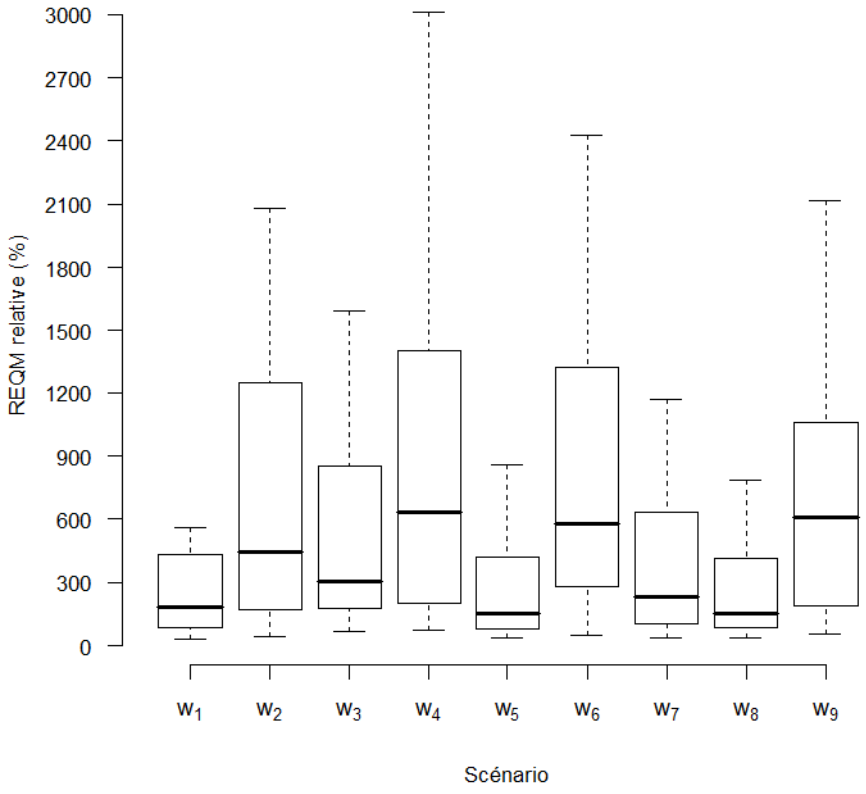


FIGURE 40 – Distribution des 33 valeurs de la REQM relative pour chaque scénario (valeurs extrêmes non tracées).



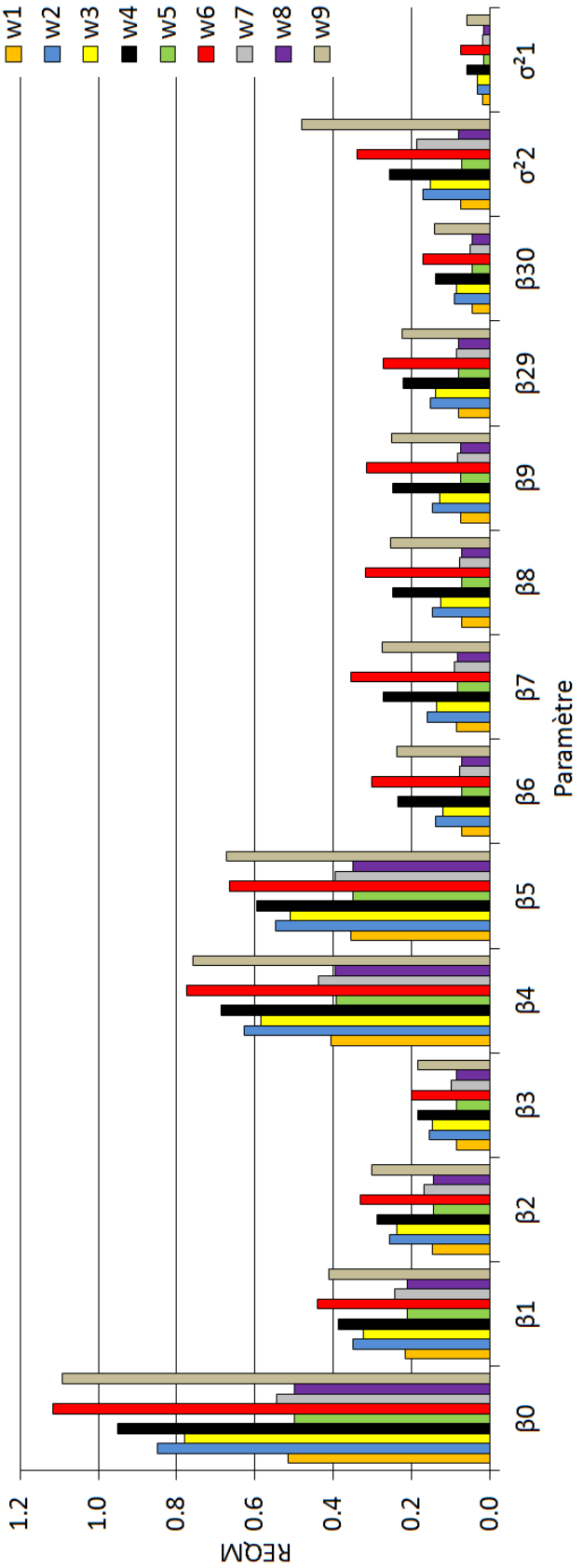


FIGURE 41 – Estimations de la REQM pour quelques paramètres.

5 Synthèse

Dans ce chapitre l'impact de la pondération dans un modèle multi-niveaux a été étudiée. **La pondération au niveau 2 a pu être précisément étudiée** dans le cadre de modèle à 2 niveaux et à 3 niveaux. L'impact de la pondération a été étudié en particulier dans la situation où les niveaux du modèle multi-niveaux ne correspondent pas totalement à la hiérarchie du plan de sondage.

Les résultats de cette étude de simulation **ont permis de valider l'utilisation** au chapitre 3 **d'un modèle à 2 niveaux non pondéré**. Ce modèle ajusté avait permis d'estimer la contribution respective des sources en plomb à contaminer la poussière intérieure au sol.

Les résultats de cette étude de simulation montrent qu'en effet, un modèle à 2 niveaux sans pondération (ce qui a été appelé $\mathbf{w_5}$) est le plus adéquate pour nos données. Néanmoins les résultats montrent que si une pondération au niveau 2 (logement) avaient été utilisée, l'utilisation de poids conditionnels (ce qui a été appelé $\mathbf{w_1}$) aurait donnée des estimations très proches.

Les résultats de cette étude de simulation montrent que si un modèle à 3 niveaux avaient été ajusté, un modèle non pondéré (ce qui a été appelé $\mathbf{w_8}$) aurait données des estimations similaires à celles obtenues avec un modèle à 2 niveaux ($\mathbf{w_5}$ ou $\mathbf{w_1}$). De même les estimations auraient été proches avec un modèle à 3 niveaux dont les poids au niveau 2 auraient été les poids conditionnels des logements (ce qui a été appelé $\mathbf{w_7}$).

Alors que des études de simulation de la littérature semblent n'avoir été faites que sur des données complètement artificielles, **notre étude de simulation s'est basée sur des données réelles**. De plus alors que les études de simulation de la littérature semblent ne s'être intéressées qu'à l'impact de la pondération au niveau 1 d'un modèle multi-niveaux, **notre étude de simulation apporte pour la première fois des résultats relatifs à l'impact de la pondération au niveau 2**.

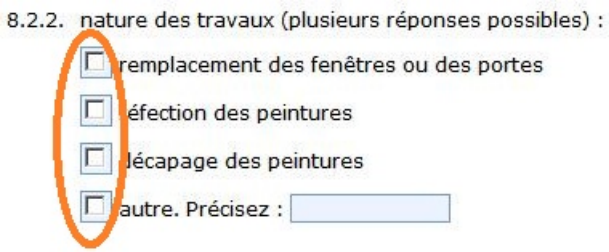
Discussion

1 Validation des données

La validation des données collectées dans le cadre de l'enquête Plomb-Habitat sur lesquelles ce travail s'est basé, a permis d'obtenir une base de données avec le moins d'erreurs possible et la plus complète possible. Pour cela, des procédés intégrés dans l'application de saisie des données ainsi qu'un processus de vérification après avoir collecté les données ont été mis en œuvre. Réalisée au fil du déroulement de l'enquête Plomb-Habitat qui, pour rappel a été menée d'octobre 2008 à août 2009, le développement des procédés de validation et le temps de correction ont demandé plusieurs mois de travail.

Néanmoins, des contraintes de calendrier et de coût de développement informatique n'ont pas permis d'atteindre une validation optimale. Les enchainements logiques auraient du être intégrés à l'application de saisie des données. De même, les cohérences auraient amélioré la qualité des données si elles avaient été programmées dans l'application et auraient permis de gagner un temps précieux lors de la validation. La programmation des enchainements et des cohérences dans l'application auraient nécessité de fournir ces enchainements et ces cohérences au pôle informatique avant le développement de l'application. En terme de calendrier, la fourniture de ces éléments n'aurait pas été possible, tout au moins de manière exhaustive. Sur ce point, même après la fin du développement de l'application, des éléments de validations nouveaux sont apparus et auraient nécessité de toute manière une validation post-collecte.

Outre ces contraintes de temps et de calendrier, de simples omissions d'éléments de validation ont induit une validation des données non optimale. Par exemple il n'a pas été demandé au pôle informatique de remplacer les *« check-box »*, comme illustré ci-dessous, par des radio-boutons.



Ces *« check-box »* ne doivent pas être utilisées car lorsqu'elles sont non cochées, elles ne permettent pas de savoir s'il s'agit d'une réponse qui signifie « Non » ou s'il s'agit d'une non-réponse. Chaque modalité de réponse possible doit être associée à un radio-bouton vierge « Oui/Non ». Alors qu'elle a été faite pour les radio-boutons justement, la demande de ne pas associer une valeur par défaut n'a pas été faite de manière exhaustive. Cela a constitué une entrave à la qualité des données, lorsque cette valeur par défaut indiquait une valeur possible de réponse. Par exemple, à la question du nombre de cuisines présentes dans le logement, la valeur 0 avait été mise par défaut. Puisque remplis par défaut, certains items n'ont pas de test associé détectant les données manquantes. Ceci explique pourquoi le nombre de tests détectant les non-réponses est inférieur au nombre d'items du questionnaire.

Bien que devant être idéalement exhaustive pour être optimale, la validation des données l'est difficilement. L'exhaustivité est difficilement atteignable dans la mesure où la validation est réalisée de manière non automatisée, et repose sur un travail humain. Les subtilités de certaines validations, la recherche d'informations annexes et l'interprétation de certains pseudo problèmes dans la validation rendent ainsi une procédure automatisée réalisée par les machines impossible. C'est pourquoi le métier de *« data-manager »* existe et est largement répandu là où les enjeux financiers sont très importants, comme dans l'industrie pharmaceutique par exemple.

2 Estimation des niveaux en plomb dans les compartiments environnementaux en milieu résidentiel

Aucun état descriptif de la contamination par le plomb dans les logements français n'avait été réalisé auparavant. Les données disponibles issues des enquêtes environnementales faisant suite à la déclaration d'un cas de saturnisme ou bien celles collectées dans le cadre de réalisation du CREP, n'ont en effet jamais été centralisées dans une base de données nationale. Même si une telle base existait, elle contiendrait des données ne permettant pas de réaliser un état descriptif national. En effet dans le cas d'enquête environnementale faisant suite à la déclaration d'un cas de saturnisme, les données relevées ne concernent que la population de logements abritant au moins un cas de saturnisme chez l'enfant. L'utilisation des mesures faites dans de tels logements conduit nécessairement à sur-estimer la contamination à laquelle sont exposés les enfants dans la population de résidences principales en Métropole. Dans le cas du CREP, la population décrite n'est que celle des logements construits avant 1949. Enfin, en se basant uniquement sur le protocole CREP il n'est pas possible de décrire les compartiments environnementaux autres que les revêtements.

2.1 Niveau en plomb dans l'eau du robinet

Dans le cadre du contrôle sanitaire, le type de prélèvement d'eau est un prélèvement aléatoire. Le type de prélèvement opéré dans ce cadre représente moins

la quantité moyenne hebdomadaire de plomb ingérée par l'usager qu'un prélèvement fait par stagnation après 30 minutes, si les conclusions du rapport européen sont suivies [European Commission et al., 1999]. Cependant, à partir de ces conclusions, l'échantillonnage aléatoire peut être considéré pour estimer plus simplement la conformité ou non-conformité d'une zone de distribution. En considérant la France comme une même zone, 5 % des prélèvements issus du contrôle sanitaire avaient une concentration en plomb supérieure à 10 μg/L en 2004 [Glorennec et al., 2007].

Des données plus récentes de la base Sise-Eaux ont été obtenues auprès de la Direction Générale de la Santé afin de comparer les concentrations du contrôle sanitaire avec les résultats obtenus dans le présent travail. Les données demandées étaient les concentrations en plomb relevées par le contrôle sanitaire réalisé sur les mêmes communes que les logements enquêtés dans Plomb-Habitat, par année de 2004 à 2009. La distribution de ces concentrations est montrée en table 43. Selon l'année, au plus 5 % (2,9 % dans Plomb-Habitat) des prélèvements avaient une concentration en plomb supérieure à 10 μg/L pour les unités de distributions desservant les communes où des logements ont été enquêtés dans Plomb-Habitat ; et ceci pour chacune des années 2004 à 2009. En se limitant aux prélèvements réalisés en 2008 et 2009 [12], le percentile 95 % est égal à 7,9 μg/L (5,4 μg/L dans Plomb-Habitat) et indique donc de même que moins de 5 % des prélèvements produisaient une concentration supérieure à 10 μg/L.

TABLE 43 – Distribution des concentrations en plomb (μg/L) dans l'eau du robinet mesurée par le contrôle sanitaire entre 2004 et 2009.

Année	***n***	**min**	**P5**	**P25**	**P50**	**P75**	**P95**	**max**
2004	872	0	0	0	0	0	8	170
2005	928	0	0	0	0	0	10	140
2006	944	0	0	0	0	0	7	65
2007	951	0	0	0	0	0	8	197
2008	959	0	0	0	0	0	7.9	500
2009	944	0	0	0	0	0	8	160

Légende. Px : Percentile d'ordre x%, n : nombre de prélèvements.
Nota Bene. Source : Ministère chargé de la santé – ARS – Sise-Eaux – Traitement : CSTB. Les concentrations en plomb de la base sont représentatives uniquement de la qualité de l'eau au niveau de l'immeuble, voire du logement, où le prélèvement a été réalisé, et non de la qualité générale de l'eau qui alimente l'unité de distribution.

Rappel du résultat obtenu dans Plomb-Habitat :

Année	***n***	**min**	**P5**	**P25**	**P50**	**P75**	**P95**	**max**
2008-09	472	< 1	< 1	< 1	< 1	1,1	5,4	74

12. Années sur lesquelles l'enquête Plomb-Habitat a été réalisée pour rappel.

Pour pouvoir respecter les seuils de la directive numéro 98/83/CE transposés dans l'arrêté du 11 janvier 2007[13], le remplacement des canalisations et des branchements en plomb est la solution la plus naturelle s'il s'agit notamment d'atteindre le seuil de 10 μg/L ; la teneur de 25 μg/L peut être en outre respectée par un traitement approprié de l'eau. Les branchements en plomb entre le réseau public et le réseau privé sont remplacées par les opérateurs privés. Dans un communiqué de presse daté du 14 octobre 2009, Lyonnaise des eaux indiquait l'achèvement du remplacement des branchements en plomb sur la rive gauche de Paris. Quant aux remplacements des canalisations intérieures la charge en revient aux propriétaires. Dès lors se pose la question du coût de réalisation de tels travaux, ce qui soulève de vives réactions de la part de certaines associations[14]. Le Haut Conseil de la Santé Publique a émis récemment un avis[15] sur l'analyse et l'évaluation de l'efficacité des actions engagées pour respecter la future limite de qualité de 10 μg/L de plomb dans l'eau du robinet. Sa recommandation n° 2 indique de privilégier une stratégie de traitement de l'eau dans des zones de dépassement de la limite de qualité de 10 μg/L.

Bien que l'information sur la présence de canalisations en plomb dans chaque logement investigué ait été fournie par l'enquêteur, la fiabilité de cette information ne peut être sûre. Cependant, sur cette base, les concentrations en plomb dans l'eau du robinet apparaissent plus élevées en présence de canalisations. De même, les estimations des percentiles d'ordre 75 % et 95 % sont plus élevées lorsque l'on compare les concentrations des logements avec canalisations en plomb et les concentrations de l'échantillon total (P75 : 4,3 vs 1,1 μg/L ; P95 : 5,7 vs 5,4 μg/L). Ce résultat semble donc cohérent, mais eu égard aux caractéristiques physico-chimiques de l'eau, temps de stagnation etc., présence de canalisations en plomb ne signifie pas toujours concentration élevée en plomb. En effet une étude réalisée entre novembre 1993 et mars 1994 [Baron, 1997] en échantillonnage proportionnel, montrait que sur 46 sites sans réseaux en plomb, 18 % avaient une concentration en plomb entre 10,1 et 25 μg/L et 2 % une concentration entre 25 et 50 μg/L, montrant ainsi une contamination autre que par le plomb massif d'une part. Sur 184 sites avec réseaux en plomb, 35 % des ces sites avaient une concentration en plomb strictement inférieure à 10 μg/L et 36 % entre 10 et 25 μg/L, justifiant l'influence des facteurs de confusion sur la solubilité du plomb dans l'eau.
Finalement les données de concentration en plomb dans l'eau du robinet de Plomb-Habitat peuvent être difficilement utilisées de prime abord pour statuer de la présence de canalisations ou branchements en plomb dans l'immeuble si ces concentrations ne peuvent pas être corrigées par un coefficient prenant en compte les différents et multiples facteurs influant sur la solubilité du plomb dans l'eau de robinet. La réglementation, via l'arrêté du 4 novembre 2002[16], prévoit au minimum d'utiliser

13. Arrêté du 11 janvier 2007 relatif aux limites et références de qualité des eaux brutes et des eaux destinées à la consommation humaine mentionnées aux articles R. 1321-2, R. 1321-3, R. 1321-7 et R. 1321-38 du code de la santé publique.

14. Par exemple de l'Institut Européen pour la Gestion Raisonnée de l'Environnement.

15. Le 14 juin 2013. Avis téléchargeable au 28 août 2013 à l'adresse http://www.hcsp.fr/Explore.cgi/Telecharger?NomFichier=hcspa20130614_plombcanalisationeaupotable.pdf.

16. Arrêté du 4 novembre 2002 relatif aux modalités d'évaluation du potentiel de dissolution du plomb pris en application de l'article 36 du décret n° 2001-1220 du 20 décembre 2001 relatif aux eaux destinées à la consommation humaine, à l'exclusion des eaux minérales naturelles.

l'information du pH mesuré sur au moins 12 mois, avec un nombre de mesures fonction du débit en m^3/jour dans l'unité de distribution. Selon le nombre de mesures de pH réalisées sur l'unité de distribution, une valeur de pH est déterminée comme valeur de référence ; une valeur inférieure ou égale à 7 indique un potentiel de dissolution du plomb très élevé ; une valeur comprise dans l'intervalle]7 ; 7,5] indique un potentiel de dissolution élevé ; une valeur comprise dans l'intervalle]7,5 ; 8] indique un potentiel de dissolution moyen ; une valeur strictement supérieure à 8 indique un potentiel de dissolution faible. Dans le cadre d'un prélèvement après stagnation de 30 minutes, l'interprétation suivante des résultats sur la présence de canalisations en plomb peut être faite selon la circulaire DGS/SD 7 A numéro 2004-45 du 5 février 2004 :

- < 5 μg/L : le réseau de distribution (réseau intérieur et branchement public) ne comporte vraisemblablement pas ou très peu de canalisations en plomb ;
- $[5; 10[$ μg/L : la présence de canalisations en plomb est possible dans le réseau de distribution (réseau intérieur et/ou branchement public). En cas de stagnation prolongée de l'eau dans le réseau, des teneurs en plomb dans l'eau supérieures à 10 μg/L peuvent être éventuellement mesurées ;
- ≥ 10 μg/L : la présence de canalisations en plomb est très probable dans le réseau de distribution (réseau intérieur et/ou branchement public).

À partir de cette interprétation, environ 93 % des logements abritant au moins un enfant âgé de 6 mois à 6 ans auraient un réseau de distribution ne comportant vraisemblablement pas ou très peu de canalisations en plomb. À noter que selon les relevés des enquêteurs, 1,9 % des logements posséderaient des canalisations en plomb (versus 7 % donc, si on utilise l'interprétation ci-dessus basée sur la valeur de 5 μg/L) ; mais dans 16 % des logements la présence ou l'absence de canalisation en plomb ne pouvaient être statuées. Dès lors il peut être raisonnablement supposé que parmi ces 16 %, des logements contenant des canalisations en plomb sont présents, faisant du chiffre de 1,9 % une borne inférieure. Ainsi le pourcentage ajusté ($>$ 1,9 % donc) et le pourcentage basé sur l'interprétation DGS (7 %) seraient finalement plutôt cohérents et permettent alors d'estimer que, de 2 à 7 % des logements de la population cible posséderaient des canalisations au plomb.

[Clement et al., 2000] indique qu'après stagnation de 30 minutes, le pH et l'alcalinité (TAC) de l'eau n'ont en fait qu'un très faible impact sur la dissolution du plomb dans l'eau en cas de présence de canalisation en plomb. D'après le modèle établi et dans le cas d'une situation la moins contaminante en plomb (diamètre du tuyau en plomb de 3 cm, TAC faible, pH élevé), la concentration prédite est de 5 μg/L ; dans le cas d'une situation la plus contaminante (diamètre du tuyau en plomb de 2 cm, TAC élevée, pH faible), la concentration prédite est de 9 μg/L. Ainsi le modèle de Clément *et al.* confirme la valeur de 5 μg/L en-deçà de laquelle la présence de canalisation en plomb dans le réseau est peu probable selon la circulaire DGS.

Alors que la Communauté Européenne impose une valeur réglementaire de 10 μg/L à partir de la fin 2013, l'U.S. EPA n'impose aucune valeur sanitaire en ce qui

concerne l'eau du robinet. L'U.S. EPA recommande un traitement dès lors que plus de 10 % d'échantillons d'eau du robinet d'un site enquêté ont une concentration en plomb supérieure à 15 μg/L [U.S. EPA, 2010]. Contrairement à la Communauté Européenne, les États-Unis ont opté pour une démarche prenant en compte les coûts, le bénéfice pour la santé et les capacités technologiques à soustraire le plomb de l'eau du robinet.

2.2 Niveau en plomb dans les poussières intérieures déposées au sol

La seule valeur de référence de la réglementation française (cf. aussi section 3.5 de la partie « De l'exposition au plomb et de la présence du plomb en milieu résidentiel »), égale à 1000μg/m^2, concerne les poussières du sol intérieur prélevées par lingette et est issue de l'arrêté du 12 juillet 1999 concernant le contrôle des locaux après réalisation de travaux d'urgence en vue de vérifier la suppression de l'accessibilité au plomb pris pour l'application de l'article R. 32-4 du code de la santé publique [Lucas, 2011]. Cette valeur n'est pas une valeur sanitaire mais uniquement une valeur à respecter lorsque des travaux ont été réalisés suite à la présence avérée d'un risque d'intoxication au plomb. Notre étude montre que 0,2 % ($IC_{95\%}$ = 0-0,4) des logements abritant au moins un enfant âgé de 6 mois à 6 ans possèdent au moins une pièce où la concentration en plomb acido-soluble est supérieure à 1000μg/m^2 (résultats non montrés). Ce seuil de 1000μg/m^2 semble être uniquement utilisé en France ce qui ne permet pas de comparaison avec d'autres pays.

La valeur américaine dans le même contexte d'après travaux (« *clearance standard for dust following an abatement* ») est de 430,5 μg/m^2 en plomb total (40 μg/ft^2) [U.S. EPA, 2001]. Elle est donc identique à la valeur de référence en termes de risque plomb dans la réglementation américaine (« *dust-lead hazard standard* »). En appliquant ce seuil aux données utilisées dans le présent travail on a obtenu que :

- Approximativement 0,21 % des logements français abritant au moins un enfant âgé de 6 mois à 6 ans ont une concentration moyenne en plomb total supérieure à ce seuil. La prévalence de tels logements en France paraît relativement faible.
- Cette prévalence ne semble pas plus élevée chez les logements les plus anciens, puisque pour les logements construits avant 1949, elle est estimée à 0,18 % ($IC_{95\%}$ = 0-0,4).
- Bien que faible, la prévalence des parties communes dont la concentration moyenne dépasse le seuil de 430,5 μg/m^2 est plus élevée : 4,1 % des parties communes d'immeuble où il existe au moins un logement abritant un enfant âgé de 6 mois à 6 ans ont une concentration moyenne en plomb dans les poussières supérieure à 430,5 μg/m^2 en plomb total.

Cependant cette valeur de 40 μg/ft^2 a été remise en cause récemment [Dixon et al., 2009]. Dixon *et al.* pensent qu'une valeur entre 6 et 12 μg/ft^2 devrait être utilisée aux États-Unis afin de protéger la plupart des enfants vivant dans des logements

datant d'avant 1978 [17] d'avoir une plombémie supérieure à 100 μg/L. Environ 1,1 % ($IC_{95\%}$ = 0,3-1,9) des logements français abritant au moins un enfant âgé de 6 mois à 6 ans ont une concentration moyenne supérieure à 12 $\mu g/ft^2$ (129,1 $\mu g/m^2$) en plomb total dans les poussières ; 4,8 % ($IC_{95\%}$ = 2,4-7,2) des logements français abritant au moins un enfant âgé de 6 mois à 6 ans ont une concentration moyenne supérieure à 6 $\mu g/ft^2$ (64,6 $\mu g/m^2$) en plomb total dans les poussières.

Une valeur non réglementaire égale à 300 $\mu g/m^2$ en plomb acido-soluble devant protéger les enfants d'avoir une plombémie supérieure à 100 μg/L avait été proposée en 1994 en France par le Comité Technique plomb. Cette valeur était appelée « seuil de positivité ». Elle avait été déterminée sur la base d'études non publiées du Laboratoire d'Hygiène de la Ville de Paris [Bretin, 2006]. Environ 0,18 % ($IC_{95\%}$ = 0-0,4) des logements abritant au moins un enfant âgé de 6 mois à 6 ans dépassent ce seuil dans notre étude.

En France, les études où des prélèvements de poussières ont été réalisés dans les logements sont rares. En 2005, sur un échantillon de 40 logements bretons volontaires, 90 % de ces logements avaient une valeur moyenne en plomb total dans leurs poussières inférieure à 30 $\mu g/m^2$ [Glorennec et al., 2005]. Bien que l'échantillon ne corresponde pas à celui de l'étude Plomb-Habitat, ces résultats obtenus sur une sous-population de logements semblent globalement analogues. En effet, à titre informatif, le percentile d'ordre 90 % des charges moyennes en plomb total de Plomb-Habitat est estimé à 38,7 $\mu g/m^2$.
Aux États-Unis, des prélèvements de poussières de sol intérieur ont été réalisées dans plusieurs études. La Table 44 rapporte les résultats présentés par Jacobs *et al.* [Jacobs et al., 2002] et Gaitens *et al.* [Gaitens et al., 2009], obtenus respectivement à partir d'une enquête sur 831 logements américains réalisée entre 1998 et 2000 par le HUD et lors de l'enquête NHANES (*National Health and Nutrition Examination Survey*) (1999-2004) dans 2065 logements. Pour l'étude du HUD, aucune restriction sur l'occupation par des enfants en bas âge n'était faite sur les logements inclus dans l'enquête ; l'enquête NHANES concerne des logements où vivent des enfants âgés de 12 à 60 mois. Les modes d'agrégation ne sont pas tout à fait identiques à ceux de Plomb-Habitat, ce qui empêche une comparaison exacte des niveaux en plomb. Néanmoins dans NHANES et Plomb-Habitat, les niveaux de charges sont voisins en termes de moyennes arithmétique ou géométrique si les charges en chambre de l'enfant dans Plomb-Habitat sont considérées. Les charges moyennes de l'étude du HUD dont les données deviennent anciennes, se rapprocheraient des concentrations maximales de Plomb-Habitat.

En dehors du fait que les niveaux en plomb dans les poussières semblent être faibles, relativement à la valeur seuil utilisée actuellement outre Atlantique, les résultats de notre travail montrent que la dispersion des niveaux en plomb dans les

17. Date à partir de laquelle le plomb a été interdit aux USA dans les peintures utilisées en logement résidentiel.

TABLE 44 – Comparaison des résultats de Plomb-Habitat avec les charges en plomb total documentées aux États-Unis.

	HUD (USA) (1998-2002)	**NHANES (USA) (1999-2004)**	**Plomb-Habitat (FR) (2008-2009)**		
Critère d'agrégation	Non précisé[(1)]	Pièce déclarée la plus fréquentée par l'enfant	Chambre de l'enfant*	Moyenne sur les pièces investiguées	Maximum sur les pièces investiguées*
n	3894 échantillons ↕ 831 lgts	2065 échantillons ↕ 2 065 lgts	455 échantillons ↕ 455 lgts	1961 échantillons ↕ 471 lgts	471 échantillons ↕ 471 lgts
Moyenne arithmétique	13,6μg/ft^2 (146,4μg/m^2)	1,34μg/ft^2 (14,4μg/m^2)	12,2μg/m^2	18,8μg/m^2	36,3μg/m^2
Erreur standard	483,5μg/ft^2 (5204,35μg/m^2)	0,14μg/ft^2 (1,54μg/m^2)	1,4μg/m^2	2μg/m^2	4,6μg/m^2
Moyenne géometrique	1,1μg/ft^2 (11,8μg/m^2)	0,52μg/ft^2 (5,6μg/m^2)	5,8μg/m^2	8,8μg/m^2	14,4μg/m^2
Erreur standard géométrique	3,8μg/ft^2 (40,9μg/m^2)	1,05μg/ft^2 (11,3μg/m^2)	1,1μg/m^2	1,1μg/m^2	1,1μg/m^2
P25	0,375μg/ft^2 (4μg/m^2)	?	3μg/m^2	3,7μg/m^2	6μg/m^2
P50	0,9μg/ft^2 (9,7μg/m^2)	?	6μg/m^2	8μg/m^2	14μg/m^2
P75	2,0μg/ft^2 (21,5μg/m^2)	?	12μg/m^2	17,3μg/m^2	29μg/m^2
P90	6,0μg/ft^2 (64,6μg/m^2)	?	22μg/m^2	38,7μg/m^2	67,8μg/m^2
P95	13,2μg/ft^2 (142,1 μg/m^2)	?	42,3μg/m^2	62,6μg/m^2	114μg/m^2

Légende. Px : Percentile d'ordre x%, [(1)] : On suppose que le critère d'agrégation est la moyenne sur les pièces de chaque logement, car une comparaison à la valeur fédérale de 40 μg/ft^2 est réalisée dans l'étude, * : résultats non décrits en section 2.2 du chapitre 2.

poussières est d'autant plus grande que les logements sont anciens. Cette constatation est *a priori* cohérente : la prévalence des logements où persiste de la peinture à base de plomb est plus grande dans les logements anciens (cf. section 2.3 à ce sujet), la peinture au plomb pouvant contaminer la poussière du logement d'après plusieurs études. *A posteriori*, les peintures contaminées ne semblent plus la raison de cette constatation. En effet les résultats du chapitre 3 et discutés ci-après ont montré que la peinture intérieure dégradée à base de plomb était peu contributrice au plomb de la poussière.

Les niveaux de charge en plomb dans les poussières intérieures (hors parties communes) ne diminuent pas systématiquement avec la période de construction des logements : les logements de la période 1975-1993 ont des concentrations en plomb dans leurs poussières du même niveau que les concentrations des logements datant d'avant 1949. Pour les parties communes, les niveaux de concentrations sont de plus en plus faibles lorsque la période de construction des bâtiments est de plus en plus récente.

Outre les peintures, d'autres sources ainsi que des facteurs de confusion peuvent jouer un rôle sur les niveaux en plomb dans les poussières intérieures. En effet les résultats indiquent que l'habitat urbain possède des niveaux en plomb dans ses poussières sensiblement plus élevés que les logements en zone rurale. Les hypothèses avancées pour expliquer ces concentrations plus élevées peuvent être l'utilisation

passée de carburant plombé, l'embellissement des villes par le blanchiment des bâtiments et la lutte contre l'humidité des murs avec de la peinture à base de céruse ou encore l'impact de la pollution industrielle.

Les résultats montrent de plus que les poussières des parties communes sont substantiellement plus contaminées que celles de l'intérieur des logements. On peut penser qu'il est plus difficile de dépenser une somme conséquente pour refaire les parties communes, votée en assemblée générale de copropriété, que pour rénover l'intérieur de son propre appartement (décision individuelle). Cependant la loi du 9 août 2004 relative à la politique de santé publique par son article L1334-8 indiquait que toute partie commune d'un immeuble collectif affecté en tout ou partie à l'habitation et datant d'avant 1949 devait avoir fait l'objet d'un constat de risque d'exposition au plomb au plus tard en août 2008. Ce dernier ne se base néanmoins pas sur des prélèvements de poussières au sol, mais sur des mesures à fluorescence X qui indiquent l'obligation de réaliser des travaux dès lors qu'au moins une unité de diagnostic a une concentration surfacique en plomb supérieure à 1 mg/cm^2 et un revêtement dégradé. Il serait donc intéressant d'avoir le bilan des diagnostics qui doivent avoir été réalisés à ce jour [18] dans les parties communes des immeubles anciens, et de connaître la part des immeubles exposant à un risque plomb du fait d'une contamination des parties communes. Puisque qu'il n'y a pas de transmission du rapport du diagnostic au représentant de l'État (sauf pour les cas d'insalubrité avérés), aucune statistique nationale ne peut toutefois être obtenue à ce sujet à ce jour.

2.3 Niveau en plomb dans les revêtements intérieurs

La table 45 résume les résultats obtenus en section 2.3 du chapitre 2 concernant la prévalence de logements français abritant au moins un enfant âgé de 6 mois à 6 ans concernés par des revêtements plombés.

Les parties communes semblent plus à risque (peinture au plomb dégradée) alors que, comme rappelé en section 2.2 ci-avant, depuis août 2008 c'est-à-dire à l'époque du début des enquêtes de Plomb-Habitat, les parties communes d'immeubles collectifs datant d'avant 1949 doivent avoir fait l'objet d'un CREP.

TABLE 45 – Prévalences de logements avec au moins une UD dont le revêtement contient du plomb ($\geq$ 1 mg/cm^2) et avec au moins une UD dont le revêtement est à risque ; prévalences en parties communes entre parenthèses.

Support du revêtement	**$\geq$ 1 mg/cm^2**	**$\geq$ 1 mg/cm^2 et dégradé**
Tout support	24,5 % (34,2)	4,7 % (7,1)
Support non métallique (céruse visée)	19,3 % (19,9)	3,3 % (7)
Support métallique (minium visé)	10,4 % (17,7)	1,4 % (0,3)

18. Mais aussi au début de l'enquête Plomb-Habitat en théorie donc.

Aucun article scientifique ne semble traiter la céruse séparément dans leur analyse de prévalence des logements ayant des revêtements à base de plomb. Tous traitent de la peinture au plomb sans distinguer le support du revêtement afin d'isoler autant que possible ce composé du plomb. Pourtant la céruse a été le sujet central de la 13ème convention adoptée en 1921 durant la Conférence Générale de l'Organisation Internationale du Travail [ILO, 1921]. Le présent travail qui a fait l'objet d'une publication en 2012 [Lucas et al., 2012] semble donc être le premier en la matière.

Concernant la céruse, l'analyse par période de construction des logements a montré un premier résultat attendu : les logements les plus anciens, c'est-à-dire construits avant 1949, ont la plus forte prévalence parmi les logements ayant encore des revêtements sur support non métallique contenant du plomb. Ce résultat était attendu dans l'hypothèse où la réglementation sur l'utilisation de la céruse dans la peinture, existant depuis la fin du XIXe siècle, a joué son rôle tout au long du siècle suivant.
Cette même prévalence, pour les logements construits à partir de 1949 jusqu'à 1974, s'est cependant avérée non négligeable puisque à hauteur de 22 %. Un tel résultat n'était pas attendu dans la mesure où, dans l'esprit commun, la céruse n'a plus été utilisée depuis 1949. L'analyse de la réglementation [Lucas, 2011] s'est avérée forte utile à ce sujet. L'analyse montrait que rien ne justifiait la considération de l'année 1949 comme date couperet à partir de laquelle la céruse ne devait plus avoir été employée dans la peinture. L'interdiction mentionnée dans l'arrêté du 30 décembre 1948 figurait déjà dans le code du travail depuis 1926. Par la refonte du code du travail qui a eu lieu en 1948, cette date de 1949 constituait tout au plus un rappel à la réglementation. Cette prévalence de 22 % a ainsi fait l'objet d'une analyse plus approfondie pour expliquer ce résultat.
Les tables 22 et 23 présentées dans la section Résultats 2.3 ont été ajoutées. Le but était d'une part d'appréhender la fréquence des UD à revêtements plombés dans les logements d'après 1949 en comptant le nombre de telles UD au delà de 2 UD : étaient-elles marginales ou pas ? D'autre part il s'agissait en remontant le seuil de définition de 1 mg/cm^2 à 2 mg/cm^2 d'éviter autant que possible de compter des UD dont le revêtement contenait des siccatifs à base de plomb et non de la céruse, la réglementation visant cette dernière. Cette prévalence passait de 22,1 % à 3,7 % dès lors que l'on comptait au moins 2 UD dont le revêtement contenait au moins 2 mg/cm^2 de plomb. Soit une diminution de 83 % de la prévalence pour les logements construits entre 1949 et 1974 alors que pour les logements construits avant 1949 cette diminution est de 54 %. Ainsi, bien que de prime abord le résultat concernant les logements construits entre 1949 et 1974 paraissait inquiétant dans le sens où la réglementation ne semblait pas avoir joué son rôle, il convient d'indiquer que les UD à base de plomb dans ces logements semblent être résiduelles et avec des charges plus faibles que dans les logements plus anciens.

Lors de la présentation de ces résultats dans le cadre d'une réunion du Comité Technique Plomb le 4 juillet 2011, les membres de ce comité souhaitaient savoir si la prévalence de la céruse diminuait au cour du XXe siècle, comme semblaient l'indiquer

ces résultats, mais en tenant compte aussi des logements construits avant 1915. Cette date de 1915 indique l'année d'un texte de loi relatif à une interdiction [19] de l'utilisation de la céruse dans la peinture (loi du 20 juillet 1909 entrant en application le 1er janvier 1915). Les données collectées dans l'enquête Plomb-Habitat permettant ce raffinement, les tables 46 et 47 ont été fournies au Comité Technique Plomb. Ces tables ont confirmé la décroissance concernant la présence de la céruse, au fil des différentes réglementations sur l'usage de la céruse dans la peinture.

TABLE 46 – Prévalence (%) de logements possédant un nombre d'unités de diagnostic (UD) à charge en plomb ≥ 1 mg/cm² selon la période de construction tenant compte de l'année 1915.

	≥1UD	**IC$_{95\%}$**	**≥2UD**	**IC$_{95\%}$**	**≥3UD**	**IC$_{95\%}$**	**≥4UD**	**IC$_{95\%}$**	**≥5UD**	**IC$_{95\%}$**	**≥10UD**	**IC$_{95\%}$**
Avant 1915	58,6	41,5-75,8	43,6	26,6-60,7	40,1	23-57,2	27,6	12-43,1	21,7	8,5-35	9,2	0-21,6
De 1915 à 1948	42	25,7-58,3	26,4	10,8-42	23,7	8,8-38,5	18	5,8-30,1	11,4	1,1-21,6	4,1	0-8,7
De 1949 à 1974	22,1	8,5-35,8	9,2	0,7-17,8	8,7	0,1-17,3	8,5	0-17	8,2	0-16,7	0,4	0-1,1
De 1975 à 1993	1,8	0-4,6	0	-	0	-	0	-	0	-	0	-
À partir de 1994	0,1	0-0,3	0	-	0	-	0	-	0	-	0	-

TABLE 47 – Prévalence (%) de logements possédant un nombre d'unités de diagnostic (UD) à charge en plomb ≥ 2 mg/cm² selon la période de construction tenant compte de l'année 1915.

	≥1UD	**IC$_{95\%}$**	**≥2UD**	**IC$_{95\%}$**	**≥3UD**	**IC$_{95\%}$**	**≥4UD**	**IC$_{95\%}$**	**≥5UD**	**IC$_{95\%}$**	**≥10UD**	**IC$_{95\%}$**
Avant 1915	47,4	29,6-65,2	24,5	12,3-36,7	22,4	10,4-34,4	18,3	7,8-28,8	12,3	4,2-20,4	2,7	0-7
De 1915 à 1948	28,6	12,4-44,7	21,8	8,3-35,2	19,2	6,5-31,9	15,6	3,8-27,4	10,8	0,6-21	3,7	0-8,2
De 1949 à 1974	13,2	0,1-26,3	3,7	0-8,1	3,3	0-7,6	1,8	0-4,7	1,8	0-4,7	0	-
De 1975 à 1993	1,2	0-3,7	0	-	0	-	0	-	0	-	0	-
À partir de 1994	0,1	0-0,3	0	-	0	-	0	-	0	-	0	-

La discussion précédente concernait les logements construits entre 1949 et 1974. Autant la date de 1949 est importante et son utilisation justifiée par rapport à la réglementation, autant la date de 1974 n'est ici qu'une contrainte imposée par la nomenclature INSEE. Il serait inexacte de déduire des résultats que, à partir de 1974 précisément il est possible de voir diminuer la présence de peintures à la céruse. L'analyse n'a pas été faite par année de construction mais par période de construction ; tout au plus il est possible d'affirmer que jusqu'au début des années 1970 il perdure des logements avec de la peinture à la céruse.

L'analyse des revêtements ne s'est pas faite en termes de charge surfacique en plomb, c'est-à-dire en considérant cette quantité comme continue, à l'instar de ce

19. L'emploi de la céruse, de l'huile de lin plombifère et de tout produit renfermant de la céruse, est interdit dans tous les travaux de peinture, de quelque nature qu'ils soient, exécutés par les ouvriers peintres, tant à l'intérieur qu'à l'extérieur des bâtiments [Lucas, 2011].

qui a été effectué pour la poussière par exemple. La raison est le protocole même, basé sur celui du CREP, permettant de relever ces charges en plomb surfacique. Pour procéder à une analyse d'une variable XRF continue, il aurait été nécessaire de connaître au moins le nombre d'UD par pièce et donc par logement et idéalement la surface de chacune des UD. Connaître le nombre d'UD par pièce aurait permis de construire une moyenne par pièce en sommant les mesures XRF et en les divisant par le nombre d'UD. Pondérer cette somme par la surface de chaque UD aurait pu permettre d'évaluert plus précisément la charge en plomb surfacique dans une pièce. Le nombre d'UD par pièce n'a cependant pas été relevé par l'enquêteur : si certaine UD étaient jugées ne pas avoir un revêtement plombée e.g. pour un revêtement neuf, l'enquêteur avait la possibilité de ne pas procéder à une mesure XRF. La surface de chaque UD devait être mesurée d'après le protocole de l'enquête Plomb-Habitat. Mais faire ces mesures est extrêmement chronophage et *in fine*, très peu de surfaces ont été relevées. Ceci montre que ce genre de quantité semble difficile à collecter dans une enquête à grande échelle. Concernant l'information donnée par les mesures XRF on pourra se reporter au complément de discussion fait en page 188.

2.4 Niveau en plomb de l'aire de jeu extérieure

Les niveaux en plomb qui ont été décrits (section 2.4 du chapitre 2) sont les niveaux en plomb des aires de jeu extérieures. Il ne s'agissait donc pas des sols extérieurs au sens large. Les aires de jeu ont été échantillonnées dans la mesure où elles doivent mieux refléter l'exposition de l'enfant, plutôt qu'un sol quelconque prélevé dans l'environnement du logement.

Les niveau en plomb des aires de jeu ont été décrits de la même manière que les autres compartiments environnementaux, c'est-à-dire en associant le poids de sondage du logement à l'aire de jeu de l'enfant habitant le logement. Ceci est problématique pour les aires de jeu non reliées aux logements, comme par exemple une aire de jeu dans un square. Mais la majorité des aires de jeu étaient constituées de places liées au logement ou très proches, comme le jardin ou la cour.

La comparaison des niveaux en plomb des sols est sujette aux types de sols mesurés ainsi qu'à la méthode de prélèvement. En France ce sont les surfaces labourées (appelés horizons) qui semblent être les plus documentées en termes de résultats [Mench & Baize, 2004]. Il est donc difficile de réaliser des comparaisons avec des études antérieures françaises. Outre les horizons, les concentrations les plus documentées doivent vraisemblablement être celles mesurées auprès de sites à risque tels que le site de Métaleurop Nord [20].

Aux États-Unis Jacobs *et al.* [Jacobs et al., 2002] indiquaient qu'en 2001-2002, 5 % des logements étaient estimés avoir un niveau en plomb dans l'aire de jeu extérieure supérieur ou égal à la valeur réglementaire américaine de 400 mg/kg. Les échantillons de sols avaient été prélevés dans cette étude par un carottage de 0,5

20. On peut prendre connaissance de résultats pour différents types de sol concernant ce site à partir de [Douay et al., 2009, Table 3].

« inch » soit environ 1,27 cm. Dans Plomb-Habitat le prélèvement a été réalisé avec un carottage d'une profondeur de 2 cm ; en France 1,4 % des aires de jeux en 2008-2009 étaient concernées par un dépassement de la valeur réglementaire américaine.

Concernant les poussières extérieures sur l'aire de jeu, aucune valeur réglementaire n'existe en France comme aux Etats-Unis. En France une limite d'exposition consensuelle égale à 300 $\mu g/m^2$ est utilisée [Bretin, 2006] comme pour les poussières intérieures. Approximativement 6,4 % ($IC_{95\%}$ = 0-15) des sols durs d'aires de jeu extérieures dépassent ce seuil à partir des données de Plomb-Habitat. Les charges extérieures en plomb dans la poussière sont approximativement 3,2 fois plus élevées que les charges les plus élevées à l'intérieur des logements.

2.5 Méthodologie statistique

À propos du plan de sondage

Pour produire les estimations décrivant les niveaux en plomb dans les compartiments environnementaux en milieu résidentiel, dans le logiciel utilisé (R package « survey ») le plan de sondage déclaré n'a pas été un plan totalement identique à celui de l'enquête Plomb-Habitat[21].
L'approximation s'est faite par un plan à degrés. La première raison est que le plan de sondage de Plomb-Habitat est un plan très complexe, difficile à manipuler théoriquement comme techniquement. Il n'est pas évident que ce genre de plan puisse être traité aisément pas certains logiciels de traitement de données d'enquête.
D'autre part la manipulation d'un tel plan dans le logiciel nécessite la détention de toutes les variables du plan de sondage. Or, toutes ces variables n'ont pas été initialement fournies pour la réalisation de l'analyse descriptive. Par exemple la stratification par niveau de plombémie utilisée au niveau des logements (cf. figure 15 page 68) n'a pas été initialement fournie.

Dès lors il aurait été naturel d'analyser le plan comme un plan à 2 degrés : un premier degré considérant les hôpitaux et un second les logements directement. Dans la fonction `svydesign` du package « survey » de R, dans laquelle le plan est déclaré, pour qu'un plan à plusieurs degrés soit considéré comme tel, il est nécessaire d'indiquer ce qui est appelé la correction de population finie (ou *fpc* en anglais) à chaque degré. Cette quantité permet de réduire les variances lorsqu'une grande fraction de la population considérée est échantillonnée. La *fpc* concernant les hôpitaux n'a pas été fournie et la *fpc* du second degré demeure inconnue. Le logiciel a donc traité le plan comme un plan à un seul degré. L'approximation à un degré des plans à plusieurs degrés est inévitable généralement lorsque l'on s'intéresse à des bases de données publiques car les informations du plan sont généralement que très partiellement fournies. De même cette approximation sera utilisée par les logiciels ne manipulant pas les plans à plusieurs degrés [Lumley, 2010a]. L'approximation par un seul degré donne les mêmes estimations des paramètres d'intérêt que celles

21. Lorsque l'on s'arrête aux logements, le plan de sondage est un plan à 2 phases dont la première phase est un plan à 2 degrés.

d'une analyse à plusieurs degrés car les poids de sondage utilisés sont les mêmes. La prise en compte des degrés permet de diminuer les variances mais il peut être acceptable dans un cadre général de perdre de la précision sur la variance pour gagner en simplicité de manipulation du plan [Lumley, 2010a].

Analyse sur cas complets

Pour chaque caractère étudié (concentration en plomb dans l'eau, charge moyenne en plomb à l'intérieur d'un logement, *etc.*), les logements qui n'avaient pas de valeur relative ont été supprimés de l'analyse. Cela explique pourquoi dans les tables de la section 2 du chapitre 2 indiquent le cas échéant un $n < 484$ et donc un $N < 3\,581\,991$ puisque les poids de sondage des logements écartés ne sont pas pris en compte dans la somme calculant N. En l'état, les analyses supposent donc que les données manquantes sont *MCAR* : l'hypothèse faite est qu'il n'y pas de différence de distribution entre les données manquantes et les données observées, ce qui est peu probable en réalité (cf. section 7.2 de la partie « Spécificités des données d'enquête »). Dans cette situation les estimations peuvent être biaisées ; pourtant les analyses sur cas complets se trouvent couramment dans les publications scientifiques. Des auteurs militent pour que dorénavant cette pratique cesse dans la mesure où les méthodes de traitement de la non-réponse sont largement implémentées dans les logiciels aujourd'hui [Sterne et al., 2009]. Cependant l'application aux données d'enquête des méthodes de traitement de la non-réponse utilisées en statistiques classiques n'est pas directe comme exposé en section 7.4 de la partie « Spécificités des données d'enquête ». Le traitement des données manquantes dans le cadre d'enquête est encore du domaine de la recherche. La problématique de la particularité des données d'enquêtes et de l'adaptation des outils de statistiques classiques est traitée dans la section suivante.
Au stade des estimations des niveaux en plomb dans les différents compartiments environnementaux résidentiels, la limite de connaissances sur le traitement des données manquantes, en particulier en ce qui concerne des données d'enquête, a été une des raisons du non traitement les données manquantes. De plus le calendrier lié au rendu contractuel a été une autre raison.
Dans le cadre d'un travail de description similaire, à l'avenir il serait préférable de mettre en œuvre un traitement des données manquantes comme celui réalisé au chapitre 3 par exemple, en tenant compte des résultats des travaux de recherche du moment obtenus dans le domaine du traitement des données manquantes en sondage.

Méthodes de statistiques classiques et données d'enquête

La méthodologie statistique utilisée pour décrire les niveaux en plomb dans les compartiments environnementaux en milieu résidentiel, repose sur les outils de la théorie de l'échantillonnage et de l'estimation en population finie. Par rapport à la statistique classique, les données d'enquête sont une contrainte à prendre en compte au moment de l'analyse. C'est une contrainte dans le sens où, sans même parler du développement théorique relatif, les méthodes de statistiques classiques ne sont techniquement pas toutes disponibles dans les logiciels pour l'analyse des données

d'enquêtes.

Par exemple, les méthodes de traitement de données manquantes telles que l'imputation multiple ne semblent pas applicables aux données d'enquête sans adaptation à ce type de données. L'information du plan doit être *a minima* prise en compte dans le modèle d'imputation mais les estimateurs peuvent tout de même avoir de mauvaises propriétés (sur ce sujet cf. section 7.4 de la partie « Spécificités des données d'enquête »).

Un autre exemple concerne les données censurées. Lors des estimations des niveaux en plomb dans chaque compartiment environnemental, les données censurées (à gauche) ont été traitées par ce qui est appelée substitution : une même valeur égale à LQ/2 a remplacé les valeurs $<$ LQ. Bien que largement employée dans les articles scientifiques (on voit aussi $\mathrm{LQ}/\sqrt{2}$ comme valeur de substitution), la méthode par substitution semble être à proscrire. Dennis R. Helsel écrit en fin du premier chapitre de son livre [Helsel, 2012], *« la substitution de valeurs assujettie à un seuil, étant encore aujourd'hui la méthode la plus communément utilisée, N'est PAS une méthode raisonnable pour interpréter les données censurées »*.
Cependant les données censurées n'ont pas été traitées car le logiciel R et son package « survey » ne peuvent traiter ce type de données. Ainsi, quand bien même on est conscient du problème sous-jacent imposé par les données censurés, si techniquement (voire même théoriquement) rien n'est implémenté dans les logiciels, alors ce qui est uniquement réalisable à travers les logiciels disponibles est utilisé en dernier recourt pour réaliser les analyses. À notre connaissance, ni le package « NADA » [22] développé par Dennis R. Helsel, ni le package « survey » développé par T. Lumley ne peuvent traiter des données censurées de données d'enquête. Il peut donc être intéressant de contribuer à ce domaine de recherche - données censurées en sondage - et d'implémenter dans les logiciels les résultats obtenus dans la littérature le cas échéant.

Intervalles de confiance

Il est fort dérangeant d'obtenir à partir des méthodes algébriques de calcul de variance, décrites en section 2.2 et tout au long de la section 3 de la partie « Spécificités des données d'enquête », des intervalles de confiance avec une borne inférieure négative pour une quantité estimée ne pouvant qu'être positive ou nulle. C'est la raison pour laquelle, par exemple, l'intervalle de confiance de la moyenne arithmétique n'a pas été affiché dans les tables 17 et 18 de la section 2.2 au chapitre 2. Néanmoins, lorsque cette borne inférieure était estimée négativement mais très proche de zéro, zéro a été affiché dans l'intervalle de confiance.

À l'avenir pour un travail descriptif similaire, il serait judicieux d'appliquer les techniques pouvant remédier à ce problème de mauvais intervalles de confiance. Les techniques pouvant remédier au problème sont les méthodes par répliques en particulier le Bootstrap. Mais il est nécessaire d'adapter cette technique aux données

22. http://www.practicalstats.com/

d'enquêtes, car les estimateurs Bootstrap classiques peuvent ne pas être sans biais [23]. Plusieurs méthodes de ré-échantillonnage sans remise e.g. *« Rao-Wu rescaling Bootstrap »* ou avec remise e.g. *« Rao-Wu rescaling Bootstrap à-la-Chipperfield-Preston »* ont donc été développées [Girard, 2012]. Cependant il est nécessaire de dériver soi-même l'expression des poids qui conviennent au plan de sondage ; la mise en oeuvre de ces méthodes n'est pas immédiate malgré le fait qu'elles utilisent principalement les capacités informatiques.

Dans cette section 2.5 les limites de l'analyse descriptive des niveaux en plomb ont été exposées. À l'avenir, ces limites doivent être écartées dans le mesure du possible, c'est-à-dire dans les limites théoriques et techniques (logiciel) imposées par le type de données manipulées - des données d'enquête.

3 Estimation de la contribution des sources en plomb à contaminer la poussière intérieure déposée au sol

3.1 Choix du type de modélisation

Le choix du type de modélisation a été en partie lié à l'intérêt porté à la corrélation entre les charges en plomb à l'intérieur d'un logement. Cette notion de corrélation n'a pas de sens en analyse de données d'enquête en approche plan puisque les données ne sont pas aléatoires. Néanmoins, comme indiqué par Thomas Lumley [24] lors d'un échange par courrier électronique, il est possible d'obtenir facilement les estimateurs de la variance inter-groupe (i.e. entre les logements) et intra-groupe (i.e. à l'intérieur des logements) et ainsi d'obtenir une estimation de la corrélation recherchée sans faire une analyse en approche modèle spécifiquement. Pour calculer cette corrélation il faut estimer la variance des charges en plomb des pièces dans la population, ainsi que la variance des moyennes de ces charges par logement dans la population [25]. Mais les charges en plomb devant être ajustées par l'introduction de variables de confusion, T. Lumley a confirmé alors que l'analyse devait se faire dans une approche modèle, dont des bases théoriques existaient pour la modélisation multi-niveaux. Bien que peu au courant des implémentations de ce type d'analyses dans les logiciels, T. Lumley a conseillé de se renseigner sur la dernière version (V12) du logiciel Stata [26]. D'autres logiciels sont capables de traiter ce genre de modélisation sur données d'enquêtes, bien qu'ils ne soient pas tous équivalents en termes de performances de calcul et surtout de caractéristiques techniques (type de variable

23. Par exemple l'estimateur de variance du Bootstrap d'Efron n'est pas sans biais dans le cas de la moyenne sous un échantillonnage aléatoire simple avec remise ;

24. *Professor of Biostatistics, University of Auckland.* Auteur de *« Complex Surveys »* [*Lumley, 2010a*].

25. Ceci peut se faire facilement à partir de la fonction « svyvar » du package « survey » [Lumley, 2004, Lumley, 2010b, Lumley, 2010a] du logiciel R.

26. StataCorp. 2011. Stata Statistical Software : Release 12. College Station, TX : StataCorp LP.

réponse, traitement des poids *etc.*) [Chantala & Suchindran, 2006].

Lors de la présentation des résultats de la section 3 du chapitre 3 au 7e colloque francophone sur les sondages [27], Louis-Paul Rivest [28] a indiqué qu'il pourrait être utile de réaliser l'analyse à partir d'un modèle de régression linéaire multiple (un seul niveau donc), étant donné que cette modélisation est relativement robuste face à la non vérification de l'hypothèse d'indépendance des observations (pièces). Cette modélisation n'avait pas été envisagée dans la mesure où la corrélation intra-classe, discutée ci-avant avec T. Lumley, ne pouvait pas être calculée. À titre indicatif et de comparaison, les résultats du modèle à 2 niveaux sans pondération, à un niveau avec pondération [29], et à un niveau sans pondération sont affichés en table 48, sur cas complets et pour le plomb acido-soluble. Les résultats pour les modèles à un niveau ont été obtenus sous Stata V12 à partir d'une approximation à un seul degré du plan de sondage (sur ce sujet cf. section 2.5 de cette partie « Discussion »). On peut constater qu'il est plus facile d'obtenir des effets avec une mise en évidence très importante à partir du modèle à un niveau (sans pondération) qu'à partir de notre modèle à 2 niveaux. En revanche, peu d'effets ont une une mise en évidence importante ou très importante à partir de l'estimation du modèle à un niveau avec pondération. Bien que les coefficients estimés à partir des 2 modèles à un niveau ne soient pas absurdes, la seule figure 42 (relative au modèle à un niveau (sans pondération)), suffit à montrer qu'une modélisation à 1 niveau ne nous est pas utile. La même figure pour le modèle avec pondération (non affichée) est similaire et indique donc un ajustement du modèle aux données tout aussi médiocre. La régression linéaire multiple n'était donc pas une modélisation adaptée pour être appliquée aux données utilisées dans le présent travail.

3.2 Sélection des covariables et choix de la forme de la relation

Les variables en lien avec la plombémie dans la littérature ont été sélectionnées pour figurer dans le modèle. L'hypothèse relative qui a été faite est que puisque la poussière contaminée par le plomb est le contributeur majeur à la plombémie il est vraisemblable que les variables reliées à la plombémie soient aussi reliées au niveau en plomb dans la poussière intérieure. Certaines de ces variables n'avaient été étudiées dans la littérature qu'en relation avec la plombémie et donc pas avec le plomb des poussières.

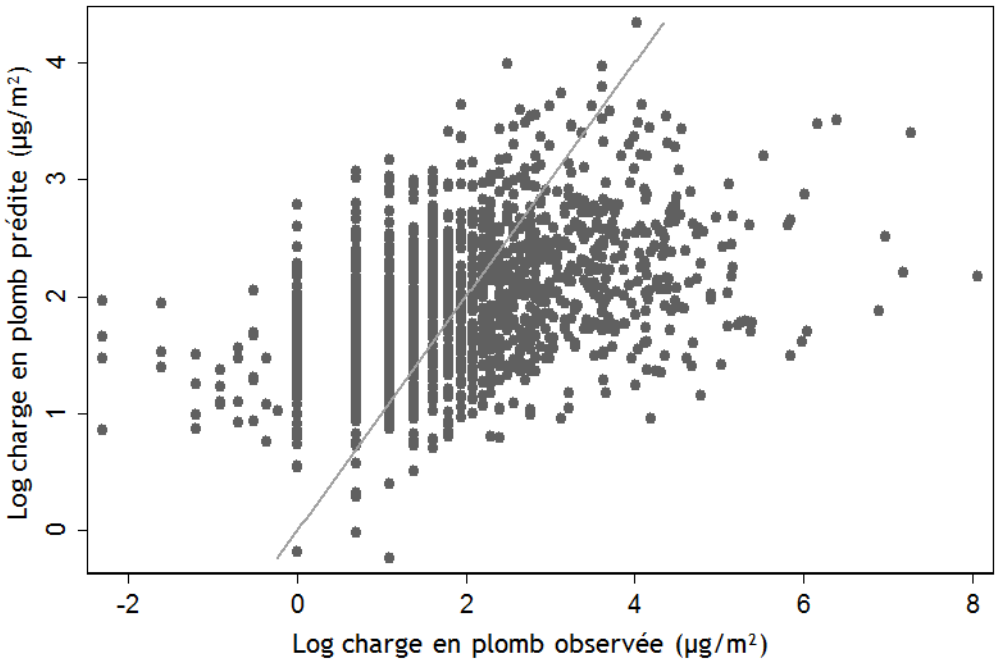
Certaines variables de confusion planifiées pour être introduites dans le modèle ont été écartées du modèle finalement. Puisque que l'on s'intéresse à l'effet de certaines sources en plomb, toute variable susceptible de « manger » une partie de ces effets ne doit pas être utilisée dans le modèle. Par exemple, la période de construction est une information qui revient régulièrement à travers la littérature comme

27. Lucas J.-P., 2012. *Modélisation multi-niveaux de données d'enquête : impact sur les estimations dû aux poids de sondage de niveau 2 introduits dans la pseudo-vraisemblance*, Ensai, Bruz.
28. Professeur, Université Laval, Québec, Canada.
29. Les poids utilisés pour les pièces sont les poids w_i - voir page 121.

TABLE 48 – Comparaison entre estimation par 2 niveaux et par 1 seul niveau.

		2 niv. sans pond.		**1 niv. sans pond.**		**1 niv. avec pond.**	
Covariables	Modalités	Coef. Estimé	p-value	Coef. Estimé	p-value	Coef. Estimé	p-value
« intercept »		0,968	0,007	0,986	0,001	1,022	0,011
Emplacement du lgt	Semi-enterré	réf.		réf.		réf.	
	RDC	0,029	0,912	0,106	0,488	-0,047	0,870
	En étage	0,327	0,284	0,405	0,021	0,097	0,762
Saison	Automne/hiver	réf.		réf.		réf.	
	Printemps/été	0,340	0,002	0,340	0,000	0,417	0,011
Lavage humide du palier	Pas de palier	réf.		réf.		réf.	
	Oui	-1,533	0,000	-1,447	0,000	-1,072	0,008
	Non	-1,684	0,000	-1,641	0,000	-1,249	0,029
Type de pièce	Chambre	-0,517	0,000	-0,460	0,000	-0,316	0,037
	Entrée	réf.		réf.		réf.	
	Salon	-0,273	0,000	-0,255	0,021	-0,147	0,304
	Cuisine	-0,039	0,450	-0,012	0,908	0,139	0,206
	Salle de jeu	-0,350	0,001	-0,521	0,001	-0,303	0,158
Fréq. Lavage humide-pièce		0,074	0,317	0,143	0,013	0,100	0,451
Fréq. Lavage sec-pièce		-0,001	0,986	-0,022	0,674	-0,078	0,497
Endroit du prélèvement	Endroit de jeu préféré	réf.		réf.		réf.	
	Centre de la pièce	-0,016	0,737	0,044	0,517	0,062	0,493
Log(Nombre d'activités à risque+1)		0,084	0,381	0,124	0,051	-0,087	0,446
Log(XRF garde-corps+1)		0,413	0,000	0,401	0,000	0,402	0,002
Log(Charge en Pb-poussière ext.+1)	Ne joue pas à l'ext.	réf.		réf.		réf.	
	Y joue souvent	0,121	0,040	0,115	0,001	0,121	0,066
	Y joue tout le tps	0,138	0,046	0,107	0,002	0,044	0,483
Log(Concentration en Pb-sol ext.+1)	Ne joue pas à l'ext.	réf.		réf.		réf.	
	Y joue souvent	0,112	0,002	0,114	0,000	0,107	0,017
	Y joue tout le tps	0,117	0,005	0,108	0,000	0,143	0,002
Log(Charge en Pb-palier+1)		0,317	0,000	0,286	0,000	0,227	0,048
Log(Trafic routier)		0,021	0,419	0,014	0,397	-0,008	0,773
Démolition	Non	réf.		réf.		réf.	
	Oui	0,317	0,010	0,289	0,001	0,323	0,083
Log(Fréquence de loisir+1)		0,042	0,743	0,054	0,421	0,205	0,082
Travaux extérieurs	Non	réf.		réf.		réf.	
	Oui	-0,090	0,549	-0,086	0,369	0,046	0,774
Travaux intérieurs	Non	réf.		réf.		réf.	
	Oui	0,123	0,267	0,091	0,168	0,316	0,062
Log(Basias+1)		0,090	0,016	0,075	0,002	0,077	0,072
Log(Basol+1)		-0,095	0,678	-0,003	0,980	0,040	0,829
Log(Bdrep+1)		0,202	0,178	0,229	0,040	0,453	0,105
Log(Tabagisme journalier+1)		0,284	0,002	0,253	0,000	0,213	0,098
Log(Somme XRF-détériorié+1)		0,157	0,000	0,330	0,000	0,375	0,001
Log(Somme XRF-état d'usage+1)		0,088	0,033	0,233	0,000	0,140	0,039

FIGURE 42 – y prédit versus y observés obtenu avec une modélisation linéaire multiple sans pondération. La droite est d'équation $y = x$.



étant un facteur de risque de plombémie [Dixon et al., 2009, par exemple] ou de contamination des poussières [Gaitens et al., 2009, par exemple]. Dès lors, c'est trop mécaniquement que cette information a été introduite dans le modèle alors qu'elle aurait accaparé une partie de l'effet des variables relatives à la charge en plomb des peintures sous l'hypothèse que période de construction et niveau en plomb des peintures sont liées. Il en a été de même pour la variable indiquant l'environnement extérieur (urbain/rural) susceptible d'accaparer une partie de l'effet des sources extérieures.

Puisqu'il s'agissait d'évaluer la contribution de toutes les sources possibles à contaminer la poussière intérieure, toutes les variables sources disponibles dans la base de données ont été introduites. L'application d'une quelconque procédure automatique itérative de sélection de variable rencontrée fréquemment dans les analyses n'était donc pas justifiée. Néanmoins certaines règles de construction de modèle existent quant au nombre maximal de variables pouvant être introduite dans un modèle. La sur-paramétrisation c'est-à-dire l'introduction d'un trop grand nombre de variables et donc de paramètres à estimer, conduit généralement à un sur-ajustement [Harrell, 2001, section 4.4]. Autrement dit le modèle estimé s'ajuste bien aux données sur lesquelles il s'est basé, mais cela ne sera pas le cas pour de nouvelles observations ajoutées (lignes dans la table de données).
Au total, 30 coefficients de régression (sources en plomb et variables de confusion) ont été estimés. Le nombre de paramètres à estimer pour qu'un modèle de régression soit fiables semble être compris entre $m/20$ et $m/10$ où m est la taille de l'échantillon [Harrell, 2001, section 4.4]. Ces recommandations semblent basées sur des

modèles à un seul niveau. Dès lors si l'on considère la valeur la plus restrictive pour m i.e. $m = 484$ (et non $m = 1834$ pièces), notre modèle ne devrait pas avoir plus de 48 ou de 24 régresseurs. Notre nombre de coefficients de régression à estimer, 30, se situait bien dans ces limites.

La seule réduction de dimensionnalité concernant les sources qui a été effectuée est liée à la présence de collinéarité à travers certaines covariables. Ceci a été réalisé pour les variables « Basias », « Basol » et « Bdrep » construites (cf. annexe 3). Concernant les variables de confusion une réduction du nombre de paramètres à estimer a été faite via le critère BIC. Le choix du critère BIC plutôt que l'AIC (*Akaike information criterion*) a été complètement arbitraire. Cependant il semble que le critère BIC soit plus conservateur dans le sens où il met moins facilement en évidence des effets [Raftery, 1998]. Le BIC suppose qu'il existe un « vrai » modèle de dimension fini. L'AIC semble plus naturel car il ne suppose pas que « le » modèle existe, sachant que l'on peut ajuster des modèles plus complexes de manière fiable quand la taille de l'échantillon augmente. Les 2 critères ont été développés pour comparer au plus 2 modèles (pas forcément emboîtés). S'ils sont utilisés comme une sélection pas à pas de variables ils conduisent à un sur-ajustement du modèle sauf si un ordre des variables à ajouter ou à enlever était pré-spécifié comme cela a été fait en section 2.3 du chapitre 3 ; la sélection de modèle pilotée par les données conduisant à des estimations biaisées des critères tels que l'AIC [Chatfield, 1995, section 4]. Il était espéré que les modèles intermédiaires seraient moins bons que le modèle sans les 3 variables de confusion considérées afin d'avoir un modèle le plus parcimonieux en termes de nombre de paramètres à estimer.

L'information a été introduite via certaines variables construites de manières particulières. La construction des variables « Basias », « Basol » et « Bdrep » s'est faite sous certaines hypothèses (cf. section 2.1 du chapitre 3). D'autres éléments concernant ces variables n'ont pas été pris en compte. En particulier l'impact des vents et notamment leurs directions qui jouent sur la dissémination vers un logement donné, d'un polluant émis d'un site ou présent sur un site. Bien qu'il faille être conscient de ce problème réel, il était impossible de prendre en considération cet aspect à travers les données disponibles. D'autres part la prise en compte de ce problème demande un traitement de modélisation du déplacement des masses d'air du ressort des compétences de la météorologie et de la géostatistique. Ceci pourrait donc être une explication du résultat non attendu pour la variable « Basol ». Outre cet aspect, il est ressorti dans le cadre de discussions avec différents partenaires du COPIL comme externes au COPIL, que l'information fournie par ces bases de données ne seraient pas idéales. Cela pourrait de plus expliquer le résultat non attendu relatif à la variable « Basol ».
L'information relative au tabagisme a été introduite via une covariable continue alors qu'elle figurait dans le questionnaire (de l'InVS) comme variable catégorielle. La raison était de limiter le nombre de paramètres à estimer (coefficients) : au lieu de 5, un seul était à estimer. Le caractère continue de l'information, la durée où l'on fume à l'intérieur, est bien une quantité continue. C'est uniquement l'utilisation dans le modèle de cette information sous forme continue qui est sujet à interrogation. Afin de s'assurer de la robustesse du modèle vis-à-vis d'une utilisation continue ou

catégorielle, le modèle a été estimé avec la variable tabagisme sous forme catégorielle ordinale et ses estimations comparées à celles du modèle utilisant la variable sous forme continue. Les résultats sont affichés en table 49. Bien qu'un gradient n'apparaisse pas en fonction des catégories ordonnées de la variable tabagisme, il semble néanmoins qu'un effet puisse être constaté entre la plus faible durée passée à fumer à l'intérieur et les plus longues durées ; tout au moins le résultat ne semble pas contredire l'effet estimé pour la covariable continue. Les contributions de la variable « Tabagisme journalier » sont indiquées par la figure 43. Les estimations (non montrées) des autres paramètres sont pratiquement inchangées. Le résultat de cette analyse de sensibilité ne semble donc pas remettre en cause l'effet du tabagisme intérieur et estimé initialement dans le modèle en utilisant la variable « Tabagisme journalier » en continue.

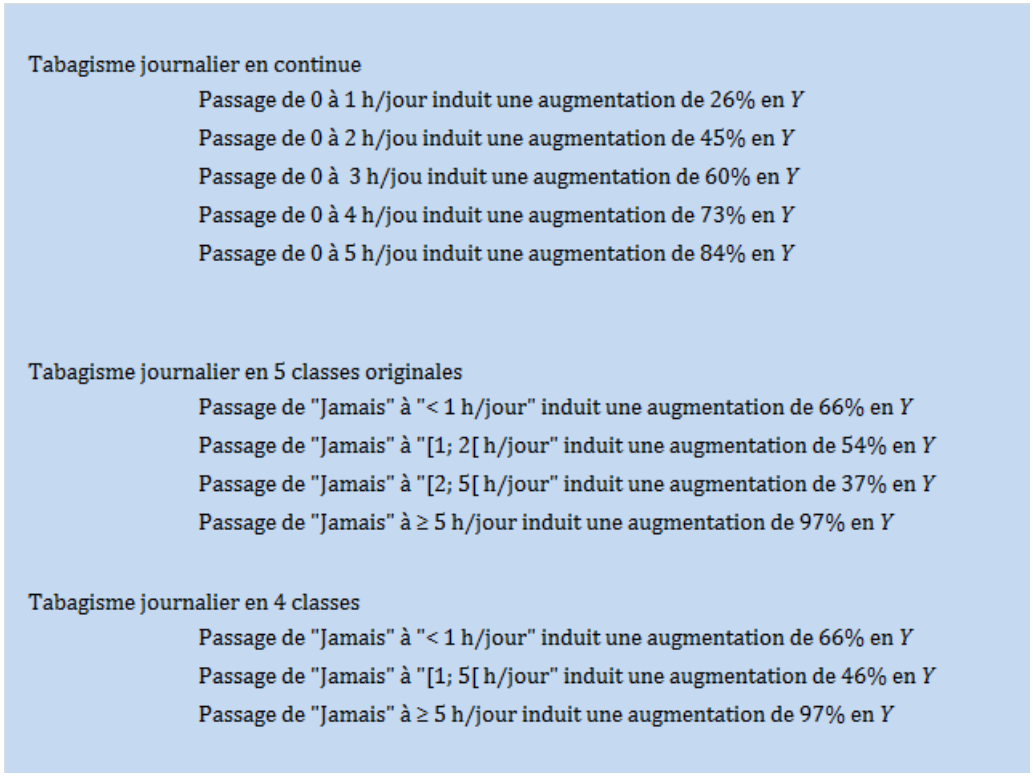
TABLE 49 – Analyse de sensibilité réalisée sur la variable Tabagisme journalier (Pb total - cas complets).

Covariable	**Modalités**	**Estimation**	***p***	**IC 95 %**
Covariable utilisée :				
Log(Tabagisme journalier+1)		0,339	0,000	(0,157 ; 0,521)
Covariable catégorielle d'origine :				
Tabagisme journalier	Jamais ou presque jamais	réf.		
	$<$ 1 h/jour	0,507	0,004	(0,164 ; 0,850)
	[1 ; 2[h/jour	0,435	0,130	(-0,128 ; 0,998)
	[2 ; 5[h/jour	0,313	0,114	(-0,075 ; 0,701)
	$\geq$ 5 h/jour	0,676	0,004	(0,216 ; 1,136)
Covariable catégorielle d'origine avec regroupement de modalités :				
Tabagisme journalier	Jamais ou presque jamais	réf.		
	$<$ 1 h/jour	0,507	0,004	(0,164 ; 0,850)
	[1 ; 5[h/jour	0,378	0,031	(0,034 ; 0,722)
	$\geq$ 5 h/jour	0,676	0,004	(0,216 ; 1,137)

Lorsque l'enfant jouait à l'extérieur sur un sol dur, par exemple une cour d'immeuble, la valeur zéro a été associée à la covariable quantifiant le plomb en sol meuble. Lorsque l'enfant jouait à l'extérieur sur un sol meuble, par exemple une pelouse de jardin, la valeur zéro a été associée à la covariable quantifiant le plomb en sol dur. Ceci implique que le phénomène de *« track-in*[30] *»* estimé à travers ces deux covariables du modèle, est considéré comme nul ou négligeable *a priori* à partir d'un des 2 types de sol. C'est une hypothèse forte mais surtout vraisemblablement fausse pour la majorité des ménages. Pour éviter de procéder ainsi, les 2 types de sols auraient du être prélevé et une valeur égale à zéro aurait du être utilisée uniquement pour décrire les cas où l'enfant ne jouait pas, de manière certaine, sur un

30. Fait d'introduire via ses vêtements et chaussures des éléments extérieurs (sols, poussières, *etc.*) à l'intérieur du logement.

FIGURE 43 – Contributions selon la variable Tabagisme journalier utilisée (Pb total - cas complets).



type de sol particulier. Cependant même cette situation n'est pas réaliste afin de quantifier le phénomène de *« track-in »*. Le tractage de sol ou de poussière contaminés à l'intérieur du logement est le lot commun de tous les occupants du ménage. Dès lors outre des prélèvements de sols extérieurs à réaliser, il s'agit de collecter le budget-espace-temps de chacun des occupants afin d'étudier finement le *« track-in »*. Sans même parler des différents types de chaussures et semelles (adhérence), les conditions d'humidité extérieure, et sans doute d'autres paramètres importants auxquels il faudrait réfléchir, il s'agit là d'une étude qui sort du cadre du présent travail de recherche. Ainsi, les choix pour ces covariables relatives aux sols extérieurs a été fait à partir des informations disponibles en base de données et en essayant de faire en sorte que ces choix soient pris en compte dans la portée de l'interprétation.

Les 2 variables XRF, quantifiant la charge en plomb surfacique des revêtements des pièces investiguées, selon le type de dégradation de ces revêtements intérieurs, ne sont pas des variables communément utilisées à travers les études publiées. La valeur maximale XRF du logement ou de la pièce est quasiment toujours utilisée. Cette valeur maximale ne peut pas quantifier l'information nécessaire à introduire dans un modèle estimant l'effet des peintures au plomb sur la contamination des poussières ou même sur la plombémie. En effet pour ce faire il est nécessaire de considérer dans l'information XRF, à la fois la charge surfacique en plomb, l'état de dégradation et la surface du revêtement. Si on s'intéressait à expliquer la plombémie et non la charge en plomb dans la poussière, il serait utile d'ajouter une notion d'accessibilité à l'enfant des revêtements. Les 2 variables utilisées dans le modèle s'approchent

de cette considération. Cependant, bien que requise par le protocole de l'enquête Plomb-Habitat, la surface de chaque UD n'est pas disponible. Plus précisément elle n'a été relevée que dans de rares cas. Étant donné le fort pourcentage de données manquantes pour cette information, il était utopique d'essayer de traiter ces données manquantes afin de pouvoir utiliser cette information pour construire une variable XRF agrégée au niveau de chaque pièce. L'explication du non relevé de la surface de chaque UD est que ce relevé devient rapidement lourd et chronophage. Le relevé d'une telle information dans le cadre d'une enquête à grande échelle semble inadapté sans allonger la durée de chaque visite à domicile et donc sans accroître le budget de l'enquête.
Chacune des 2 variables XRF est donc construite à partir de la somme des mesures XRF de chaque UD d'une même pièce. Il n'était pas possible de faire une moyenne en divisant par le nombre d'UD de la pièce, à défaut de leur surface totale, car ce nombre n'était pas connu. Dès lors que l'enquêteur réalisant les mesures XRF jugeait que le revêtement d'une UD ne contenait pas de plomb, par exemple si le revêtement était neuf, il était autorisé à ne pas réaliser de mesure XRF. Ceci dans un souci de gain de temps de réalisation d'enquête. Ces UD non mesurées n'ayant pas été listées - ce n'était pas prévu dans le protocole[31] - le nombre d'UD par pièce n'était pas connu. Dès lors le fait de sommer les valeurs XRF de chaque UD était dans l'idée de faire apparaitre le nombre d'UD dans la variable construite, et par là même de se rapprocher de la surface contaminée, surface qui aurait été idéale. L'utilisation usuelle de la valeur maximale XRF pose alors la question de la sur-estimation de l'effet des peintures au plomb dans la contamination des poussières possiblement faite dans les études passées.

Il a été choisi sur la base de la littérature d'estimer une relation, entre la variable réponse et les covariables, ayant une forme $\text{Log}(Y) = \sum \beta \text{Log}(X)$ afin de répondre à l'objectif fixé, à savoir estimer la contribution des sources en plomb à contaminer la poussière intérieure. On a donc fait l'hypothèse que cette relation aura été utile pour faire des comparaisons entre les contributions des sources, et ainsi des interprétations pour donner aux pouvoir publics la possibilité de prendre des décisions relatives à la réduction des expositions liées aux poussières intérieures. D'autre part il est à noter que les 2 études [Jiang & Succop, 1996, Rust et al., 1997] sur lesquelles le choix de la relation a été basée, n'indiquent pas que cette forme de relation est la meilleure en soi pour modéliser des données environnementales relatives au plomb. Ces études ont seulement indiqué que, entre ce qui est appelée une modélisation *Log-Additive*[32] et une modélisation appelée Log-Log dans [Jiang & Succop, 1996] et *Log-Linear*[33] dans [Rust et al., 1997], le modèle *Log-Linear* expliquait une plus grande part de variabilité de la variable réponse. Il faut de plus souligner que la variable réponse de leurs modèles était la Log plombémie. On a donc fait l'hypothèse que leurs résultats étaient encore valables pour modéliser la Log charge en plomb des poussières. Enfin,

31. Le protocole de mesurage XRF de Plomb-Habitat était basé sur le protocole CREP. Ce dernier requiert que le technicien liste toutes les UD recouvertes d'un revêtement ou non (paragraphe 6 de l'annexe I de l'arrêté du 25 avril 2006 relatif au constat de risque d'exposition au plomb). La raison pour laquelle cette liste n'a pas été demandée dans l'enquête Plomb-Habitat m'est inconnue.
32. $\text{Log}(Y) = \text{Log}(\sum \beta X)$.
33. $\text{Log}(Y) = \sum \beta \text{Log}(X)$.

leurs modèles sont des modèles à un niveau et l'impact de l'introduction d'un niveau supplémentaire sur la validité de leur résultats serait à étudier.

3.3 Données censurées et imputation des données manquantes

Les données censurées (< LQ) des niveaux en plomb dans la poussière intérieure ont été substituées par la valeur LQ/2. L'impact négatif de la méthode de substitution pour traiter les données censurées a déjà été discutée en section 2.5 de cette partie « Discussion ».

Néanmoins vu le pourcentage de données censurées pour une variable, au plus de 10 % environ, l'utilisation de la substitution ne devait pas induire des estimateurs biaisés selon [Lubin et al., 2004]. Il faut cependant noter qu'il n'y avait que 1,7 % de données censurées pour le dosage en plomb acido-soluble des poussières intérieures mais que ce pourcentage était de 11,2 % pour le dosage en plomb total.

Il aurait été cependant possible de traiter ces données censurées à l'étape de l'imputation des données manquantes. En effet il est possible (dans Stata V12 utilisé pour l'imputation) de spécifier pour chaque variable avec données censurées un intervalle dans lequel les valeurs obtenues pour remplacer les données censurées seront issues. Ceci est appelée imputation multiple par la régression par intervalle (*interval regression imputation*). Cependant cela aurait d'une part considérablement alourdi le modèle d'imputation des données manquantes, déjà complexe, et cela aurait compliqué l'organisation du travail d'autre part. En effet il aurait fallu procéder à l'imputation avant de réaliser l'application numérique pour l'étude de l'impact des poids de niveau 2 (section 3 du chapitre3). De plus étant donné que les résultats du modèle sur données imputées ont été volontairement comparés à ceux du modèle sur cas complets, il aurait été nécessaire d'imputer, de manière simple, les seules données censurées pour disposer d'une table « cas complets » sans données censurées.

L'imputation multiple n'est pas conseillée par certains auteurs pour des données d'enquête alors que selon d'autres elle est encourageante (cf. section 7.4 de la partie « Spécificités des données d'enquête »). Puisqu'il n'y a pas encore de recommandations finales au sujet de l'imputation des données manquantes, le choix a été d'utiliser l'imputation multiple notamment à partir de la méthode ICE car cette dernière était très flexible en pratique.

3.4 Contributions des sources à contaminer la poussière et interprétation des résultats

Les résultats de la section 6.3 du chapitre 3 ont été obtenus à partir d'une modélisation à 2 niveaux, sans pondération ; 30 coefficients de régression ont été estimés dont 18 concernant 16 covariables sources.

Ces résultats ont montré qu'il n'était pas possible de mesurer l'effet et donc la contribution d'une source de manière très précise, les intervalles de confiance étant assez larges. Néanmoins les résultats peuvent être assez précis pour prendre des décisions quant à l'utilité d'agir sur une ou plusieurs sources afin de réduire l'exposition au plomb à travers l'abaissement des niveaux en plomb dans les poussières intérieures déposées au sol. On entend par « résultats assez précis » une gamme de valeurs pour la contribution des sources qui permet de prendre une décision allant dans le même sens quelque soit la contribution dans la gamme.
Le résultat de chacune des sources est discuté ci-après. Puisque le dosage en plomb acido-soluble est le dosage de la réglementation française, la discussion se focalisera sur les résultats concernant ce dosage ; les différences de résultats entre le plomb total et le plomb acido-soluble ne se situent de toute manière qu'à la marge.

Sources liées à l'activité des occupants du logement

Nombre d'activités professionnelles à risque pratiquées par les membres du foyer.
Ces sources n'ont été étudiées par le passée que dans le cas d'une relation avec la plombémie [Sanborn et al., 2002, Schapiro & Bretin, 2006].
Le nombre d'activités professionnelles a été évalué comme une source l'effet relatif n'a été que pauvrement mis en évidence. De plus il faut atteindre un seuil de 2 voire même 4 à 5 activités pratiquées à travers les membres d'un même foyer pour mettre en évidence une contribution (au maximum de 10 % mais pouvant aller de -20 % à 51 %) sur la charge en plomb des poussières intérieures. Au plus 5 % des ménages dans lesquelles un enfant âgé de 6 mois à 6 ans vit sont concernés par un nombre d'activités égal à 4 ou 5.

Nombre de loisirs à risque pratiqués par les membres du foyer à l'intérieur du logement.
La discussion autour de cette quantité est similaire à celle du nombre d'activités professionnelles à risque pratiquées par les membres du foyer.

Tabagisme à l'intérieur du logement.
Le plomb provenant de la fumée de tabac a été peu étudié dans la contamination par le plomb dans l'habitat. Il a été estimé que la charge en plomb des poussières intérieures augmentait de 40% dès lors que l'on fume environ 1,5 h/jour à l'intérieur du logement. La mise en évidence de cet effet était très fiable. Le résultat trouvé va dans le même sens que celui de Gaitens *et al.* [Gaitens et al., 2009]. L'information portée par notre covariable « Tabagisme » n'est cependant pas idéale comme discuté en section 3.2 de cette partie discussion. La durée de 1,5 h/jour semble importante pour un seul fumeur ; cette information mis en évidence reflète probablement plus le temps cumulé sur l'ensemble des fumeurs du ménage. Inclure dans le modèle le nombre moyen de cigarettes fumées par jour à l'intérieur du logement permettrait probablement de préciser l'effet du tabagisme sur la contamination des poussières. Cette étude plus précise serait importante dans le contexte où un concept de *« third-*

hand smoke[34] » émerge [Matt et al., 2011].

Sources caractéristiques de l'intérieur du bâtiment

Charge en plomb des poussières du palier.
La charge en plomb des poussières au sol du palier d'appartement a été identifiée comme la source majeure de contamination des poussières intérieures au sol. On avait montré au chapitre 2 que les charges en plomb des poussières des parties communes étaient environ 4 fois plus élevées qu'à l'intérieur si on considère les moyennes géométrique et jusqu'à 15 fois plus élevées si on considère les valeurs maximales. Le présent résultat est en accord avec celui établi par Dixon *et al.* [Dixon et al., 2005b] qui ont étudié l'influence du risque plomb lié aux parties communes sur l'intérieur du logement. Les poussières des parties communes sont introduites à l'intérieur des logements par le flux normal des personnes, en particulier via les vêtements et les chaussures. Dixon *et al.* [Dixon et al., 2005b] faisaient l'hypothèse que les peintures au plomb des parties communes devaient être une source de contamination des poussières, tout comme les sols et les poussières extérieures contaminés introduites dans les parties communes. Pour vérifier cette hypothèse la covariable « Charge en Pb-palier » a été remplacée par une covariable indiquant la charge en plomb des revêtements du palier[35]. Son coefficient a été estimé dans le modèle en plomb acido-soluble à 0,023 (IC à 95 % = -0.107 ; 0,154) alors que la variable « Charge en Pb-palier » avait un coefficient estimé à 0,412 (IC à 95 % = 0,240 ; 0,585). Son coefficient a été estimé dans le modèle en plomb total à 0,004 (IC à 95 % = -0.130 ; 0,138) alors que la variable « Charge en Pb-palier » avait un coefficient estimé à 0,369 (IC à 95 % = 0,189 ; 0,549). Ce changement de variable n'a produit qu'un faible impact sur les estimations des coefficients des autres sources introduites dans le modèle. Ce résultat montre donc que le plomb des poussières du palier ne provient pas des peintures au plomb du palier. Ainsi, on peut faire l'hypothèse que la majeure partie du plomb des poussières des parties communes proviendraient des sols et poussières extérieurs. Néanmoins on ne peut pas exclure que les poussières à l'intérieur des logements elles-mêmes contaminent les poussières des parties communes. Cette dernière hypothèse sortait du champ d'étude du travail de recherche. Pour étudier de tels aspects à partir des données utilisées dans le présent travail, une modélisation multi-niveaux par équations structurelles pourrait être utilisée [Rabe-Hesketh et al., 2007].

Travaux à l'intérieur du logement dans les 6 mois précédant l'enquête.
La réalisation de travaux généraux de rénovation à l'intérieur du logement sans mesure spécifique de protection, a été estimée comme étant capable d'augmenter la charge en plomb des poussières au sol d'environ 16 à 17%. Cependant l'effet de cette source n'a été que faiblement prouvé à partir des données utilisées dans le présent travail. Dans l'étude de Dixon *et al.* [Dixon et al., 2012] seuls des travaux spécifiques (e.g. le remplacement de fenêtres) et non généraux, ont été montrés comme contri-

34. Le *« thirdhand smoke »* est la contamination des surfaces par les composés de la fumée de tabac.
35. Le maximum des mesures XRF a été simplement considéré.

buant à la contamination des poussières intérieures avec un effet prouvé. Le résultat obtenu dans le présent travail va donc dans le sens de ces précédents résultats. Le fait de considérer une variable « travaux (généraux) à l'intérieur » ne permet sans doute pas de discriminer suffisamment les travaux capables de disperser des poussières contenant du plomb. Dans le questionnaire de Plomb-Habitat, le type de travaux effectués a été précisé. Cependant les différentes modalités n'avaient que de trop faibles effectifs pour pouvoir étudier finement l'impact de chaque type de travaux. Il semble donc difficile sur une enquête à grande échelle, basée sur seulement 500 individus statistiques, de pouvoir étudier les différents types de travaux intérieurs capables de contaminer la poussière déposée au sol.

Somme des charge en plomb (XRF) des revêtements de la pièce. Les résultats indiquent que l'effet des peintures au plomb à travers les 2 covariables introduites dans le modèle, basées sur la somme des mesures XRF maximales de chaque UD d'une pièce, existe encore. Cependant la contribution de ces peintures, dégradées ou en état d'usage, n'est plus à même d'être importante en population générale de logements relativement aux contributions des autres sources identifiées. Pour que leur contribution apparaisse comme relativement importante, il est nécessaire que leurs valeurs (i.e. la somme de la mesure XRF des UD d'une pièce) dépassent 60 mg/cm^2 pour la variable en revêtement dégradé, et 140 mg/cm^2 pour la variable en revêtement en état d'usage. Ainsi aujourd'hui, seuls les logements avec des charges en plomb cumulées importantes, dans leurs revêtements intérieurs, sont à même d'être concernés par une contribution substantielle du plomb de leurs peintures pour contaminer leur poussière au sol. Ces logements appartiennent probablement à l'ensemble des logements anciens et peu rénovés. En effet dans notre échantillon, les logements ayant au moins une pièce avec une somme des mesures XRF sur UD dégradée supérieure à seulement 30 mg/cm^2, sont au nombre de 4 : 3 d'entre eux ont été construits avant 1915 et un logement a été construit entre 1915 et 1948. Pour les UD à revêtements en état d'usage dont la somme des mesures est supérieure à 30 mg/cm^2, 13 logements sont concernés : un seul a été construit entre 1949 et 1974, 5 ont été construits entre 1915 et 1948 et 7 ont été construits avant 1915.

Sources caractéristiques de l'extérieur du bâtiment

Mesure XRF sur le garde-corps de la terrasse/loggia/balcon.
La charge en plomb des garde-corps a été identifiée comme l'un des contributeurs majeurs à contaminer la poussière au sol à l'intérieur des logements dès lors que cette charge dépasse 2,6 mg/cm^2 (P95). Dans ce cas la charge en plomb dans les poussières intérieures au sol est augmentée d'environ 50%. Le minium de plomb a été largement utilisé en France dans les peintures extérieures comme agent antirouille jusqu'au milieu des années 1990 [Lucas, 2011]. À notre connaissance aucun étude ne s'était intéressée auparavant à l'association entre cette variable et le niveau en plomb dans les poussières intérieures. Cependant dès le début du 20^{e}siècle J. Lockhart Gibson croyait fortement en la responsabilité de ces barreaudages dans les cas de saturnsime (Gibson, 2005). La contribution de la peinture de ces barreaudages n'est pas évidente à comprendre de prime abord dans la mesure où ces surfaces peintes sont faibles relativement à la surface pouvant être peinte à l'inté-

rieur. Ainsi la première interprétation qui a été faite était plutôt celle d'un proxy. C'est-à-dire que la mesure XRF positive de ces barreaudages indiquait plus simplement la présence d'une terrasse, d'un balcon ou d'une loggia, autrement dit une aire extérieure sur laquelle se déposent et s'accumulent des poussières contaminées, qui contaminent ensuite l'intérieur du logement. Cette hypothèse avait été aussi faite dans une précédente étude [Tong & Lam, 2000]. Le remplacement dans le modèle de cette variable « XRF-garde corps », par une variable binaire « présence/absence » d'une terrasse, d'un balcon ou d'une loggia, n'a pas confirmé cette hypothèse. Dès lors, on peut penser que se trouvant à l'extérieur, ces surfaces peintes peuvent être moins souvent considérées comme à risque et donc moins souvent rénovées.

Travaux à l'extérieur du logement dans les 6 mois précédant l'enquête.
Il a été trouvé une contribution mineure et négative de tels travaux (de -17 à -12 %) alors que l'hypothèse faite étaient qu'ils constituaient une source de contamination. Un résultat analogue avait été trouvé dans l'étude de *Reissman et al.* [Reissman et al., 2002] et celle de Dixon *et al.* [Dixon et al., 2012]. Cependant dans leurs études et dans le présent travail, la mise en évidence de cet effet protecteur était faible. Par contre, Clark *et al.* [Clark et al., 2011] ont démontré que les travaux extérieurs pouvaient être bénéfiques pour faire diminuer la plombémie chez l'enfant lorsque que ces travaux concernent des revêtements contenant au moins 7 mg/cm^2 de plomb.

Sources liées à l'environnement extérieur du bâtiment

Charge en Pb-poussière extérieure et Concentration en Pb-sol extérieur (selon la fréquentation de l'aire de jeu par l'enfant).
Le sol de l'aire de jeu extérieure semble contribuer à contaminer les poussières intérieures du logement. L'augmentation de la charge en plomb des poussières due au sol extérieur va de 30 à 75 % environ. Cette contribution est visible sans même atteindre les percentiles les plus hauts de la distribution des concentrations en plomb dans le sol. Il n'a pas été possible à partir des données utilisées dans le présent travail de mettre en évidence une contribution en fonction de la fréquentation de l'aire de jeux. On pouvait s'attendre à ce que la contribution quand l'enfant joue « tout le temps » sur l'aire de jeu soit supérieure à celle où l'enfant joue « souvent ». La non vérification de cette attente peut être due à la collecte de l'information de la fréquentation de l'aire sous forme qualitative d'une part, et d'autre part il s'agissait d'une information assez subjective évaluée part l'enquêteur.
Les sols extérieurs avaient déjà été suspectés comme étant une source majeure de la contamination des poussières intérieures [Hunt et al., 2012]. Cependant il semble qu'aucune étude n'avait auparavant quantifié la contribution des sols extérieurs dans cette contamination. La plupart des études concernaient la relation plomb des sols extérieurs/plombémie. Lanphear et Roghmann [Lanphear & Roghmann, 1997] avaient néanmoins trouvé que la contribution des peintures à base de plomb était plus importante que celle des sols extérieurs contaminés. Mais plus de 15 ans se sont écoulés entre la collection de leurs données et notre enquête et les relations entre les compartiments environnementaux ont pu changer depuis.
De même, pour les poussières extérieures, aucune étude ne semble avoir quantifié la relation avec la poussière intérieure déposée au sol. Clark et al. [Clark et al., 2004]

ont étudié une relation analogue mais leur étude concernait les poussières extérieures à l'entrée du logement, et donc pas celles de l'aire de jeu de l'enfant. D'autre part, la relation n'a pas été quantifiée dans leur étude. À partir des données utilisées dans le présent travail, il n'a pas été possible de mettre en évidence l'effet des poussières extérieures. Cela peut être dû au plus faible nombre d'échantillons de poussières extérieures qui ont été collectés par rapport au nombre d'échantillons de sols (53 charges en plomb disponibles versus 315 concentrations). De plus la faible mise en évidence de l'effet peut être due au fait que les poussières du palier d'appartement peuvent avoir accaparé une partie de l'effet des poussières extérieures (cf. discussion à propos du palier d'appartement ci-après). Néanmoins l'importance de leur contribution (produisant environ 40 % d'augmentation en la charge en plomb des poussières) lorsque des valeurs en plomb dans la poussière extérieure atteignent des seuils de l'ordre du percentile 97,5 devrait être considérée avec attention.

Trafic routier sur la route la plus proche du logement.
Étant donné que les mesures du plomb dans le sol et la poussière extérieure sur les aires de jeu des enfants n'apportaient pas une information exhaustive, l'idée était de compléter cette information à travers la variable « Trafic routier » . De compléter cette information en particulier à travers l'impact sur les sols du trafic routier passé pour lequel, avant l'année 2000, les véhicules roulaient encore à l'essence à base de plomb. L'hypothèse était que ce trafic routier passé et le trafic actuel étaient corrélés. La contribution de cette variable « Trafic routier » a été estimée comme mineure, au plus 10% pour le modèle en plomb acido-soluble et 5% pour le plomb total. De plus son effet n'a été que faiblement mis en évidence à partir des données utilisées dans le présent travail. Il semble ainsi difficile de jouer en pratique sur l'effet indirect du trafic routier pour réduire la contamination des poussières intérieures. Agir sur d'autres sources discutées ici semble plus adéquate pour accomplir cet objectif.

Démolitions dans le passé ou rénovations de bâtiments anciens du voisinage du logement .
Le fait d'avoir subi des démolitions autour de son logement contribue à augmenter la charge en plomb des poussières intérieures au sol d'environ 30% quel que soit le type de plomb. Un tel effet a été observé par Rabito *et al.* [Rabito et al., 2007] mais sur les niveaux de plombémie et pas sur la charge en plomb des poussières intérieures. En outre Dixon *et al.* [Dixon et al., 2012] ont trouvé dans leur étude une contribution plus importante pour contaminer la poussière intérieure. Cependant leur covariable renseignait aussi de la présence d'usines à risque et pas seulement de démolitions. Néanmoins les résultats de ces 2 études vont le même sens que le résultat du présent travail, à savoir que les démolitions autour de logements sont à même de produire une contamination des poussières intérieures de ces logements. Dans notre étude, cette information étant qualitative, et la contribution de démolitions passées, elle ne peut donc pas être néanmoins évaluée finement.

Anciens sites industriels et activités de service (Basias).
Il a été montré que la réduction des niveaux en plomb dans les sols proches de sites industriels émetteurs de plomb permet de réduire la charge en plomb des poussières intérieures au sol [Lanphear et al., 2003]. Les résultats du présent travail confirment

l'influence de tels sites : selon la valeur de la variable « Basias » introduite dans le modèle, une augmentation entre 40 et 50% de la charge en plomb des poussières se produit. Un changement de la variable « Basias » passant de son percentile d'ordre 25 à son percentile d'ordre 75 produit une augmentation en la charge en plomb des poussières intérieures faisant partie des 3 plus fortes contributions à travers les sources étudiées.

Sites et sols pollués (Basol).
La contribution de la variable relative aux sites et sols pollués a été estimée comme négative. Un tel résultat n'était pas attendu. Néanmoins à partir des données utilisées dans le présent travail, la mise en évidence de cette contribution n'a pas du tout été établie.

Émissions polluantes (Bdrep).
La contribution de cette variable a certes été estimée positivement mais comme très mineure et son effet n'a été que très pauvrement mis en évidence à partir des données utilisées dans le présent travail. La concentration en plomb dans l'air a considérablement chuté depuis l'interdiction en 2000 de l'essence au plomb. Selon Layton et Beamer [Layton & Beamer, 2009] le plomb atmosphérique était la source principale de contamination de poussières intérieures au sol à Sacramento en Californie aux États-Unis, au début des années 1980. Ils ont fait l'hypothèse qu'après l'interdiction de l'essence au plomb, la contribution du sol extérieur tracté à l'intérieur du logement serait la source majeure de contamination des poussières intérieures au sol. Les résultats du présent travail sont en fait assez cohérents avec cette hypothèse.

Les distributions des covariables continues étaient dissymétriques à droite (*right-skewed*) et leur contribution n'est dès lors quantifiable que pour des valeurs atteignant leurs hauts percentiles. Dorénavant l'influence sur la contamination par le plomb des poussières intérieures déposées au sol semble n'être apparente qu'à partir des valeurs extrêmes des sources potentielles. Ceci doit néanmoins être relativisé pour le sol extérieur de l'aire de jeu, les sites polluants (Basias) et la poussière du palier où des effets sont visibles pour des gammes de valeurs plus faibles (P25-P75).

Il a été souligné que les poussières des parties communes peuvent avoir accaparé une partie de l'effet des sources extérieures. Il n'aurait pourtant pas été prudent de ne pas introduire toutes ces sources, poussières du palier et sources extérieures, dans le modèle à ajuster. En effet il semble invraisemblable qu'il n'y ait pas d'effet direct des sources extérieures sur les poussières intérieures des logements, c'est-à-dire sans être tributaire du palier. De plus les logements individuels, type maison, n'ont de toute manière pas de partie commune. Pour au moins ce type de logement, les sources extérieures devaient donc figurer dans le modèle. Ainsi il n'y a pas eu de double prise en compte d'une source à travers différentes variables utilisées dans le modèle.

Il n'y a pas de coefficient mesurant la qualité d'ajustement, tel que le coefficient détermination (R^2) en régression linéaire, qui est disponible pour la modélisation

multi-niveaux. Un tel coefficient a l'avantage d'être largement connu et interprétable dans la communauté scientifique. Il a l'avantage d'être un bon outil de communication lors de présentations de résultats. Le fait que le R^2 ne puisse pas être calculé simplement dans le cadre d'un modèle hiérarchique est que, dans ce cas, les données sont corrélées. La variance de la variable réponse est dès lors expliquée d'une part par les effets fixes de la partie « moyenne » du modèle, et d'autre part par les effets aléatoires dans la partie modélisation de la covariance du modèle multi-niveaux. De plus s'ajoutait à cette question, la manipulation des données d'enquête et donc l'introduction des poids de sondage dans la formule de calcul d'un coefficient de détermination adapté à ce type de données. Sur ce sujet aucune littérature disponible n'a été identifiée.
Bien que plusieurs méthodes ait été développées pour déterminer un tel coefficient [Kramer, 2005] pour la modélisation multi-niveaux, il semble qu'aucune méthode n'ait été largement adoptée. Les logiciels, en tout cas Stata V12, ne fournissent pas un tel coefficient. Afin de donner une information sur la qualité du modèle, la figure 44 montre la distribution des résidus et la figure 45 montre les (Log) charges en plomb des poussières (Y) prédites par le modèle en fonction des (Log) charges observées. La difficulté pour pouvoir tracer ces graphiques est de prendre en compte l'imputation multiple, c'est-à-dire de travailler sur $M = 100$ jeux de données. Les figures 44 et 45 ne concernent que l'estimation du modèle pour le premier jeu de données. Des graphiques obtenus à partir d'autres jeux de données sont en fait très proches. L'hypothèse de normalité des erreurs semble être valide. L'ajustement des données les plus élevées et les plus faibles de la variable réponse Y ne sont que modérément ajustées mais l'ajustement global est correct.

FIGURE 44 – Distribution des résidus du modèle (en plomb acido-soluble) sur le jeu de données imputées $M = 1$.

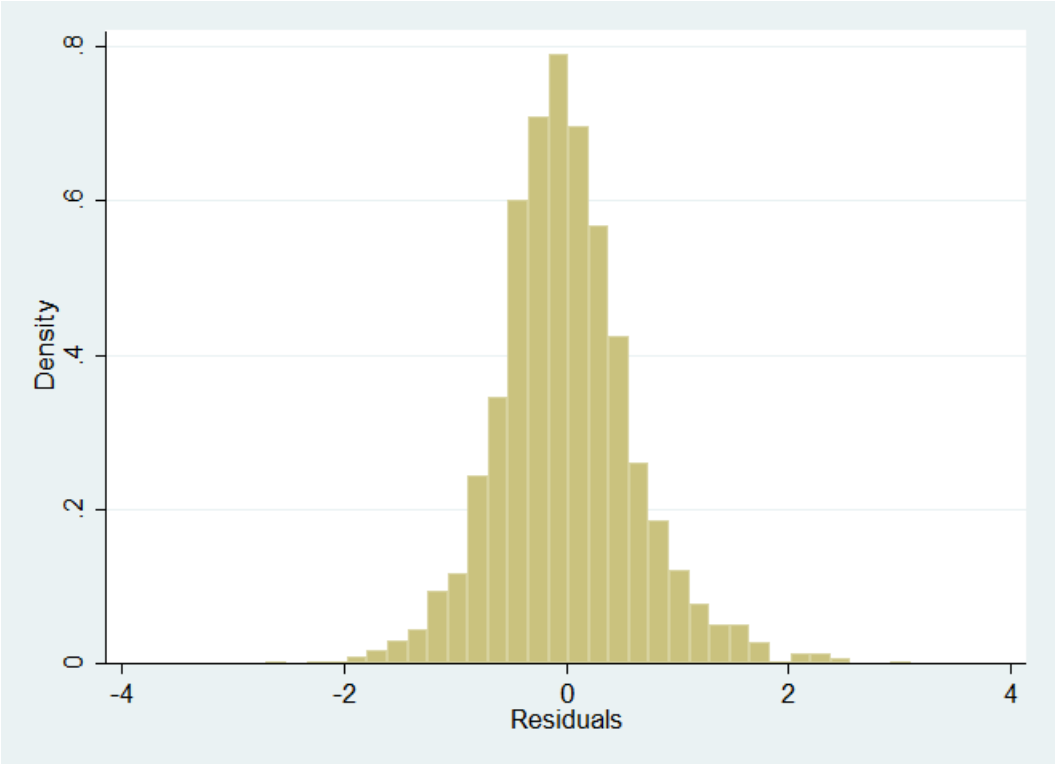
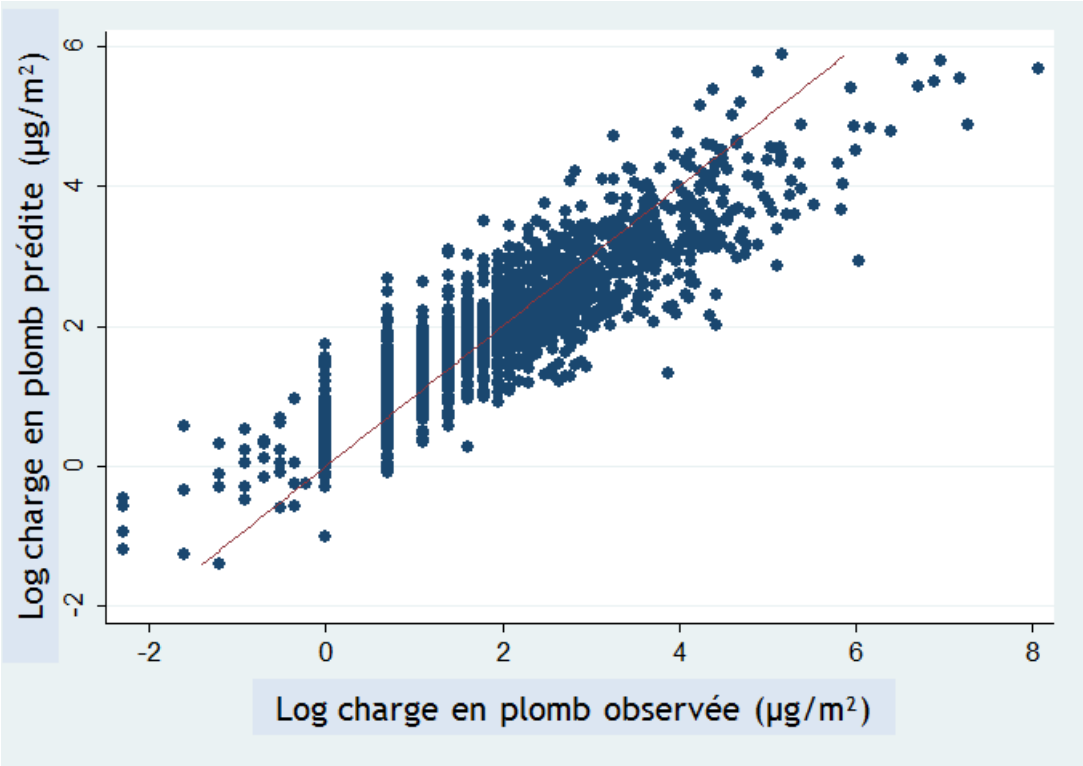


FIGURE 45 – y prédit versus y observés (jeu de données imputées $M = 1$). La droite est d'équation $y = x$.



3.5 Corrélation entre les charges en Pb de la poussière

La corrélation entre 2 (Log) charges en plomb au sein d'un même logement (ICC) a été estimée approximativement à $0,62$ (cf. section 6.4 du chapitre 3). Il est rappelé que le calcul de cette corrélation était lié au nombre de prélèvements des poussière par lingettes à réaliser afin de représenter au mieux le niveau de contamination par le plomb d'un logement.

Cette corrélation a été calculée à partir des charges en plomb transformées par un logarithme. L'ICC concerne donc les charges transformées et non les charges « directes » qui seules nous intéressent. Il n'était pas évident, en tout cas en qui me concerne, de savoir si le calcul de l'ICC fait à partir d'une transformation Log était alors d'une quelconque utilité. L'article [Euser et al., 2008] traite de ce sujet et plus généralement des indicateurs de reproductibilité calculés à partir de données Log-transformées, et en particulier de l'ICC. Il est indiqué dans cet article : « Bien que les ICC après une transformation logarithmique peuvent être calculés directement en estimant les composantes de la variance sur les données Log-transformées, un problème survient avec le calcul et l'interprétation des autres indicateurs de reproductibilité, à la fois en termes de fiabilité et de concordance. » Dès lors l'interprétation de l'ICC peut se faire directement semble-t-il pour les charges en plomb des poussières. Afin de s'assurer de la bonne compréhension de ce qui était écrit dans

l'article, un des auteurs de [Euser et al., 2008] a été contacté. Saskia le Cessie[36] a répondu sur ce sujet : « Le point principal sur lequel nous voulions insister dans cet article est que l'*ICC* sur données Log-transformées peut encore être interprété comme un *ICC*, mais que le coefficient de variation devient non interprétable. »

Il n'a pas été fourni d'intervalle de confiance pour les estimations de l'*ICC* dans la table 37 du chapitre 3. La raison est que les variances des estimateurs de σ_1^2 et σ_2^2, tout comme les estimateurs des coefficients de régression, sont estimées de façon « robuste[37] ». Il ne semble pas possible alors de fournir des estimations relatives à l'*ICC* avec l'utilisation d'estimateurs de variance robustes ; Stata V12 en tout cas ne le permet pas : `estat icc not allowed after estimation with robust standard errors`. Il semble que seule la construction d'un estimateur de la variance pour l'*ICC* n'est pas possible. En revanche pour ce qui est de l'*ICC* lui-même, une fois les estimations de σ_1^2 et σ_2^2 calculées, rien ne s'oppose au calcul algébrique de l'*ICC*. Les raisons algébriques qui font que cet estimateur de la variance de l'*ICC* ne puisse pas être calculé n'ont pas été identifiées. Les méthodes de ré-échantillonnage, type Bootstrap, devraient permettre de pouvoir fournir un intervalle de confiance pour l'*ICC*.

Habituellement, une valeur d'*ICC* comprise entre 40 % et 75 % est interprétée comme indiquant une reproductibilité passable à bonne [Rosner, 2006]. Cependant il est difficile à partir d'une telle valeur bâtarde d'*ICC* de conclure quant au fait qu'un seul prélèvement de poussière par lingette puisse suffire afin d'évaluer la contamination par le plomb d'un logement en ce qui concerne ses poussières. D'autre part toutes les pièces d'un logement n'ont pas été investiguées et certaines pièces n'ont donc pas eu de prélèvement de poussières. Le protocole de l'enquête Plomb-Habitat n'avait pas été spécialement dédié à cette question de l'*ICC*. Une future étude est donc nécessaire afin de confirmer cette valeur à partir de données collectées dans une enquête où toutes les pièces des logements doivent subir un prélèvement voire même plusieurs prélèvements de poussières.

Bien que quelques rares études telles que [Wilson et al., 2007], semblaient posséder des données permettant de fournir un *ICC* relatif à la charge en plomb des poussières d'un logement, il semble que l'estimation obtenu dans le présent travail soit la première à être fournie dans la littérature internationale (via l'article relatif ; cf. annexe 12).

36. *Associate professor of Medical Statistics, Leids Universitair Medisch Centrum, Leiden, Netherlands.*

37. Un estimateur de la variance est « robuste » dans le sens où cet estimateur peut calculer des variances correctes sans vérification de certaines hypothèses du modèle. En particulier l'estimateur robuste de la variance peut calculer des variances sans faire l'hypothèse d'indépendance des observations. Dans le modèle à 2 niveaux qui a été ajusté, c'est le fait d'avoir indiqué au logiciel que les logements (niveau 2) étaient inclus dans des unités supérieures (les hôpitaux), qui produit l'utilisation de cet estimateur pour le calcul des erreurs standards.

4 Évaluation par simulation de l'impact des poids de niveaux 2 introduits dans la pseudo-vraisemblance

Dans le chapitre 4 l'objectif était de déterminer quelle pondération dans un modèle multi-niveau, à 2 ou à 3 niveaux, était à utiliser afin d'estimer les contributions de sources en plomb dans la contamination des poussières intérieures déposées au sol par une expérience de Monte-Carlo. L'étude de la pondération au niveau 2 a pu être faite précisément puisque la problématique du type de pondération au niveau 1, pouvant introduire des estimateurs biaisés et inefficaces selon la littérature, n'existait pas sur les données utilisées dans le présent travail.

Il semble que les auteurs de la littérature du domaine de la modélisation multi-niveaux sur données d'enquête se soient jusqu'alors focalisés sur le niveau 1 des modèles (composant les unités échantillonnées en dernier). L'intérêt premier du niveau 1 semble naturel dans la mesure où ce sont les valeurs y_{ij}, d'une variable de niveau 1 qui sont modélisées.

Les études de simulation étudiant la pondération au niveau 1 dont les résultats ont été publiés dans la littérature ne concernent que des modèles à 2 niveaux, souvent où seul l'*« intercept »* est aléatoire. Il s'agit de plus de modèles relativement simples. Par exemple, Pferffermann *et al.* [Pfeffermann et al., 1998] ont réalisé une étude de simulation où le modèle est $y_{ij} = \beta_0 + \zeta_j + \epsilon_{ij}$, c'est-à-dire où 3 paramètres sont à estimer, avec les hypothèses de Normalité standards pour les 2 effets aléatoires. Rabe-Hesketh et Skrondal [Rabe-Hesketh & Skrondal, 2006] ont réalisé une étude de simulation où le modèle était un modèle logistique (Y binaire) à 2 niveaux à *« intercept »* aléatoire où seuls 2 coefficients de régression étaient à estimer, l'un concernant une covariable de niveau 1, l'autre concernant une covariable de niveau 2. Dans une étude de simulation utilisant des méthodes Bayésiennes, Pferffermann *et al.* [Pfeffermann et al., 2006] estiment un plus grand nombre de paramètres ; 9 paramètres dont 7 coefficients de régression. Les études de simulation ne peuvent résumer la « vraie vie » dans laquelle on est amené à estimer un plus grand nombre de coefficients de régression. Dans le chapitre 4 nous avons réalisé une étude où un nombre plus conséquent de paramètres étaient estimés car cette étude par simulation était motivée par l'ajustement d'un modèle sur données réelles : 33 paramètres dont 30 coefficients de régression.

4.1 Génération de populations

Dans le processus de simulation mis en place au chapitre 4, les réalisations de la variable réponse Y, c'est-à-dire la charge en plomb des poussière, ont été générées à partir d'un modèle à 2 niveaux à *« intercept »* aléatoire. La raison d'utiliser un modèle à 2 niveau est similaire à celle faite concernant le modèle multi-niveaux que l'on cherchait à ajuster au chapitre 3 : la variabilité des EPCI (resp. des hôpitaux) n'était pas d'intérêt. Cependant il n'est pas à exclure que dans le processus de

simulation, la génération des y_{ij} n'ait pas eu d'impact sur les résultats, en particulier en ce qui concerne le classement relatif des 9 scénarios les uns par rapport aux autres. La simulation pourrait être réitérée afin d'étudier cet aspect de la génération des y_{ij} en ajoutant donc un effet aléatoire pour les EPCI. On serait alors dans la situation d'un modèle à 3 niveaux sans covariables relatives aux EPCI (niveau 3), comme indiqué par les équations 6.2, 6.3, 6.4 en page 50. Dans ce cas, l'effet aléatoire ζ_k au niveau 3 suivra une loi $\mathcal{N}(0, \sigma_3^2)$ avec σ_3^2 pouvant être fixé à une valeur de $0,31$ [38].

4.2 Plan de sondage

La stratification par zone à risque plomb utilisée au niveau des PSU (hôpitaux) dans le plan de sondage Saturn-Inf/Plomb-Habitat n'a pu être reproduite comme indiqué en section 2 du chapitre 4. Ces zones à risque ont été construites par l'InVS relativement au bassin de population de chaque hôpital. Après discussion avec l'InVS il s'est révélé qu'il n'était pas faisable de reproduire facilement cette stratification pour les EPCI du plan de sondage utilisé dans la simulation.
De même la stratification par niveaux de plombémie, présente au niveau de la phase 2 du plan des enquêtes Saturn-Inf/Plomb-Habitat (cf. section 8.2 de la partie « Spécificités des données d'enquête ») n'a pas pu être reproduite dans le plan de sondage utilisé dans la simulation. La raison était que les enfants inclus dans la deuxième phase, i.e. dans l'enquête Plomb-Habitat, ont été remplacés par des logements dans la simulation. Cette stratification par niveaux de plombémie n'avait donc plus de sens.
Une stratification de remplacement aurait du être faite. Pour cela une variable auxiliaire en lien avec la charge en plomb des poussières (Y) aurait du être disponible dans le fichier logement généré (cf. section 1 du chapitre 4). Une variable auxiliaire naturelle est la période de construction du logement possible construite en trois strates : « $<$ 1949 » en remplacement de « $\geq 100\mu$g/L » en tant que strate la plus à risque plomb, « 1949-1974 » en remplacement de « $]30 - 100]\mu$g/L » en tant que strate moyennement à risque et « $\geq$ 1975 » en remplacement de « $< 30\mu$g/L » en tant que strate faiblement à risque. La période de construction n'a pas été utilisée finalement en tant que variable de stratification car d'une part elle est utilisée pour la post-stratification qui devait suivre. D'autre part et surtout, pour imiter la stratification de Saturn-Inf/Plomb-Habitat et l'échantillonnage dans la strate initiale « $\geq 100\mu$g/L », il aurait été nécessaire de n'avoir que 4 logements dans la strate « $<$ 1949 » dans le plan de la simulation. En effet seuls 4 enfants avec une plombémie $\geq 100\mu$g/L étaient présents parmi les 3623 enfants à l'issue de la phase 1 de Saturn-Inf/Plomb-Habitat. Or n'obtenir que 4 logements dans la strate « $<$ 1949 » du plan de sondage utilisé dans la simulation est hautement improbable. Pour aboutir à un tel résultat, il aurait été nécessaire de stratifier en amont dans le plan de sondage utilisé dans la simulation, c'est-à-dire stratifier selon la période de construction lors du tirage des logements (degré 2) de la phase 1 du plan de simulation. Mais une telle stratification n'aurait alors pas calqué le tirage des unités secondaires de la phase 1 de Saturn-Inf/Plomb-Habitat. Le choix a donc été de préférer perdre des stratifications du plan de sondage de Saturn-Inf/Plomb-Habitat plutôt que d'ajouter des

38. Cette valeur a été obtenue comme la moyenne des estimations de σ_3^2 obtenues à partir des 3 scénarios $\mathbf{w_7}$, $\mathbf{w_8}$ et $\mathbf{w_9}$ sur cas complets (cf. section 3 du chapitre 3).

stratifications artificielles dans le plan de sondage de la simulation.

Les résultats obtenus au chapitre 4 pour les scénarios $\mathbf{w_7}$, $\mathbf{w_8}$ et $\mathbf{w_9}$ représentant différents modèles à 3 niveaux, ne sont pas relatifs à 3 « vrais » degrés du plan de sondage. En effet le niveau 1 de ces 3 scénarios correspond aux pièces qui constituent bien un degré et de même pour le niveau des 484 logements. Par contre les EPCI utilisés dans la simulation ou de manière analogue les hôpitaux dans Saturn-Inf/Plomb-Habitat, ne constituent pas un niveau au dessus des logements. Ils sont bien hiérarchiquement au dessus des logements, mais un sous-échantillonnage les séparent plutôt qu'un degré dans le plan de sondage. Cela peut donc avoir induit un résultat différent en termes de classement des scénarios les uns par rapport aux autres, du résultat qui aurait été obtenu avec 3 « vrais » degrés dans le plan de sondage.
Il serait néanmoins possible d'étudier l'impact de la pondération au niveau 2 dans le cadre d'un plan avec 3 « vrais » degrés, en adaptant facilement notre programme de simulation.

Le plan de sondage de Saturn-Inf/Plomb-Habitat était un plan informatif pour Y (cf. section 5 de la partie « Spécificités des données d'enquête »). Il était informatif puisque les PSU (hôpitaux) ont été sur-échantillonnés dans les régions administratives à risque plomb ainsi que dans les zones géographiques à risque plomb (cf. section 8.2 de la partie « Spécificités des données d'enquête »). La stratification selon la plombémie de l'enfant discutée précédemment rendait de plus le plan informatif pour Y car il y a eu sur-échantillonnage (inclusion systématique en fait) des enfants avec une plombémie « $\geq 100\mu$g/L ». Dans le plan de sondage utilisé dans la simulation, les sur-échantillonnages dans les zones à risque plomb et dans la strate « $\geq 100\mu$g/L » ont été perdus. Ces sur-échantillonnages rendaient le plan informatif pour Y, la charge en plomb des poussières. Cependant le plan de sondage utilisé dans la simulation est resté informatif grâce au sur-échantillonnage fait pour les EPCI par région administrative à risque plomb. Ce caractère informatif du plan de sondage de la simulation est assuré par le fait d'avoir simulé certaines covariables source plomb extérieures [39] selon la région administrative (cf. annexe 9). *In fine*, le plan de sondage utilisé dans la simulation est tout au plus « moins informatif » que le plan de sondage de Saturn-Inf/Plomb-Habitat.

D'autres paramètres supplémentaires tels que la fraction de sondage ou la taille des *« clusters »* ont un impact sur la qualité des estimateurs d'un modèle multiniveaux. Par exemple si la taille des *« clusters »* est trop petite, il a été montré que cela pouvait induire un biais des estimateurs de la pseudo-vraisemblance pour les coefficients de régression lorsque des poids au niveau 1 sont utilisés [Pfeffermann et al., 1998]. Ce type de paramètres n'ont pas été étudiés puisque l'objectif n'était pas d'observer leur effet mais simplement de déterminer quelle pondération pour les unités du niveau 2 était à privilégier sur les données utilisées dans le présent travail. Cependant on a pu montré que même si aucun poids n'était utilisé au niveau 1, un biais pouvait survenir si des poids non adéquates étaient utilisés au niveau 2 tel que

39. 4 covariables : trafic routier, Basol, Basol et Bdrep.

dans les scénarios $\mathbf{w_3}$, $\mathbf{w_2}$, $\mathbf{w_4}$, $\mathbf{w_9}$ et $\mathbf{w_6}$ (cf. section 4 du chapitre 4). Une étude de simulation pourrait être réalisée afin d'étudier l'impact de tels facteurs sur les estimateurs selon la pondération utilisée au niveau 2.

Aucun poids n'a été en effet utilisé au niveau 1 (pièce) car les pièces n'ont pas été échantillonnées avec des probabilités de sélection inégales ; ces probabilités étaient toutes égales à 1, quelque soit la pièce et quelque soit le logement. Dès lors quel est l'impact des poids de niveaux 2 lorsqu'une pondération est utilisée au niveau 1 ? D'après les résultats obtenus (cf. section 4 du chapitre 4) il semble plus précautionneux dans ce cas d'utiliser classiquement les poids conditionnels des unités du niveau 2, car les résultats obtenus pour les scénarios avec poids conditionnels (ou assimilés conditonnels) au niveau 2 ($\mathbf{w_1}$ et $\mathbf{w_8}$) étaient proches des résultats du scénarios $\mathbf{w_5}$ jugé le meilleur. Néanmoins cela reste à confirmer par une étude dédiée.

4.3 Stratification et pseudo-vraisemblance

Dans la commande de StataV12 `xtmixed` qui estime les paramètres d'un modèle multi-niveaux en maximisant la pseudo-vraisemblance, il n'est pas possible de déclarer la stratification (au niveau des EPCI ou de manière équivalent des hôpitaux). Cela aurait pu améliorer l'efficacité des estimateurs et changer alors possiblement le classement relatif des scénarios les un par rapport aux autres. Bien que l'estimation basée sur la pseudo-vraisemblance semble pouvoir prendre en compte la stratification (au degré 1 i.e. pour les PSU) comme expliqué dans [Rabe-Hesketh & Skrondal, 2006, section 5], la commande `xtmixed` ne le permet pas. Pour l'utilisateur il n'est pas évident de savoir comment manipuler cette stratification en pratique même si cela est expliqué en théorie par Rabe-Hesketh et Skrondal [Rabe-Hesketh & Skrondal, 2006]. Dans le manuel d'utilisation du programme `gllamm` [Rabe-Hesketh et al., 2004] que Rabe-Hesketh et Skrondal ont développé pour fonctionner sous StataV12 et qui traite la stratification selon [Rabe-Hesketh & Skrondal, 2006, section 5], la stratification n'est pas abordée [40]. Il serait donc bon de permettre aux utilisateurs, aguerris comme non aguerris à la modélisation multi-niveaux sur données d'enquête, d'identifier plus clairement comment en pratique indiquer la stratification dans les logiciels.

4.4 Résultats et recommandations

Dans l'étude de simulation réalisée au chapitre 4 le comportement des estimateurs de pseudo-vraisemblance a été étudié notamment lorsque les niveaux du modèle multi-niveaux ne correspondent pas aux degrés du plan de sondage. En particulier la situation étudiée est celle où il existe des unités hiérarchiquement au dessus du plus haut niveau utilisé dans la modèle (cas des scénarios $\mathbf{w_1}$ à $\mathbf{w_6}$). Parmi ces 6 scénarios, les scénarios $\mathbf{w_2}$, $\mathbf{w_3}$, $\mathbf{w_4}$ et $\mathbf{w_6}$ se sont révélés les plus mauvais : la distribution du biais relatif de leurs estimateurs possède une plus forte dispersion (figure

40. Le mot « stratification » n'apparaît même pas dans le manuel.

38 de la section 4 du chapitre 4) et l'efficatité des estimateurs (leur variance puis leur REQM) de chaque paramètre du modèle est moins bonne. On rappelle la définition des pondérations au niveau 2 pour ces 6 modèles à 2 niveaux :

1. $\mathbf{w_1} : 1/\pi_j^b$
2. $\mathbf{w_2} : w_j^b$
3. $\mathbf{w_3} : \widetilde{w}_j^b$
4. $\mathbf{w_4} : 1/(\pi_j^a \times \pi_j^b)$
5. $\mathbf{w_5} : 1$
6. $\mathbf{w_6} : 1/(\pi_{j|k} \times \pi_j^b)$

Ainsi les résultats permettent de déduire qu'introduire des poids au niveaux 2 qui ne sont pas des poids conditionnels ou assimilés, n'est pas une bonne idée lorsque les niveaux du modèle multi-niveaux ne correspondent pas à la hiérarchie du plan. En effet, hormis le poids $\mathbf{w_5}$ particulier (analyse non pondérée), les poids $\mathbf{w_2}$, $\mathbf{w_3}$ et $\mathbf{w_4}$ sont des poids de sondage finaux à une ou plusieurs post-stratifications près (cf. section 2 du chapitre 4). Le cas $\mathbf{w_6}$ n'est pas un poids de sondage final (il manque au numérateur un facteur π_k, relatif aux EPCI) mais n'est pas non plus un « véritable » poids conditionnel dans la mesure où il y a un « facteur de trop », $\pi_{j|k}$ (probabilité conditionnelle entre les logements et les EPCI à la phase 1.).
Il peut sembler néanmoins naturel d'introduire un poids de sondage final au niveau 2, lorsque des unités existent hiérarchiquement au dessus du plus haut niveau utilisé dans la modèle. En effet, cela se base sur la réflexion suivante : en ne considérant pas les unités hiérarchiquement au dessus du plus haut niveau du modèle, on perd de l'information (du plan de sondage), et on tente alors par l'introduction de poids finaux de récupérer cette information perdue. C'est ce qui a été fait dans l'application numérique de Rabe-Hesketh et Skrondal [Rabe-Hesketh & Skrondal, 2006], basée sur des données de la base PISA [41]. Les données étaient issues d'un plan à 3 degrés (niv. 3 : région, niv. 2 : école, niv. 1 : élève). Les auteurs ont ajusté un modèle à 2 niveaux (élève + école). Le poids introduits pour les unités de niveau 2 est le poids de sondage final pour les écoles, ce qui correspond à notre scénario $\mathbf{w_3}$. Or ce scénario s'est révélé être parmi les 5 plus mauvais. La non utilisation du niveau 3 (les régions) a été justifiée par les auteurs, d'abord par le fait que la variance entre les régions n'était pas d'intérêt puis par le fait que les probabilités de sélection des régions n'étaient pas disponibles [42]. La non fourniture des probabilités de sélection est certainement en fait la première raison de la non utilisation d'un modèle à 3 niveaux bien que, la variance entre les unités de niveaux supérieurs puisse en effet ne pas être d'intérêt. C'était aussi le cas pour les données utilisées dans le présent travail : la variance due aux hôpitaux n'était pas recherchée (cf. section 1 du chapitre 3). Cette base de données publique PISA illustre donc bien ce qui peut se passer lorsque l'on ajuste un modèle multi-niveaux sur des données publiques : ne pas pouvoir ajuster un modèle dont les niveaux correspondent à ceux du plan de sondage. Dans ce cas, les résultats du présent travail montrent qu'il faut être prudent sur la pondération à introduire

41. *Programme for International Student Assessment.*
42. Ni même les identifiants des régions initialement ; ces identifiants ont été obtenus après une requête auprès du détenteur des données.

pour les unités du plus haut niveau du modèle.
En effet dans l'expression de la Log pseudo-vraisemblance :

$$\sum_{j=1}^{n^{(2)}} w_j^{(2)} \log \int \exp \left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j}^{(1)} \log \left(f(y_{ij} \mid \zeta_j) \right) \right\} g(\zeta_j) \mathrm{d}\zeta_j$$

l'introduction des poids à chaque niveau doit se faire avec des poids basés sur des probabilités de sélection conditionnelles. Dans l'expression ci-dessus concernant un modèle à 2 niveaux, les poids $w_j^{(2)}$ ne sont pas notés sous forme conditionnelle i.e. $w_{j|k}^{(2)}$ car il n'y a pas de niveau supérieur supposé relatif aux unités k. $w_j^{(2)}$ est donc un poids de sondage final pour les unités du niveau 2 mais il est aussi un poids conditionnel car relatif au premier degré d'un plan où poids final et poids conditionnel se confondent naturellement. $w_j^{(2)}$ est donc un poids conditionnel. Dès lors, rien ne justifie d'introduire autre chose que des poids conditionnels aux différents niveaux apparaissant dans l'expression de la Log pseudo-vraisemblance.
Concernant cette particularité d'ajuster un modèle avec moins de niveaux que de degrés du plan, des explications ont été demandées à plusieurs auteurs de la littérature du domaine avant de réaliser l'étude de simulation : quelles étaient les recommandations dans cette situation en termes de pondération au niveau 2 ? À deux reprises Sophia Rabe-Hesketh [43] a été contactée par email ainsi que Adam C. Carle [44]. Aucun retour n'a été obtenu de leur part.

Les résultats du présent travail montrent de plus que les analyses pondérées ne sont pas toujours plus performantes que les analyses non pondérées. Cela indique donc qu'il est possible d'ajuster des modèles multi-niveaux sur données publiques lorsque certaines informations ne sont pas fournies, en particulier des informations concernant les PSU.

D'après les résultats obtenus au chapitre 4, lorsque l'on souhaite ajuster un modèle multi-niveaux, il est recommandé de ne pas utiliser les poids mis à disposition lorsque ces poids sont des poids de sondage finaux, ou bien lorsque l'on n'est pas assuré que ces poids soient de « purs » poids conditionnels.

De plus il semble que procéder à une analyse de sensibilité en ajustant le modèle recherché selon plusieurs pondérations puisse être utile, en particulier en incluant une analyse non pondérée. Cette recommandation de réaliser une analyse non pondérée avait déjà été faite pour le cas des poids au niveau 1 [Carle, 2009]. Si des différences existent entre les estimations induisant alors des différences en termes décisionnels, il est préférable de réaliser une étude de simulation à partir de ses données : simuler une population finie à partir du modèle que l'on cherche à ajuster, sélectionner un échantillon en imitant le plan de sondage de l'enquête ayant fourni les données, et étudier si les paramètres du modèles sont correctement estimés selon

43. *University of California, Berkeley, USA, and Institute of Education, London, UK.*
44. *Department of Psychology, University of North Florida, 1 UNF Drive, Jacksonville, FL, 32224, USA.*

les différentes méthodes de pondération. Rabe-Hesketh et Skrondal [Rabe-Hesketh & Skrondal, 2006] recommandaient aussi ceci après avoir étudié le cas de la pondération au niveau 1.

Enfin, l'étude de simulation s'est basée sur des données réelles, celles collectées dans l'enquête Plomb-Habitat. L'objectif de l'étude de simulation ne permet pas à ses résultats d'être généralisables comme discuté précédemment. Néanmoins les résultats obtenus sur une situation particulière permettent de mettre en garde les utilisateurs ajustant des modèles multi-niveaux sur d'autres données d'enquêtes. De plus, étant basé sur des données non artificielles, les résultats peuvent se transposer à d'autres domaines en sciences environnementales tels que la qualité de l'air intérieure. En effet dans ce dernier domaine, les distributions des polluants sont très souvent Log-Normales comme les distributions des niveaux en plomb dans les différents compartiments environnementaux résidentiels. De plus dans le domaine de la pollution de l'air intérieur, les sources de contamination sont à la fois externes et internes aux logements, comme le cas de la contamination intérieure par le plomb.

Conclusion

Ce travail avait pour objectif d'établir un état de la contamination par le plomb dans les résidences principales abritant au moins un enfant âgé de 6 mois à 6 ans en France métropolitaine. Il avait également pour objectif de déterminer la contribution des sources de contamination des poussières intérieures déposées au sol. Ce travail s'est basé sur des données collectées dans le cadre d'une enquête environnementale réalisée dans près de 500 logements. Il a intégré la préparation et la validation des données par des procédures de *« data management »*.

Ce travail met en évidence les principaux résultats suivants :

- Les niveaux en plomb dans les différents compartiments environnementaux en milieu résidentiel ont été estimés pour la première fois en France :
  - Les niveaux exprimés en moyenne géométrique sont de :
    - moins de 1 μg/L pour l'eau du robinet ;
    - 8,8 μg/m^2 en plomb total et 6,9 μg/m^2 en plomb acido-soluble pour la poussière intérieure déposée au sol ;
    - 32,1 μg/m^2 en plomb total et 27,5 μg/m^2 en plomb acido-soluble pour la poussière déposée au sol en parties communes ;
    - 33,9 mg/kg en plomb total et 21,7 mg/kg en plomb acido-soluble pour les sols meubles de l'aire de jeu extérieure ;
    - 44,4 μg/m^2 en plomb total et 36,9 μg/m^2 en plomb acido-soluble pour les sols durs de l'aire de jeu extérieure.
  - Environ 2,5 % des logements ont une concentration en plomb dans l'eau du robinet $\geq 10\mu$g/L, le seuil réglementaire européen en vigueur fin 2013.
  - Approximativement 0,21 % des logements et 4,1 % des parties communes ont une charge en plomb dans leurs poussières déposées au sol $\geq 430,5\mu$g/m^2, le seuil réglementaire américain[45]. Les charges en plomb en parties communes sont 73 % plus élevées que les charges à l'intérieur des logements.
  - Les niveaux des charges en plomb dans les poussières intérieures sont approximativement les mêmes dans les logements construits avant 1949 et ceux construits à partir 1949 jusqu'au début des années 1990. Ils sont plus

45. Il n'y a pas de seuil réglementaire européen à ce jour.

faibles dans les logements construits après le début des années 1990.

- Environ 24,5 % des logements et 34,2 % des parties communes possèdent encore des revêtements à base de plomb [46]. Environ 4,7 % des logements et 7,1 % des parties communes possèdent des revêtements dégradés à base de plomb.

- Les logements construits à partir de 1949 jusqu'au début des années 1970 contiennent encore de la peinture au plomb mais de manière résiduelle.

- Approximativement 1,4 % des sols des aires de jeu extérieures des enfants ont une concentration en plomb ≥ 400 mg/kg, le seuil réglementaire américain.

- La corrélation entre deux charges en plomb dans la poussière intérieure au sol d'un même logement est égale approximativement à 0,60.

- Les contributions conjointes des sources pouvant contaminer en plomb la poussière du sol à l'intérieur des logements ont été estimées pour la première fois pour un nombre important de sources potentielles :

  - La poussière contaminée du palier d'appartement déposée au sol est le contributeur majeur.

  - Les contributeurs secondaires sont le sol contaminé de l'aire de jeu extérieure de l'enfant, la peinture au plomb des garde-corps extérieurs et le tabagisme à l'intérieur du logement.

  - Les peintures intérieures à base de plomb dans les logements abritant des enfants en bas âge, ne semblent plus contribuer de manière substantielle à contaminer la poussière intérieure. Les logements anciens ($<$ 1949 et surtout $<$ 1915) constituent des situations exceptionnelles dans lesquelles cependant, la contribution des peintures à base de plomb existe mais reste inférieure à celles des contributeurs majoritaires et secondaires.

- Il a été mis en évidence sur des données réelles d'enquête [47] qu'une pondération non adéquate au niveau 2 d'un modèle multi-niveaux (à 2 ou à 3 niveaux) pouvait induire des estimateurs biaisés. Ceci a été étudié dans la situation où les niveaux du modèle ne se calquent pas totalement sur les degrés du plan de sondage ; situation souvent rencontrée avec des bases de données publiques.

46. Dans le sens d'au moins une unité de diagnostic dont le revêtement contient une charge en plomb ≥ 1 mg/cm^2.

47. Dans le cadre d'un modèle expliquant la variabilité de la charge en plomb des poussières mesurée dans les pièces (niveau 1) de logements (niveau 2).

- La pondération au niveau 2 (par exemple un niveau « logement ») d'un modèle multi-niveaux a été étudiée par une expérience de Monte Carlo alors que jusqu'alors seule la pondération au niveau 1 (par exemple un niveau « pièce ») avait été étudiée par simulation dans la littérature.

Les perspectives de ce travail sont les suivantes :

- Définir le concept de logement à risque plomb pour les occupants, sur la base d'une agrégation des niveaux en plomb des différents compartiments environnementaux du milieu résidentiel.

- Proposer un diagnostic permettant d'évaluer le risque plomb « global » d'un logement. Le Constat de Risque d'Exposition au Plomb (CREP) diagnostique dans un logement le risque plomb essentiellement lié aux revêtements à base de plomb. On pourrait proposer un modèle prédictif du risque plomb (i.e. pas seulement lié aux revêtements à base de plomb) afin de fournir un « nouveau CREP ». La qualité de ce modèle prédictif du risque plomb pourrait être diagnostiquée à partir des niveaux en plomb dans la poussière puisque ces dernières ont été jugées [48] comme un bon indicateur du risque plomb global d'un logement. Ce « nouveau CREP » ne devrait pas être plus contraignant *in situ* que le CREP actuel et ne devra pas être plus coûteux : à ce jour il semble que la seule mesure instantanée du plomb surfacique [49], utilisée le cas échéant, puisse respecter ces impératifs. Mais il pourrait intégrer d'autres informations que des mesures, informations faciles à collecter à faible coût (e.g. situation du logement, date de construction du logement). En ce qui concerne le protocole de mesurage du plomb dans les revêtements, il devrait agréger dans sa conclusion l'ensemble des niveaux des charges en plomb, avec l'état de dégradation des revêtements, et avec la quantité de surfaces mesurées ; le cas échéant on pourrait y associer l'accessibilité des revêtements si on cherche à évaluer le risque plomb pour l'enfant.

- Compléter l'étude des voies de contamination entre les différentes pièces d'un logement, notamment entre l'intérieur des logements et les parties communes, et plus généralement avec les autres de sources de contamination (notamment extérieures). Ceci pourrait être fait dans le cadre d'une étude par ce qui est appelé réseaux de « causalité », au moyen d'une modélisation multi-niveaux par équations structurelles par exemple.

- Étudier plus généralement l'impact de la pondération au niveau 2 dans le cadre d'un modèle multi-niveaux [50] ajusté [51] sur des données d'enquête issues d'un plan de sondage à 3 degrés (et plus), dans la situation où le nombre de niveaux du modèle est inférieur au nombre de degrés du plan. L'impact sur les estimations du nombre d'unités de niveau 1 contenu dans les unités de niveau

48. Par exemple dans [Lanphear et al., 2005].
49. Par fluorescence X.
50. Appelé encore classiquement modèle mixte ou modèle hiérarchique.
51. Par pseudo maximum de vraisemblance.

2 pourrait être évalué. On pourrait également étudier l'impact de la fraction de sondage.

Annexes

1 Codifications INSEE des 22 régions métropolitaines françaises

REGION	NCC
11	ÎLE-DE-FRANCE
21	CHAMPAGNE-ARDENNE
22	PICARDIE
23	HAUTE-NORMANDIE
24	CENTRE
25	BASSE-NORMANDIE
26	BOURGOGNE
31	NORD-PAS-DE-CALAIS
41	LORRAINE
42	ALSACE
43	FRANCHE-COMTE
52	PAYS DE LA LOIRE
53	BRETAGNE
54	POITOU-CHARENTES
72	AQUITAINE
73	MIDI-PYRENEES
74	LIMOUSIN
82	RHÔNE-ALPES
83	AUVERGNE
91	LANGUEDOC-ROUSSILLON
93	PROVENCE-ALPES-COTE D'AZUR
94	CORSE

2 Études concernant les informations en lien avec le plomb des poussières ou avec la plombémie

Information	Références
Emplacement du logement	[Tong & Lam, 2000]
Saison	[Laxen et al., 1988, Yiin et al., 2000] [Laidlaw et al., 2005, Laidlaw & Filippelli, 2008] [Mielke et al., 2010, Clark et al., 2011] [Dixon et al., 2012]
Nettoyage humide possible sur le sol du palier d'appartement	[Gaitens et al., 2009]
Type de pièce	[Wilson et al., 2007]
Fréquence hebdomadaire de nettoyage du sol par aspiration ou balayage Fréquence hebdomadaire de nettoyage du sol par serpillère ou balai-éponge	[Yiin et al., 2003, Bretin, 2006] [Gaitens et al., 2009]
Endroit du prélèvement dans la pièce	[U.S. HUD, 2003, Bretin, 2006] [Wilson et al., 2007]
Nb. d'activités prof. à risque pratiquées par les membres du foyer	[Sanborn et al., 2002, Schapiro & Bretin, 2006]
Mesure XRF sur le garde-corps de la terrasse/loggia/balcon	[Gibson, 2005]
[Pb] sol_dur (× Fréquentation de l'aire de jeu par l'enfant) [Pb] sol_meuble (× Fréquentation de l'aire de jeu par l'enfant)	[Fergusson et al., 1986, Thornton et al., 1990] [Succop et al., 1998, Clark et al., 2004] [Dixon et al., 2005b, Caravanos et al., 2006] [Dixon et al., 2008, Hunt et al., 2006] [Hunt et al., 2012]
[Pb] poussière du palier	[Dixon et al., 2005b]
Trafic routier. Nb. de véhicules/an sur la route la plus proche du logement	[Sheets et al., 2001, Mielke et al., 2010] [Mielke et al., 2011]
Démolitions ou rénovations dans le passé de bâtiments du voisinage du logement	[Farfel et al., 2003, Farfel et al., 2005] [Rabito et al., 2007, Dixon et al., 2012]
Loisirs à risque pratiqués par les membres du foyer à l'intérieur du logement	[Hozhabri et al., 2004, Schapiro & Bretin, 2006]
Travaux sur l'extérieur du logement dans les 6 mois passés Travaux à l'intérieur du logement dans les 6 mois passés	[Rabinowitz et al., 1985, U.S. EPA, 2000] [Reissman et al., 2002, Dixon et al., 2005a] [Clark et al., 2011, Dixon et al., 2012]
Anciens sites industriels et activités de service (Basias)	[Davies et al., 1985, Cook et al., 1993] [Lanphear et al., 2003]
Sites et sols pollués (Basol)	[Davies et al., 1987, Lanphear & Roghmann, 1997] [Lanphear et al., 2003, Clark et al., 2004] [Hunt et al., 2006]
Emissions polluantes (Bdrep)	[Layton & Beamer, 2009]
Tabagisme	[Gaitens et al., 2009]
Charge en plomb (XRF) des revêtements de la pièce : somme des XRF des UD à dégradées Charge en plomb (XRF) des revêtements de la pièce : somme des XRF des UD en état d'usage	[Sturges & Harrison, 1985, U.S. HUD, 1995b] [Lanphear & Roghmann, 1997, Farley, 1998] [Succop et al., 1998, CDC, 2007] [Beauchemin et al., 2011, Dixon et al., 2012]

3 Variables introduites dans le modèle multi-niveaux

Label	Type	Source/ Confusion	Description et niveau d'information	Modalité	Transformation
Charge en Pb	num	-	Charge en Pb dans la poussière au sol (µg/m²) prélevée par lingette. Niveau 1 (pièce).	-	Log(y)
Emplacement du logement	disc	Confusion	Étage de l'entrée du logement. Niveau 2 (logement).	- 1=Semi-enterré (réf.) - 2=Rez de chaussée - 3=En étage	-
Saison	disc	Confusion	Période de l'année où a lieu l'enquête. Niveau 2 (logement).	- 0=Automne/hiver (réf.) - 1=Printemps/été	-
Lavage humide du palier	disc	Confusion	Si oui ou non un moyen humide peut être appliqué sur le palier d'appartement. Oui pour serpillère ou balai éponge, Non pour balai ou aspirateur. Niveau 2 (logement).	- 0=Pas de palier (réf.) - 1=Oui - 2=Non	-
Type de pièce	disc	Confusion	Type de la pièce investiguée. Niveau 1 (pièce).	- 1=Chambre de l'enfant ou d'un autre enfant - 2=Entrée (réf.) - 3=Salon - 4=Cuisine - 5=Salle de jeu	-
Fréq. Lavage humide-pièce	num	Confusion	Fréquence hebdomadaire de nettoyage du sol par moyen humide. Niveau 1 (pièce).	-	Log(x+1)
Fréq. Lavage sec-pièce	num	Confusion	Fréquence hebdomadaire de nettoyage du sol par moyen non humide. Niveau 1 (pièce).	-	Log(x+1)
Endroit du prélèvement poussière	disc	Confusion	Endroit dans la pièce où a eu lieu le prélèvement de poussière. Niveau 1 (pièce).	- 0=Endroit de jeu préféré de l'enfant (réf.) - 1=Au centre de la pièce	-
Nombre d'activités à risque	num	Source	Nombre d'activités à risque [a] (comme profession ou loisir éventuellement) pratiquées par les membres du ménages. Niveau 2 (logement).	-	Log(x+1)
XRF garde-corps	num	Source	Mesure XRF (mg/cm²) réalisée sur le garde-corps du balcon/loggia/terrasse; mise à 0 si	-	Log(x+1)

Label	Type	Source/ Confusion	Description et niveau d'information	Modalité	Transformation
			pas de terrasse/loggia/balcon. Niveau 2 (logement).		
Concentration en Pb-sol extérieur	num×disc	Source	Concentration en Pb (mg/kg) du sol (meuble) de l'aire de jeu extérieure de l'enfant en interaction (×) avec la fréquence de fréquentation de cette aire. La concentration est mise à 0 si l'enfant ne jouait pas à l'extérieur ou s'il jouait sur une surface dure. Niveau 2 (logement).	- 0=L'enfant ne joue pas à l'extérieur (réf.) - 1=L'enfant y joue souvent - 2=L'enfant y joue tout le temps	Log(x+1)
Charge en Pb-poussière extérieure	num×disc	Source	Charge en Pb (mg/kg) du sol (dur) de l'aire de jeu extérieure de l'enfant en interaction (×) avec la fréquence de fréquentation de cette aire. La charge est mise à 0 si l'enfant ne jouait pas à l'extérieur ou s'il jouait sur une surface meuble. Niveau 2 (logement).	- 0=L'enfant ne joue pas à l'extérieur (réf.) - 1=L'enfant y joue souvent - 2=L'enfant y joue tout le temps	Log(x+1)
Charge en Pb-palier	num	Source	Charge en Pb dans la poussière au sol (µg/m²) du palier d'appartement prélevée par lingette; mise à 0 si pas de partie commune. Niveau 2 (logement).	-	Log(x+1)
Trafic routier	num	Source	Flux annuel de véhicule de la route la plus proche du logement divisée par la distance (km) entre la route et le logement. Voir après le tableau pour sa construction. Niveau 2 (logement).	-	Log(x)
Démolition	disc	Source	Si de vieux bâtiments ont été détruits ou rénovés par le passé dans un rayon de 50m autour du logement. Niveau 2 (logement).	-1=Oui -2=Non (réf.)	-
Fréquence de loisir[b]	num	Source	Nombre de fois par an où un loisir à risque est pratiqué à l'intérieur du logement. Niveau	-	Log(x+1)

Label	Type	Source/ Confusion	Description et niveau d'information	Modalité	Transformation
			2 (logement).		
Travaux extérieurs[c]	disc	Source	Si des travaux de rénovation ont été faits à l'extérieur dans les 6 derniers mois. Niveau 2 (logement).	-1=Oui -2=Non (réf.)	-
Travaux intérieurs[d]	disc	Source	Si des travaux de rénovation ont été faits à l'intérieur dans les 6 derniers mois. Niveau 2 (logement).	-1=Oui -2=Non (réf.)	-
Sites polluants (Basias)	num	Source	Score relatif aux sites industriels et aux activités de service dans un rayon de 2km, actuels ou anciens, ayant une activité potentielle polluante (Pb). Voir après le tableau pour sa construction. Niveau 2 (logement).	-	Log(x+1)
Sols pollués (Basol)	num	Source	Score relatif aux sites et sols pollués potentiellement contaminés par le plomb dans un rayon de 2km, impliquant une action gouvernementale préventive ou curative. Voir après le tableau pour sa construction. Niveau 2 (logement).	-	Log(x+1)
Pb dans l'air (Bdrep)	num	Source	Score relatif aux usines sujettes à autorisation (industrie et élevage). Voir après le tableau pour sa construction. Niveau 2 (logement).	-	Log(x+1)
Tabagisme journalier	num	Source	Durée moyenne journalière où au moins une personne fume à l'intérieur. Niveau 2 (logement).	-	Log(x+1)
Somme XRF-détérioré	num	Source	Somme des mesures XRF maximale de chaque UD dans une pièce. Uniquement les UD au revêtement dégradé[e]. Niveau 1 (pièce).	-	Log(x+1)
Somme XRF-état d'usage	num	Source	Somme des mesures XRF maximale de chaque UD dans une pièce. Uniquement les UD au revêtement en été d'usage[f]. Niveau 1 (pièce).	-	Log(x+1)

Légende. UD : unité de diagnostic ; num : numérique ; disc : discrète.
a : Fabrication de fils ou de bâtons de soudure (en revanche, leur utilisation est, en principe, sans danger car les températures de mise en œuvre sont insuffisantes pour produire une exposition notable) ; Fabrication de batteries d'accumulateurs ; Fabrication de pigments, peintures, vernis contenant des dérivés inorganiques du plomb, ainsi que leur application en aérosol (pistolet) ou leur usinage ; Typographie et linotypie (procédés d'imprimerie en voie d'abandon) ; Fabrication de protections contre les radiations ionisantes ; Fabrication et utilisation de munitions ; Production de verre (en particulier, de cristal) ; Production et utilisation d'émaux ; Fabrication ou rénovation de vitraux ; Production ou usinage de matières plastiques contenant du plomb, employé comme pigment ou stabilisant ; Production et utilisation de lubrifiants contenant du plomb ; Réparation de radiateurs automobiles ; Fonte, ciselage ou usinage de bronzes au plomb ; Pose ou dépose de canalisations en plomb ; Démolition de bâtis anciens ; Décapage thermique ou par ponçage de vieilles peintures ou de peintures antirouille ; Pose et dépose d'ouvrages en plomb sur des toitures, terrasses ou balcons ; Utilisation de films ou de plaques de plomb pour l'isolation contre le bruit, les vibrations et/ou l'humidité ; Découpage au chalumeau de ferrailles peintes ; Pose et dépose de protecteur de câbles d'acier ou de lignes téléphoniques.
b : poterie, émaillage ; travail sur vitraux ; chasse, tir sportif, pêche ; fabrication de soldats de plomb, de modèles réduits ou d'objets décoratifs comportant des pièces en plomb ou revêtus d'un peinture au plomb ; fonte de plombs de chasse, de pêche, de plongée,... ; décapage de peintures de mobiliers anciens, véhicules, bateaux,...
c : remplacement des fenêtres ou des portes ; réfection des peintures ; décapage des peintures ; sablage des peintures ; ravalement de façade.
d : remplacement des fenêtres ou des portes ; réfection des peintures ; décapage des peintures.
e : pulvérulence ; écaillage; cloquage ; fissures ; faïençage ; traces de grattage ; lézardes.
f : traces de chocs ; micro-fissures.

Les 2 covariables relatives aux sites polluants (Basias) et aux sols pollués (Basol) ont été construites par $\sum_k 1/d_k$ où d_k est la distance au logement du k-ème site « Basias », respectivement « Basol », identifié à partir des bases du même nom [MEEDDTL & BRGM, 2011, MEEDDTL, 2011]. Le rayon retenu autour du logement a été de 2 km.

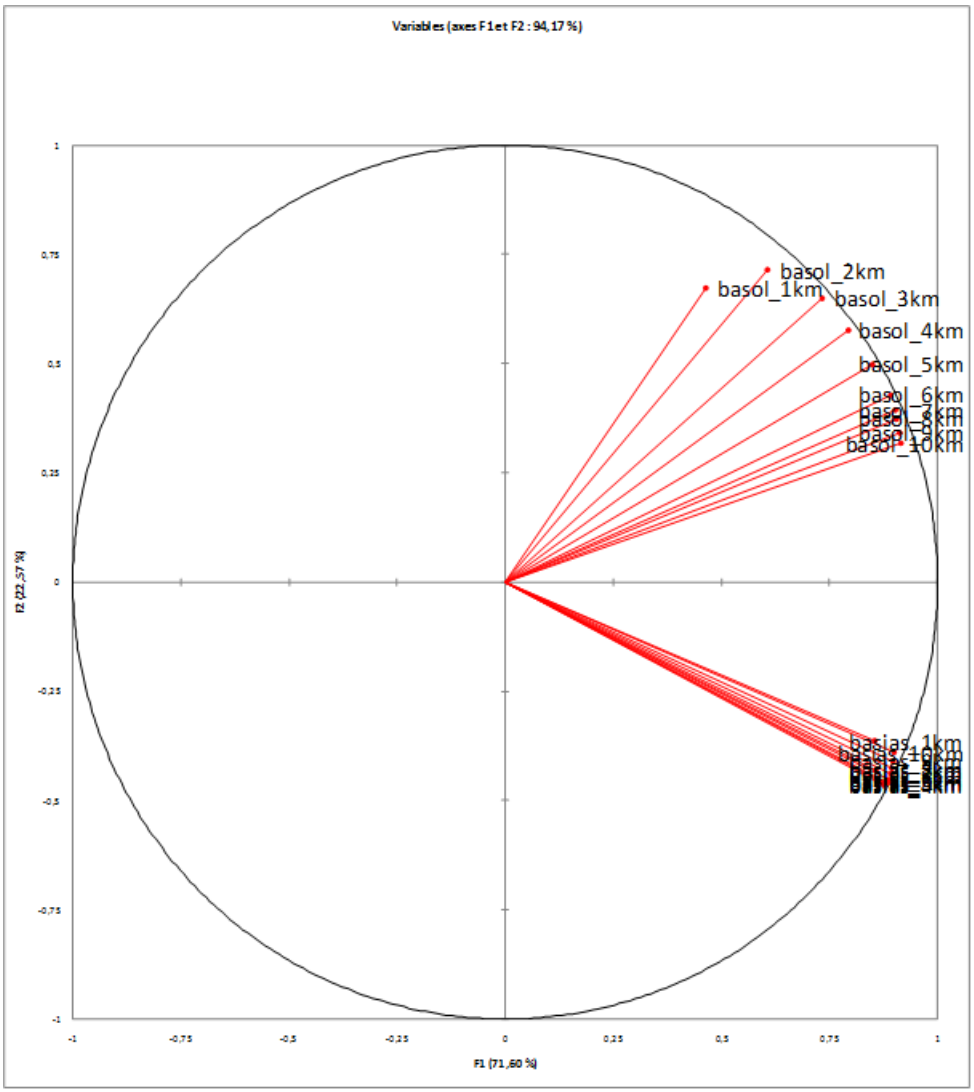
Les données ont été fournies au CSTB par l'INERIS (Institut National de l'Environnement industriel et des Risques) après avoir communiqué les coordonnées géodésiques des 484 logements dans un cadre contractuel. Les sites et sols pollués ont été obtenus jusqu'à 10 km autour de chaque logement. Ainsi 10 variables relatives à chacune des 2 bases ont été initialement construites, correspondant aux 10 rayons allant de 1 km à 10 km. Afin d'analyser les corrélations entre les 10 variables « Basias » et les 10 variables « Basol », dont les informations qu'elles indiquent ne sont pas forcément évidentes à différencier de prime abord, une analyse en composantes principales a été réalisée. Son résultat est montré par la figure c-dessous. L'ACP restitue plus de 94 % de l'inertie dont la première composante restitue plus de 71 %. Les 10 variables « Basias » donnent pratiquement la même information. Ce n'est pas le cas des 10 variables « Basol » : de « basol_1km » à « basol_5km » les variables sont de plus en plus corrélées avec la première composante ; à partir de « basol_6km » les variables donnent presque la même information. L'information était donc différentes entre ce que fournissaient les variables « Basol » et « Basias ». Les 2 variables « basol_2km » et « basias_2km » ont été retenues car d'une part il pouvait y avoir une incertitude de 1000 mètres sur l'identification des sites autour d'un logement ; les variables « basol_1km » et « basias_1km » pouvaient ne pas être alors idéales. D'autre part parmi les variables « Basol » restantes, « basol_2km » est la moins corrélées aux variables « Basias ».

Concernant « Bdrep », de même 10 variables ont été intialement construites ; elles indiquaient pratiquement la même information. La variable « bdrep_2km » a été retenu par homogénéité avec « Basol » et « Basias » et en faisant l'hypothèse que plus les sites émissifs sont proches du logement plus ils devraient voir un impact sur la contamination du logement. La variable « Bdrep » retenue a été construite de la manière suivante : $\sum_k f_k/d_k$ où f_k est l'émission en plomb (kg/an) et d_k la distance en km entre le site et le logement [MEEDDTL & INERIS, 2003].

La variable « Trafic routier » a été construite par n/d où n est le flux annuel de voiture de la route la plus proche du logement et d la distance en km entre la route et le logement ; les données utilisées sont celles de la base *European Open Street Map* [Geofabrik, 2008].

La conservation d'une seule variable par base de données (Basias, Basol, Bdrep) a permis d'être parcimonieux en termes de dimensionnalité (nombre de paramètres à estimer), et a donc participé à éviter le surajustement d'une part, et a participé à éviter la multicolinéarité d'autre part.

ACP des 10 variables Basias et des 10 variables Basol.



4 Prédicteurs des variables à imputer

Sur fond saumon les variables niveau pièce dont les suffixes 1 à 5 indique l'indexation de la pièce.
En police verte, les variables entrant dans le modèle d'analyse, en plomb acido-soluble (« AS » ou « as »), respectivement en plomb total (« TOT » ou « tot »).

Le préfixe « L » ou « NL » indique que la variable est sous forme transformée logarithmique.

ologit : régression logistique ordinale.

pmm : regression *« predictive mean matching »* ; la valeur prédite est parmi les valeurs observées.

intreg : régression par intervalle. Cette méthode ne concerne que les variables indiquant un niveau en plomb total, imputées après leur homologue respectif acido-soluble. La borne inférieure de l'intervalle est la valeur en plomb acido-soluble disponible ou préalablement imputée car, par définition la valeur en plomb total est supérieure ou égale à la valeur en plomb acido-soluble. La borne supérieure de l'intervalle est la valeur maximale observée pour la variable.

Variable à imputer	Méthode	Condition	LFUMEE2	LMesure_as_impute01	LMesure_as_impute02	LMesure_as_impute03	LMesure_as_impute04	LMesure_as_impute05	LPbA_impute_palier	LSol_PbAS_dur_freq_enfant1	LSol_PbAS_dur_freq_enfant2	LSol_PbAS_meuble_freq_enf1	LSumXrfUDPbSup1Degrad1	LSumXrfUDPbSup1Degrad2	LSumXrfUDPbSup1Degrad3	LSumXrfUDPbSup1Degrad4	LSumXrfUDPbSup1Degrad5	LSumXrfUDPbSup1Use1	LSumXrfUDPbSup1Use2
emplacement_logement3mod	ologit		×	×					×	×	×	×							
Lscore_km	pmm		×	×					×	×	×	×							
LSol_PbAS_meuble_freq_enf1	pmm	Imputé si la fréquentation de l'enfant est "souvent"	×	×															
NLSol_PbTOT_meu_freq_enf1	intreg	Imputé si la fréquentation de l'enfant est "souvent"	×									×							
LFUMEE2	pmm			×					×	×	×	×							
LSol_PbAS_dur_freq_enfant1	pmm	Imputé si la fréquentation de l'enfant est "souvent"	×	×															
NLSol_PbTOT_dur_fq_enfant1	intreg	Imputé si la fréquentation de l'enfant est "souvent"	×							×									
LSol_PbAS_dur_freq_enfant2	pmm	Imputé si la fréquentation de l'enfant est "tout le temps"	×	×															
NLSol_PbTOT_dur_fq_enfant2	intreg	Imputé si la fréquentation de l'enfant est "tout le temps"	×								×								
Lmax_xrf_palier	pmm	Imputé si présence de palier		×									×					×	
LPbA_impute_palier	pmm	Imputé si présence de palier	×	×															
NLPb_impute_palier	intreg	Imputé si présence de palier	×						×										
LMesure_as_impute01	pmm	Imputé si une pièce indexée en n°1 existe	×						×	×	×	×	×					×	
NLMesure_tot_impute01	intreg	Imputé si une pièce indexée en n°1 existe	×	×															
LMesure_as_impute02	pmm	Imputé si une pièce indexée en n°2 existe	×	×					×	×	×	×		×					×
NLMesure_tot_impute02	intreg	Imputé si une pièce indexée en n°2 existe	×		×														
LMesure_as_impute03	pmm	Imputé si une pièce indexée en n°3 existe	×	×	×				×	×	×	×			×				
NLMesure_tot_impute03	intreg	Imputé si une pièce indexée en n°3 existe	×			×													
LMesure_as_impute04	pmm	Imputé si une pièce indexée en n°4 existe	×	×	×	×			×	×	×	×				×			
NLMesure_tot_impute04	intreg	Imputé si une pièce indexée en n°4 existe	×				×												
LMesure_as_impute05	pmm	Imputé si une pièce indexée en n°5 existe	×	×	×	×	×		×	×	×	×					×		
NLMesure_tot_impute05	intreg	Imputé si une pièce indexée en n°5 existe	×					×											

Variable à imputer	LSumXrfUDPbSup1Use3	LSumXrfUDPbSup1Use4	LSumXrfUDPbSup1Use5	Lbasias_2km	Lbasol_2km	Lbdrep_2km	Lloisir_int_freq	Lmax_xrf_palier	Lmoyen_humide_frequence1	Lmoyen_humide_frequence2	Lmoyen_humide_frequence3	Lmoyen_humide_frequence4	Lmoyen_humide_frequence5	Lmoyen_sec_frequence1	Lmoyen_sec_frequence2	Lmoyen_sec_frequence3	Lmoyen_sec_frequence4	Lmoyen_sec_frequence5	Lprof_risque	Lscore_km	Lterrasse_garde_corps_XRF	NLMesure_tot_impute01	NLMesure_tot_impute02	NLMesure_tot_impute03	NLMesure_tot_impute04	NLMesure_tot_impute05	NLPb_impute_palier	NLSol_PbTOT_dur_fq_enfant1	NLSol_PbTOT_dur_fq_enfant2	NLSol_PbTOT_meu_freq_enf1	emplacement_logement3mod	hauteur_max1
emplacement_logement3mod								×												×												
Lscore_km				×	×	×		×																							×	
LSol_PbAS_meuble_freq_enf1				×		×														×							×			×	×	
NLSol_PbTOT_meu_freq_enf1																				×		×					×				×	
LFUMEE2								×												×											×	
LSol_PbAS_dur_freq_enfant1				×		×														×							×	×			×	
NLSol_PbTOT_dur_fq_enfant1																				×		×					×				×	
LSol_PbAS_dur_freq_enfant2				×																×							×		×		×	
NLSol_PbTOT_dur_fq_enfant2																				×		×					×				×	
Lmax_xrf_palier																					×						×	×	×	×	×	
LPbA_impute_palier				×		×		×												×							×				×	
NLPb_impute_palier								×												×		×									×	
LMesure_as_impute01				×	×	×	×	×	×					×					×	×	×	×									×	×
NLMesure_tot_impute01								×												×							×	×	×	×	×	
LMesure_as_impute02				×	×	×	×	×		×					×				×	×	×		×								×	
NLMesure_tot_impute02								×												×		×					×	×	×	×	×	
LMesure_as_impute03	×			×	×	×	×	×			×					×			×	×	×			×							×	
NLMesure_tot_impute03								×												×		×	×				×	×	×	×	×	
LMesure_as_impute04		×		×	×	×	×	×				×					×		×	×	×				×						×	
NLMesure_tot_impute04								×												×		×	×	×			×	×	×	×	×	
LMesure_as_impute05			×	×	×	×	×	×					×					×	×	×	×					×					×	
NLMesure_tot_impute05								×												×		×	×	×	×		×	×	×	×	×	

Variable à imputer	hauteur_max2	hauteur_max3	hauteur_max4	hauteur_max5	Environ2mod	REGION	absence_cour	absence_hall	age4mod	aucun_systeme	canal_int	chauffage	cspmer	cspper	demolition	frequentation_cour_hiver	frequentation_esc_ete	frequentation_hall_ete	grattage	humidite_condensations1	humidite_condensations2	humidite_condensations3	humidite_condensations4	humidite_condensations5	humidite_degats1	humidite_degats2	humidite_degats3	humidite_degats4	humidite_impression1	humidite_impression2	humidite_impression3	humidite_impression4
emplacement_logement3mod							×		×																							
Lscore_km					×										×																	
LSol_PbAS_meuble_freq_enf1					×										×	×	×		×													
NLSol_PbTOT_meu_freq_enf1																																
LFUMEE2									×	×			×	×																		
LSol_PbAS_dur_freq_enfant1					×	×									×	×	×	×	×													
NLSol_PbTOT_dur_fq_enfant1																																
LSol_PbAS_dur_freq_enfant2					×										×				×													
NLSol_PbTOT_dur_fq_enfant2																																
Lmax_xrf_palier					×			×	×	×	×																					
LPbA_impute_palier					×			×	×	×	×				×																	
NLPb_impute_palier																																
LMesure_as_impute01					×				×		×	×	×	×	×	×	×	×	×	×					×				×			
NLMesure_tot_impute01																																
LMesure_as_impute02	×				×				×		×	×	×	×	×	×	×	×	×		×					×				×		
NLMesure_tot_impute02																																
LMesure_as_impute03		×			×				×		×	×	×	×	×	×	×	×	×			×					×				×	
NLMesure_tot_impute03																																
LMesure_as_impute04			×		×				×		×	×	×	×	×	×	×	×	×				×					×				×
NLMesure_tot_impute04																																
LMesure_as_impute05				×	×				×		×	×	×	×	×	×	×	×	×					×								
NLMesure_tot_impute05																																

Variable à imputer	humidite_impression5	humidite_infiltrations1	humidite_infiltrations2	humidite_infiltrations3	humidite_infiltrations4	humidite_infiltrations5	humidite_rats1	humidite_rats2	humidite_rats3	humidite_rats4	humidite_taches1	humidite_taches2	humidite_taches3	humidite_taches4	humidite_taches5	idstratebis	mastic	montant_accepte	nettoyage_humide_palier	pays_naissance_mere	periode2mod	plafond	plancher	statut_hebergement	surface_ech2mod1	surface_ech2mod2	surface_ech2mod3	surface_ech2mod4	surface_ech2mod5	terrasse	travaux_exterieur	travaux_interieur
emplacement_logement3mod																×														×		
Lscore_km																×																
LSol_PbAS_meuble_freq_enf1																×	×				×											
NLSol_PbTOT_meu_freq_enf1																																
LFUMEE2																				×	×			×								
LSol_PbAS_dur_freq_enfant1																					×											
NLSol_PbTOT_dur_fq_enfant1																																
LSol_PbAS_dur_freq_enfant2																																
NLSol_PbTOT_dur_fq_enfant2																																
Lmax_xrf_palier																×															×	×
LPbA_impute_palier																×			×		×										×	×
NLPb_impute_palier																																
LMesure_as_impute01		×					×				×					×	×	×		×	×	×	×	×	×					×	×	×
NLMesure_tot_impute01																																
LMesure_as_impute02			×					×				×				×	×	×		×	×	×	×	×		×				×	×	×
NLMesure_tot_impute02																																
LMesure_as_impute03				×					×				×			×	×	×		×	×	×	×	×			×			×	×	×
NLMesure_tot_impute03																																
LMesure_as_impute04					×					×				×		×	×	×		×	×	×	×	×				×		×	×	×
NLMesure_tot_impute04																																
LMesure_as_impute05	×					×									×	×		×		×	×	×	×	×					×	×	×	×
NLMesure_tot_impute05																																

Variable à imputer	type_lgt	ventilo	vmc	vnat	nb_personne	nombre_pieces_investiguees	nombre_pieces_principales	superficie_piece1	superficie_piece2	superficie_piece3	superficie_piece4	superficie_piece5	taille_psu_ds_pop
emplacement_logement3mod							×						×
Lscore_km													×
LSol_PbAS_meuble_freq_enf1													×
NLSol_PbTOT_meu_freq_enf1													×
LFUMEE2	×		×	×	×		×						×
LSol_PbAS_dur_freq_enfant1													×
NLSol_PbTOT_dur_fq_enfant1													×
LSol_PbAS_dur_freq_enfant2													×
NLSol_PbTOT_dur_fq_enfant2													×
Lmax_xrf_palier			×	×									×
LPbA_impute_palier			×	×									×
NLPb_impute_palier													×
LMesure_as_impute01	×	×	×	×	×	×	×	×					×
NLMesure_tot_impute01													×
LMesure_as_impute02	×	×	×	×	×	×	×		×				×
NLMesure_tot_impute02													×
LMesure_as_impute03	×	×	×	×	×	×	×			×			×
NLMesure_tot_impute03													×
LMesure_as_impute04	×	×	×	×	×	×	×				×		×
NLMesure_tot_impute04													×
LMesure_as_impute05		×	×	×	×		×					×	×
NLMesure_tot_impute05													×

Prédicteur	Description
LFUMEE2	Tabagisme journalier (cf. annexe 2)
LMesure_as_impute01 à 05	Charge en Pb AS (cf. annexe 2)
LPbA_impute_palier	Charge en Pb-palier Pb AS (cf. annexe 2)
LSol_PbAS_dur_freq_enfant1	Charge en Pb-poussière extérieure Pb AS (cf. annexe 2)
LSol_PbAS_dur_freq_enfant2	Charge en Pb-poussière extérieure Pb AS (cf. annexe 2)
LSol_PbAS_meuble_freq_enf1	Concentration en Pb-sol extérieur Pb AS (cf. annexe 2)
LSumXrfUDPbSup1Degrad1 à 5	Somme XRF-détérioré (cf. annexe 2)
LSumXrfUDPbSup1Use1 à 5	Somme XRF-état d'usage (cf. annexe 2)
Lbasias_2km	Basias (cf. annexe 2)
Lbasol_2km	Basol (cf. annexe 2)
Lbdrep_2km	Bdrep (cf. annexe 2)
Lloisir_int_freq	Fréquence de loisir (cf. annexe 2)
Lmax_xrf_palier	Mesure XRF maximal du palier
Lmoyen_humide_frequence1 à 5	Fréq. Lavage humide-pièce (cf. annexe 2)
Lmoyen_sec_frequence1 à 5	Fréq. Lavage sec-pièce (cf. annexe 2)
Lprof_risque	Nombre d'activités à risque (cf. annexe 2)
Lscore_km	Trafic routier (cf. annexe 2)
Lterrasse_garde_corps_XRF	XRF garde-corps (cf. annexe 2)
NLMesure_tot_impute01 à 05	Charge en Pb TOT (cf. annexe 2)
NLPb_impute_palier	Charge en Pb-palier Pb TOT (cf. annexe 2)
NLSol_PbTOT_dur_fq_enfant1	Charge en Pb-poussière extérieure Pb TOT (cf. annexe 2)
NLSol_PbTOT_dur_fq_enfant2	Charge en Pb-poussière extérieure Pb TOT (cf. annexe 2)
NLSol_PbTOT_meu_freq_enf1	Concentration en Pb-sol extérieur Pb TOT (cf. annexe 2)
emplacement_logement3mod	Emplacement du logement (cf. annexe 2)
hauteur_max1 à 5	Hauteur maximale (cm) de la pièce
Environ2mod	Environnement urbain ou rural du logement
REGION	Région administrative où se trouve le logement
absence_cour	Absence d'une cour d'immeuble
absence_hall	Absence d'un hall d'immeuble
age4mod	Période de construction en 4 modalités
aucun_systeme	Si le logement ne possède aucun système de ventilation
canal_int	Si le logement possède des canalisations en plomb
chauffage	Type de chauffage en 3 modalités
cspmer	Catégorie socio-profressionnelle de la mère
cspper	Catégorie socio-profressionnelle du père
demolition	Démolition (cf. annexe 2)
frequentation_cour_hiver	Fréquence de fréquentation par l'enfant de la cour d'immeuble en période hivernale.
frequentation_esc_ete	Fréquence de fréquentation par l'enfant de la cage d'escalier de l'immeuble d'immeuble en période estivale.
frequentation_hall_ete	Fréquence de fréquentation par l'enfant du hall d'escalier de l'immeuble d'immeuble en période estivale.
grattage	Si l'enfant à tendance à gratter, mordiller ou sucer les revêtements des fenêtre, des portes ou des murs.
humidite_condensations1 à 5	Si la pièce a un problème de condensation persistante.
humidite_degats1 à 5	Si la pièce a un dégât des eaux.
humidite_impression1 à 5	Si une impression d'humidité se fait ressentir en entrant dans la pièce.
humidite_infiltrations1 à 5	Si la pièce a des infiltrations d'eau.
humidite_rats1 à 4	Si la pièce a des nuisibles.
humidite_taches1 à 5	Si la pièce a des tâches d'humidité.
idstratebis	Identifiant des strates des PSUs (hôpitaux).
mastic	Si l'enfant à tendance à gratter le mastic des fenêtres.
montant_accepte	Si le ménage a accepté d'indiquant le montant des ressources du ménages.
nettoyage_humide_palier	Lavage humide du palier (cf. annexe 2)
pays_naissance_mere	Le pays de naissance de la mère est la France ou Autre.
periode2mod	Saison (cf. annexe 2)
plafond	Le plafond menace-t-il de s'écrouler dans une ou plusieurs pièces ?
plancher	Le plancher menace-t-il de s'écrouler dans une ou plusieurs pièces ?
statut_hebergement	Si le ménage est propriétaire, locataire, hégergé ou a un autre statut d'occupation.
surface_ech2mod1 à 5	Endroit du prélèvement poussière (cf. annexe 2)
terrasse	Présence d'une terrasse /balcon / logia ?
travaux_exterieur	Travaux extérieurs (cf. annexe 2)
travaux_interieur	Travaux intérieurs (cf. annexe 2)
type_lgt	Logement individuel ou collectif
ventilo	Le système de ventilation est par ventilateurs mécaniques.
vmc	Le système de ventilation est par ventilation mécanique contrôlée.
vnat	Le système de ventilation est par ventilation naturelle.
nb_personne	Nombre de personnes habitant dans le logement.
nombre_pieces_investiguees	Nombre de pièces investiguées dans le logement lors de l'enquête.
nombre_pieces_principales	Nombre de pièces principales dans le logement.
superficie_piece1 à 5	Surface de la pièce (m²).
taille_psu_ds_pop	Somme des poids de sondage post-stratifiés des logements par PSU (hôpital).

5 Commande Stata V12 du modèle d'imputation

Note : les « i. » indiquent au logiciel une variable catégorielle.

mi impute chained

(**ologit,** include(i.absence_cour i.age4mod i.idstratebis i.terrasse nombre_pieces_principales) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05 NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier) augment) **emplacement_logement3mod**

(**pmm,** include(i.demolition i.idstratebis Lbasias_2km Lbasol_2km Lbdrep_2km i.Environ2mod) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05 NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **Lscore_km**

(**pmm** if freq_enfant2meuble == 1, include(i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete i.grattage i.idstratebis i.mastic i.periode2mod Lbasias_2km Lbdrep_2km i.Environ2mod) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05 LSol_PbAS_dur_freq_enfant1 LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier Lmax_xrf_palier NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **LSol_PbAS_meuble_freq_enf1**

(**intreg** if freq_enfant2meuble == 1, ll(ll_LSol_PbTOT_meu_fq_enf1) ul(ul_LSol_PbTOT_meu_fq_enf1) omit(LMesure_as_impute01 LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05 LSol_PbAS_dur_freq_enfant1 LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier Lmax_xrf_palier NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **NLSol_PbTOT_meu_freq_enf1**

(**pmm ,** include(i.age4mod i.aucun_systeme i.cspmer i.cspper i.pays_naissance_mere i.periode2mod i.statut_hebergement i.type_lgt i.vmc i.vnat nb_personne nombre_pieces_principales) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05 NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **LFUMEE2**

(**pmm** if freq_enfant2dur == 1, include(i.demolition
i.frequentation_cour_hiver i.frequentation_esc_ete i.frequentation_hall_ete
i.grattage i.periode2mod Lbasias_2km Lbdrep_2km i.Environ2mod i.REGION)
omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04
LMesure_as_impute05 LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant2
Lmax_xrf_palier LPbA_impute_palier NLMesure_tot_impute01 NLMesure_tot_impute02
NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05
NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier))
LSol_PbAS_dur_freq_enfant1

(**intreg** if freq_enfant2dur == 1, ll(ll_LSol_PbTOT_dur_fq_enf1)
ul(ul_LSol_PbTOT_dur_fq_enf1) omit(LMesure_as_impute01 LMesure_as_impute02
LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05
LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant2 Lmax_xrf_palier
LPbA_impute_palier NLMesure_tot_impute02 NLMesure_tot_impute03
NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **NLSol_PbTOT_dur_fq_enfant1**

(**pmm** if freq_enfant2dur == 2, include(i.demolition i.grattage Lbasias_2km
i.Environ2mod) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04
LMesure_as_impute05 LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1
LPbA_impute_palier Lmax_xrf_palier NLMesure_tot_impute01 NLMesure_tot_impute02
NLMesure_tot_impute03
NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant1 NLPb_impute_palier)) **LSol_PbAS_dur_freq_enfant2**

(**intreg** if freq_enfant2dur == 2, ll(ll_LSol_PbTOT_dur_fq_enf2)
ul(ul_LSol_PbTOT_dur_fq_enf2) omit(LMesure_as_impute01 LMesure_as_impute02
LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05
LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1 LPbA_impute_palier
Lmax_xrf_palier NLMesure_tot_impute02 NLMesure_tot_impute03
NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant1 NLPb_impute_palier)) **NLSol_PbTOT_dur_fq_enfant2**

(**pmm** if absence_palier==2, include(i.absence_hall i.age4mod i.aucun_systeme
i.canal_int i.idstratebis i.travaux_exterieur i.travaux_interieur
i.vmc i.vnat Lterrasse_garde_corps_XRF i.Environ2mod LSumXrfUDPbSup1Degrad1
LSumXrfUDPbSup1Use1) omit(LMesure_as_impute02 LMesure_as_impute03
LMesure_as_impute04 LMesure_as_impute05 LPbA_impute_palier
LSol_PbAS_meuble_freq_enf1 LFUMEE2 LSol_PbAS_dur_freq_enfant1
LSol_PbAS_dur_freq_enfant2 Lscore_km NLMesure_tot_impute01
NLMesure_tot_impute02 NLMesure_tot_impute03 NLMesure_tot_impute04
NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1
NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) **Lmax_xrf_palier**

(**pmm** if absence_palier==2, include(i. nettoyage_humide_palier i.absence_hall
i.age4mod i.aucun_systeme i.canal_int i.demolition i.idstratebis i.periode2mod
i.travaux_exterieur i.travaux_interieur i.vmc i.vnat Lbasias_2km Lbdrep_2km
i.Environ2mod) omit(LMesure_as_impute02 LMesure_as_impute03 LMesure_as_impute04

```
LMesure_as_impute05 LSol_PbAS_dur_freq_enfant1 LSol_PbAS_dur_freq_enfant2
LSol_PbAS_meuble_freq_enf1 NLMesure_tot_impute01 NLMesure_tot_impute02
NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05
NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1
NLSol_PbTOT_dur_fq_enfant2)) LPbA_impute_palier

    (intreg if absence_palier==2, ll(ll_LPb_impute_palier)
ul(ul_LPb_impute_palier) omit(LMesure_as_impute01 LMesure_as_impute02
LMesure_as_impute03 LMesure_as_impute04 LMesure_as_impute05
LSol_PbAS_dur_freq_enfant1 LSol_PbAS_dur_freq_enfant2
LSol_PbAS_meuble_freq_enf1 NLMesure_tot_impute02 NLMesure_tot_impute03
NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2)) NLPb_impute_palier

    (pmm if piece1 != ., include(i.age4mod i.canal_int i.chauffage i.cspmer
i.cspper i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete
i.frequentation_hall_ete i.grattage i.idstratebis i.mastic
i.montant_accepte i.pays_naissance_mere i.periode2mod i.plafond i.plancher
i.statut_hebergement i.terrasse i.travaux_exterieur i.travaux_interieur
i.type_lgt i.ventilo i.vmc i.vnat Lbasias_2km Lbasol_2km Lbdrep_2km
Lloisir_int_freq Lprof_risque Lterrasse_garde_corps_XRF nb_personne
nombre_pieces_investiguees nombre_pieces_principales i.Environ2mod
hauteur_max1 i.humidite_condensations1 i.humidite_degats1
i.humidite_impression1 i.humidite_infiltrations1 i.humidite_rats1
i.humidite_taches1 i.surface_ech2mod1 Lmoyen_humide_frequence1
Lmoyen_sec_frequence1 LSumXrfUDPbSup1Degrad1 LSumXrfUDPbSup1Use1
superficie_piece1) omit(LMesure_as_impute02 LMesure_as_impute03
LMesure_as_impute04 LMesure_as_impute05 NLMesure_tot_impute02
NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05
NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1
NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) LMesure_as_impute01

    (intreg if piece1 != ., ll(ll_LMesure_tot_impute1)
ul(ul_LMesure_tot_impute1) omit(LMesure_as_impute02
LMesure_as_impute03
LMesure_as_impute04 LMesure_as_impute05 NLMesure_tot_impute02
NLMesure_tot_impute03 NLMesure_tot_impute04 NLMesure_tot_impute05
LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1
LSol_PbAS_dur_freq_enfant2
LPbA_impute_palier)) NLMesure_tot_impute01

    (pmm if piece2 != ., include(i.age4mod i.canal_int i.chauffage i.cspmer
i.cspper i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete
i.frequentation_hall_ete i.grattage i.idstratebis i.mastic
i.montant_accepte i.pays_naissance_mere i.periode2mod i.plafond i.plancher
i.statut_hebergement i.terrasse i.travaux_exterieur i.travaux_interieur
i.type_lgt i.ventilo i.vmc i.vnat Lbasias_2km Lbasol_2km Lbdrep_2km
Lloisir_int_freq Lprof_risque Lterrasse_garde_corps_XRF nb_personne
nombre_pieces_investiguees nombre_pieces_principales i.Environ2mod
```

```
hauteur_max2 i.humidite_condensations2 i.humidite_degats2
i.humidite_impression2 i.humidite_infiltrations2 i.humidite_rats2
i.humidite_taches2 i.surface_ech2mod2 Lmoyen_humide_frequence2
Lmoyen_sec_frequence2 LSumXrfUDPbSup1Degrad2 LSumXrfUDPbSup1Use2
superficie_piece2) omit(LMesure_as_impute03 LMesure_as_impute04
LMesure_as_impute05 NLMesure_tot_impute01 NLMesure_tot_impute03
NLMesure_tot_impute04 NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier))
LMesure_as_impute02

    (intreg if piece2 != ., ll(ll_LMesure_tot_impute2) ul(ul_LMesure_tot_impute2)
omit(LMesure_as_impute01 LMesure_as_impute03 LMesure_as_impute04
LMesure_as_impute05 NLMesure_tot_impute03 NLMesure_tot_impute04
NLMesure_tot_impute05 LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1
LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier)) NLMesure_tot_impute02

    (pmm if piece3 != ., include(i.age4mod i.canal_int i.chauffage i.cspmer
i.cspper i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete
i.frequentation_hall_ete i.grattage i.idstratebis i.mastic
i.montant_accepte i.pays_naissance_mere i.periode2mod i.plafond i.plancher
i.statut_hebergement i.terrasse i.travaux_exterieur i.travaux_interieur
i.type_lgt i.ventilo i.vmc i.vnat Lbasias_2km Lbasol_2km Lbdrep_2km
Lloisir_int_freq Lprof_risque Lterrasse_garde_corps_XRF nb_personne
nombre_pieces_investiguees nombre_pieces_principales i.Environ2mod
hauteur_max3 i.humidite_condensations3 i.humidite_degats3
i.humidite_impression3 i.humidite_infiltrations3 i.humidite_rats3
i.humidite_taches3 i.surface_ech2mod3 Lmoyen_humide_frequence3
Lmoyen_sec_frequence3 LSumXrfUDPbSup1Degrad3 LSumXrfUDPbSup1Use3
superficie_piece3) omit(LMesure_as_impute04 LMesure_as_impute05
NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute04
NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1
NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) LMesure_as_impute03

    (intreg if piece3 != ., ll(ll_LMesure_tot_impute3) ul(ul_LMesure_tot_impute3)
omit(LMesure_as_impute01 LMesure_as_impute02 LMesure_as_impute04
LMesure_as_impute05 NLMesure_tot_impute04 NLMesure_tot_impute05
LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1
LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier) ) NLMesure_tot_impute03

    (pmm if piece4 != ., include(i.age4mod i.canal_int i.chauffage i.cspmer
i.cspper i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete
i.frequentation_hall_ete i.grattage i.idstratebis i.mastic
i.montant_accepte i.pays_naissance_mere i.periode2mod i.plafond i.plancher
i.statut_hebergement i.terrasse i.travaux_exterieur i.travaux_interieur
i.type_lgt i.ventilo i.vmc i.vnat Lbasias_2km Lbasol_2km Lbdrep_2km
Lloisir_int_freq Lprof_risque Lterrasse_garde_corps_XRF nb_personne
nombre_pieces_investiguees nombre_pieces_principales i.Environ2mod
hauteur_max4 i.humidite_condensations4 i.humidite_degats4
i.humidite_impression4 i.humidite_infiltrations4 i.humidite_rats4
```

```
i.humidite_taches4 i.surface_ech2mod4 Lmoyen_humide_frequence4
Lmoyen_sec_frequence4 LSumXrfUDPbSup1Degrad4 LSumXrfUDPbSup1Use4
superficie_piece4) omit( LMesure_as_impute05
NLMesure_tot_impute01 NLMesure_tot_impute02 NLMesure_tot_impute03
NLMesure_tot_impute05 NLSol_PbTOT_meu_freq_enf1 NLSol_PbTOT_dur_fq_enfant1
NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier)) LMesure_as_impute04

    (intreg if piece4 != ., ll(ll_LMesure_tot_impute4)
ul(ul_LMesure_tot_impute4) omit(LMesure_as_impute01 LMesure_as_impute02
LMesure_as_impute03 LMesure_as_impute05
NLMesure_tot_impute05 LSol_PbAS_meuble_freq_enf1 LSol_PbAS_dur_freq_enfant1
LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier)) NLMesure_tot_impute04
(pmm if piece5 != ., include(i.age4mod i.canal_int i.chauffage i.cspmer
i.cspper i.demolition i.frequentation_cour_hiver i.frequentation_esc_ete
i.frequentation_hall_ete i.grattage i.idstratebis i.montant_accepte
i.pays_naissance_mere i.periode2mod i.plafond i.plancher
i.statut_hebergement i.terrasse i.travaux_exterieur i.travaux_interieur
i.ventilo i.vmc i.vnat Lbasias_2km Lbasol_2km Lbdrep_2km Lloisir_int_freq
Lprof_risque Lterrasse_garde_corps_XRF nb_personne
nombre_pieces_principales i.Environ2mod hauteur_max5
i.humidite_condensations5 i.humidite_impression5 i.humidite_infiltrations5
i.humidite_taches5 i.surface_ech2mod5 Lmoyen_humide_frequence5
Lmoyen_sec_frequence5 LSumXrfUDPbSup1Degrad5 LSumXrfUDPbSup1Use5
superficie_piece5 ) omit(NLMesure_tot_impute01 NLMesure_tot_impute02
NLMesure_tot_impute03 NLMesure_tot_impute04 NLSol_PbTOT_meu_freq_enf1
NLSol_PbTOT_dur_fq_enfant1 NLSol_PbTOT_dur_fq_enfant2 NLPb_impute_palier))
LMesure_as_impute05

    (intreg if piece5 != ., ll(ll_LMesure_tot_impute5)
ul(ul_LMesure_tot_impute5) omit(LMesure_as_impute01 LMesure_as_impute02
LMesure_as_impute03 LMesure_as_impute04 LSol_PbAS_meuble_freq_enf1
LSol_PbAS_dur_freq_enfant1 LSol_PbAS_dur_freq_enfant2 LPbA_impute_palier)
) NLMesure_tot_impute05

= taille_psu_ds_pop, add(100) orderasis noisily
```

6 Commandes Stata V12 du modèle d'analyse

L'estimation du modèle à 2 niveaux à *« intercept »* aléatoire de l'application numérique en section 3 du chapitre 3 a été obtenue à partir de la commande Stata suivante :

```
    xtmixed Y liste_des_covariables [pw = w_ij] || code_enquete : ,
variance ml pweight(w_j) vce(cluster id_centre)
```

où `w_ij` indique la variable de pondération des pièces (ici tous les poids sont égaux à 1), `code_enquete` indique l'identifiant des logements, `variance` indique l'affichage

des estimations des variances σ_1^2 et σ_2^2 en variance plutôt qu'en écart-type, `ml` indique que l'on souhaite utiliser le maximum de vraisemblance mais en fait cet argument peut être enlevé car l'estimation par maximum de vraisemblance restreint n'est pas supportée en analyse de données d'enquête, `w_j` indique la variable des poids des logements et indique donc un des poids $\mathbf{w_j}$, et `vce(cluster id_centre)` permet d'obtenir des estimateurs robustes des variances en indiquant que les logements sont *« clusterisés »* dans les `id_centre` (hôpitaux) sans considérer ces `id_centre` comme des unités d'un niveau 3.

L'estimation du modèle à 3 niveaux à *« intercept »* aléatoire a été obtenue à partir de la commande Stata suivante :

```
xtmixed Y liste_des_covariables [pw = w_ij] || id_centre :,
pweight(w_k) || code_enquete : , variance ml pweight(w_j)
```

où `w_k` indique la variable de pondération des unités du niveau 3 (les hôpitaux).

L'estimation du modèle à 2 niveaux à *« intercept »* aléatoire à partir des données imputées a été obtenue à partir de la commande Stata suivante :

```
mi estimate : xtmixed Y liste_des_covariables [pw = w_ij] ||
code_enquete : , variance ml pweight(w_j) vce(cluster id_centre)
```

7 Contributions des sources calculées sur données imputées

Covariable *X*		**Percentile**						**% d'augmentation en *Y* quand *X* passe de son percentile d'ordre q_0 à celui d'ordre q_1 (q_0-q_1)**							
Plomb acido-soluble		**0.25**	**0.50**	**0.75**	**0.90**	**0.95**	**0.975**	**0.25-0.75**	**95% IC**	**0.50-0.90**	**95% IC**	**0.50-0.95**	**95% IC**	**0.50-0.975**	**95% IC**
Nombre d'activités à risque		0	0	0	2	4	5	0	-	6	(-13;29)	9	(-19;45)	10	(-20;51)
XRF garde-corps (mg/cm²)		0	0	0	0.2	2.6	5.1	0	-	6	(3;9)	52	(21;89)	80	(31;146)
Charge en Pb-poussière ext. (µg/m²)*	Y joue souvent	0	0	0	0	16	48	0	-	0	-	26	(-5;66)	37	(-6;101)
	Y joue tout le tps	0	0	0	0	0	82	0	-	0	-	0	-	44	(-14;141)
Concentration en Pb-sol ext. (mg/kg)*	Y joue souvent	0	0	14.5	42.4	82.9	243.7	32	(11;58)	47	(15;87)	57	(18;109)	75	(22;150)
	Y joue tout le tps	0	0	5	39.1	65.5	100.1	21	(6;38)	49	(13;95)	57	(15;114)	64	(17;131)
Charge en Pb-palier (µg/m²)*		0	0	6	30	44	148	128	(62;221)	313	(128;646)	383	(150;832)	687	(232;1764)
Trafic routier (millier/an /km)*		18	44	135	627	1210	2222	5	(-4;15)	6	(-6;20)	8	(-7;26)	10	(-8;31)
Démolition	Oui				-			28	(3;61)	28	(3;61)	28	(3;61)	28	(3;61)
Fréquence de loisir (nb. fois/an)		0	0	0	0	1	1	0	-	0	-	5	(-10;24)	5	(-10;24)
Travaux extérieurs	Oui				-			-17	(-37;10)	-17	(-37;10)	-17	(-37;10)	-17	(-37;10)
Travaux intérieurs	Oui				-			17	(-5;45)	17	(-5;45)	17	(-5;45)	17	(-5;45)
Basias (./km)		0	3.5	29.6	100.2	252.0	314.3	30	(3;64)	27	(3;57)	37	(4;80)	39	(4;86)
Basol (./km)		0	0	0	1.2	2.2	3.1	0	-	-11	(-32;18)	-15	(-43;27)	-18	(-50;34)
Bdrep (kg/an /km)		0	0	0	0.6	1.0	1.3	0	-	3	(-7;14)	4	(-11;22)	5	(-12;25)
Indoor smoking time (h/jour)*		0	0	0	1.5	3.5	5	0	-	33	(12;58)	59	(20;111)	74	(24;143)
Somme XRF-détériorié (mg/cm²)		0	0	0	0	0	0	0	-	0	-	0	-	0	-
Somme XRF-état d'usage (mg/cm²)		0	0	0	0	1.1	5.9	0	-	0	-	7	(2;13)	19	(4;36)
Plomb total															
Nombre d'activités à risque		0	0	0	2	4	5	0	-	11	(-10;38)	17	(-15;60)	19	(-16;68)
XRF garde-corps (mg/cm²)		0	0	0	0.2	2.6	5.1	0	-	5	(2;9)	48	(17;86)	74	(25;141)
Charge en Pb-poussière ext. (µg/m²)*	Y joue souvent	0	0	0	0	20	53	0	-	0	-	14	(-16;55)	19	(-21;78)
	Y joue tout le tps	0	0	0	0	0	91	0	-	0	-	0	-	27	(-23;111)
Concentration en Pb-sol ext. (mg/kg)*	Y joue souvent	0	0	23.0	64.6	97.6	267.5	27	(2;57)	36	(3;81)	40	(3;91)	51	(4;121)
	Y joue tout le tps	0	0	10.3	53.4	102.1	118.5	19	(-1;44)	34	(-2;82)	40	(-2;101)	42	(-2;105)
Charge en Pb-palier (µg/m²)*		0	0	7.7	41.1	51.7	172.5	122	(51;228)	298	(103;679)	332	(112;781)	570	(165;1593)
Trafic routier (millier/an /km)*		18	44	135	627	1210	2222	2	(-7;13)	3	(-9;17)	4	(-11;22)	5	(-13;26)
Démolition	Oui				-			31	(3;66)	31	(3;66)	31	(3;66)	31	(3;66)
Fréquence de loisir (nb. fois/an)		0	0	0	0	1	1	0	-	0	-	7	(-8;24)	7	(-8;24)
Travaux extérieurs	Oui				-			-12	(-35;18)	-12	(-35;18)	-12	(-35;18)	-12	(-35;18)
Travaux intérieurs	Oui				-			16	(-7;45)	16	(-7;45)	16	(-7;45)	16	(-7;45)
Basias (./km)		0	3.5	29.6	100.2	252.0	314.3	40	(9;79)	36	(8;70)	49	(11;99)	52	(12;106)
Basol (./km)		0	0	0	1.2	2.2	3.1	0	-	-12	(-34;16)	-17	(-45;25)	-21	(-52;31)
Bdrep (kg/an /km)		0	0	0	0.6	1.0	1.3	0	-	2	(-8;13)	3	(-13;21)	3	(-14;25)
Indoor smoking time (h/jour)*		0	0	0	1.5	3.5	5	0	-	39	(17;65)	71	(29;128)	90	(36;167)

Covariable *X*	**Percentile**						**% d'augmentation en *Y* quand *X* passe de son percentile d'ordre q_0 à celui d'ordre q_1 (q_0-q_1)**							
Plomb acido-soluble	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**	**0.975**	**0.25-0.75**	**95% IC**	**0.50-0.90**	**95% IC**	**0.50-0.95**	**95% IC**	**0.50-0.975**	**95% IC**
Somme XRF-détériorié (mg/cm²)	0	0	0	0	0	0	0	-	0	-	0	-	0	-
Somme XRF-état d'usage (mg/cm²)	0	0	0	0	1.1	5.9	0	-	0	-	7	(2;13)	20	(4;37)

Légende. *: covariable imputée, leurs quantiles ont été estimés à partir des données imputées; IC: Intervalle de confiance.

8 Contributions des sources calculées sur cas complets

Covariable *X*		**Percentile**						**% d'augmentation en *Y* quand *X* passe de son percentile d'ordre q_0 à celui d'ordre q_1 (q_0-q_1)**							
Plomb acido-soluble		**0.25**	**0.50**	**0.75**	**0.90**	**0.95**	**0.975**	**0.25-0.75**	**95% CI**	**0.50-0.90**	**95% CI**	**0.50-0.95**	**95% CI**	**0.50-0.975**	**95% CI**
Nombre d'activités à risque		0	0	0	2	4	5	0	-	10	(-11;35)	15	(-15;55)	16	(-17;63)
XRF garde-corps (mg/cm²)		0	0	0	0.2	2.6	5.1	0	-	7	(4;12)	70	(29;123)	111	(44;210)
Charge en Pb-poussière ext. (µg/m²)	Y joue souvent	0	0	0	0	11	21	0	-	0	-	35	(1;80)	45	(2;108)
	Y joue tout le tps	0	0	0	0	0	28	0	-	0	-	0	-	59	(1;152)
Concentration en Pb-sol ext. (mg/kg)	Y joue souvent	0	0	14.5	42.4	80.3	243.7	36	(12;65)	52	(17;99)	63	(20;123)	85	(25;173)
	Y joue tout le tps	0	0	5	39.1	65.5	100.1	23	(7;42)	54	(14;107)	63	(16;129)	71	(18;149)
Charge en Pb-palier (µg/m²)		0	0	4	29	39	78	67	(38;101)	194	(98;338)	223	(110;396)	300	(140;567)
Trafic routier (millier/an /km)		18	44	135	627	1210	2222	4	(-6;15)	6	(-7;21)	7	(-9;26)	8	(-11;32)
Démolition	Oui				-			37	(8;75)	37	(8;75)	37	(8;75)	37	(8;75)
Fréquence de loisir (nb. fois/an)		0	0	0	0	1	1	0	-	0	-	3	(-13;22)	3	(-13;22)
Travaux extérieurs	Oui				-			-9	(-32;23)	-9	(-32;23)	-9	(-32;23)	-9	(-32;23)
Travaux intérieurs	Oui				-			13	(-9;40)	13	(-9;40)	13	(-9;40)	13	(-9;40)
Basias (./km)		0	3.5	29.6	100.2	252.0	314.3	36	(6;74)	32	(5;66)	44	(7;93)	47	(7;100)
Basol (./km)		0	0	0	1.2	2.2	3.1	0	-	-7	(-35;32)	-10	(-47;50)	-13	(-53;64)
Bdrep (kg/an /km)		0	0	0	0.6	1.0	1.3	0	-	10	(-4;26)	15	(-6;42)	18	(-7;51)
Indoor smoking time (h/jour)		0	0	0	1.5	3.5	5	0	-	30	(10;53)	53	(17;101)	66	(20;129)
Somme XRF-détériorié (mg/cm²)		0	0	0	0	0	0	0	-	0	-	0	-	0	-
Somme XRF-état d'usage (mg/cm²)		0	0	0	0	1.1	5.9	0	-	0	-	7	(1;13)	19	(1;39)
Plomb total															
Nombre d'activités à risque		0	0	0	2	4	5	0	-	17	(-7;46)	26	(-10;74)	29	(-11;86)
XRF garde-corps (mg/cm²)		0	0	0	0.2	2.6	5.1	0	-	7	(3;11)	68	(27;123)	109	(41;210)
Charge en Pb-poussière ext. (µg/m²)	Y joue souvent	0	0	0	0	13	24	0	-	0	-	18	(-14;63)	23	(-17;82)
	Y joue tout le tps	0	0	0	0	0	32	0	-	0	-	0	-	39	(-11;115)
Concentration en Pb-sol ext. (mg/kg)	Y joue souvent	0	0	23	64.3	97.4	267.5	27	(1;59)	37	(2;84)	41	(2;95)	52	(2;126)
	Y joue tout le tps	0	0	10.3	53.4	102.1	118.5	19	(-2;45)	34	(-4;85)	40	(-4;104)	42	(-4;109)
Charge en Pb-palier (µg/m²)		0	0	5	36	46	82	58	(26;98)	151	(60;295)	167	(65;332)	208	(77;436)
Trafic routier (nb/an /km)		18	44	135	627	1210	2222	1	(-9;13)	2	(-11;17)	2	(-14;22)	3	(-16;27)
Démolition	Oui				-			37	(5;78)	37	(5;78)	37	(5;78)	37	(5;78)
Fréquence de loisir (nb. fois/an)		0	0	0	0	1	1	0	-	0	-	7	(-9;24)	7	(-9;24)
Travaux extérieurs	Oui				-			-7	(-31;28)	-7	(-31;28)	-7	(-31;28)	-7	(-31;28)
Travaux intérieurs	Oui				-			11	(-11;39)	11	(-11;39)	11	(-11;39)	11	(-11;39)
Basias (./km)		0	3.5	29.6	100.2	252.0	314.3	45	(11;89)	41	(10;79)	55	(14;112)	59	(14;121)
Basol (./km)		0	0	0	1.2	2.2	3.1	0	-	-8	(-37;35)	-11	(-49;54)	-14	(-56;70)
Bdrep (kg/year /km)		0	0	0	0.6	1.0	1.3	0	-	8	(-7;26)	13	(-10;42)	15	(-12;50)
Indoor smoking time (h/jour)		0	0	0	1.5	3.5	5	0	-	36	(16;61)	67	(27;119)	84	(33;154)

Covariable *X*	**Percentile**						**% d'augmentation en *Y* quand *X* passe de son percentile d'ordre q_0 à celui d'ordre q_1 (q_0-q_1)**							
Plomb acido-soluble	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**	**0.975**	**0.25-0.75**	**95% CI**	**0.50-0.90**	**95% CI**	**0.50-0.95**	**95% CI**	**0.50-0.975**	**95% CI**
Somme XRF-détériorié (mg/cm²)	0	0	0	0	0	0	0	-	0	-	0	-	0	-
Somme XRF-état d'usage (mg/cm²)	0	0	0	0	1.1	5.9	0	-	0	-	7	(0;15)	21	(1;44)

Légende. *: covariable imputée, leurs quantiles ont été estimés à partir des données imputées; IC: Intervalle de confiance.

9 Détails de simulation pour chaque covariable

<table>
<tr><th>Variable</th><th>Type</th><th>Modalité</th><th>Génération</th></tr>
<tr><td>Charge en Pb</td><td>num</td><td>-</td><td>En dernière, selon l’équation du modèle à 2 niveaux à « intercept » aléatoire.</td></tr>
<tr><td>Emplacement du logement</td><td>disc</td><td>- 1=Semi-enterré (réf.)
- 2=Rez de chaussée
- 3=En étage</td><td>Selon le type de logement* (collectif (COL) versus Individuel (IND)). Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>COL</td><td>0,0008</td><td>0,1335</td><td>0,8657</td></tr>
<tr><td>IND</td><td>0,0836</td><td>0,8242</td><td>0,0922</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Saison</td><td>disc</td><td>- 0=Automne/hiver (réf.)
- 1=Printemps/été</td><td>Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
<table>
<tr><td>0</td><td>1</td></tr>
<tr><td>0,5</td><td>0,5</td></tr>
</table></td></tr>
<tr><td>Lavage humide du palier</td><td>disc</td><td>- 0=Pas de palier (réf.)
- 1=Oui
- 2=Non</td><td>Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
<table>
<tr><td>0</td><td>1</td><td>2</td></tr>
<tr><td>0,1104</td><td>0,7523</td><td>0,1373</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Type de pièce</td><td>disc</td><td>- 1=Chambre de l’enfant ou d’un autre enfant
- 2=Entrée (réf.)
- 3=Salon
- 4=Cuisine
- 5=Salle de jeu</td><td>Pour chaque logement, le nombre de pièces à investiguées a été simulé en premier. Cela a été fait en fonction du nombre de pièces principales au sein du logement (séjour/salon ; chambre ; bureau ; cuisine si > 12m²).
Puis le type de pièce a été simulé selon le nombre de pièces investiguées estimés à partir des données de PH en prenant en compte les poids de sondage. Voir les détails après le tableau.</td></tr>
<tr><td>Fréq. Lavage humide-pièce</td><td>num</td><td>-</td><td>Log $\mathcal{N}(1.007;\ 0.606^2)$. Les 2 paramètres de la distribution Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Fréq. Lavage sec-pièce</td><td>num</td><td>-</td><td>Log $\mathcal{N}(1.422;\ 0.621^2)$. Les 2 paramètres de la distribution Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Endroit du prélèvement poussière</td><td>disc</td><td>- 0=Endroit de jeu préféré de l’enfant (réf.)
- 1=Au centre de la pièce</td><td>
<table>
<tr><td>0</td><td>1</td></tr>
<tr><td>0,4020</td><td>0,5980</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Nombre d’activités à risque</td><td>num</td><td>-</td><td>
<table>
<tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9 +</td></tr>
<tr><td>0,7870</td><td>0,1067</td><td>0,0347</td><td>0,0188</td><td>0,0266</td><td>0,0105</td><td>0,0072</td><td>0,0046</td><td>0,0017</td><td>0,0020</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
</table>

<table>
<tr><th>Variable</th><th>Type</th><th>Modalité</th><th>Génération</th></tr>
<tr><td>XRF garde-corps</td><td>num</td><td>-</td><td>Premièrement, la présence ou l’absence d’une terrasse a été simulée. Cela a été fait selon le type de logement (collectif (COL) versus Individuel (IND)). Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
<table><tr><td></td><td>Présence</td><td>Absence</td></tr><tr><td>COL</td><td>0,4301</td><td>0,5699</td></tr><tr><td>IND</td><td>0,3016</td><td>0,6984</td></tr></table>Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.
Si « Absence » est tiré, la valeur pour la covariable est 0. Sinon le tirage s’est fait selon la période de construction* du logement :
<table><tr><td>< 1949</td><td>$\text{Log}\,\mathcal{N}(0{,}154;\ 0{,}537^2)$</td></tr><tr><td>1949-1974</td><td>$\text{Log}\,\mathcal{N}(0{,}328;\ 0{,}714^2)$</td></tr><tr><td>1974-1993</td><td>$\text{Log}\,\mathcal{N}(0{,}036;\ 0{,}185^2)$</td></tr><tr><td>≥ 1994</td><td>$\text{Log}\,\mathcal{N}(0{,}005;\ 0{,}036^2)$</td></tr></table>Les 2 paramètres de la distribution Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Concentration en Pb-sol extérieur</td><td>num×disc</td><td>- 0=L’enfant ne joue pas à l’extérieur (réf.)
- 1=L’enfant y joue souvent
- 2=L’enfant y joue tout le temps</td><td>Premièrement, la fréquentation de l’aire de jeu extérieure de l’enfant a été simulée si la fréquentation n’était déjà pas égale à 1 ou 2 pour la poussière extérieure (voir covariable ci-dessous). Un seul type (sol ou poussière) est simulé. Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
<table><tr><td>0</td><td>1</td><td>2</td></tr><tr><td>0,1846</td><td>0,5094</td><td>0,3060</td></tr></table>Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.
Deuxièmement la valeur de la covariable a été simulée selon l’urbanisation* :
<table><tr><td>Urbain</td><td>$\text{Log}\,\mathcal{N}(3.274;\ 1.255^2)$</td></tr><tr><td>Rural</td><td>$\text{Log}\,\mathcal{N}(2.516;\ 0.647^2)$</td></tr></table>Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Charge en Pb-poussière extérieure</td><td>num×disc</td><td>- 0=L’enfant ne joue pas à l’extérieur (réf.)
- 1=L’enfant y joue souvent
- 2=L’enfant y joue tout le temps</td><td>Premièrement, la fréquentation de l’aire de jeu extérieure de l’enfant a été simulée. Le tirage des modalités a été fonction des probabilités affichées dans la table ci-dessous :
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.
Deuxièmement la valeur de la covariable a été simulée selon l’urbanisation* :
<table><tr><td>Urbain</td><td>$\text{Log}\,\mathcal{N}(3{,}741;\ 1{,}247^2)$</td></tr><tr><td>Rural</td><td>$\text{Log}\,\mathcal{N}(3{,}304;\ 0{,}894^2)$</td></tr></table>Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Charge en Pb-palier</td><td>num</td><td>-</td><td>Si le logement est individuel , la valeur de la covariable est 0. Sinon le tirage s’est fait selon la période de construction* du logement :</td></tr>
</table>

<table>
<tr><th>Variable</th><th>Type</th><th>Modalité</th><th>Génération</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><td>< 1949</td><td>Log $\mathcal{N}(2{,}570;\ 1{,}936^2)$</td></tr>
<tr><td>1949-1974</td><td>Log $\mathcal{N}(3{,}007;\ 1{,}111^2)$</td></tr>
<tr><td>1974-1993</td><td>Log $\mathcal{N}(3{,}011;\ 1{,}039^2)$</td></tr>
<tr><td>≥ 1994</td><td>Log $\mathcal{N}(1{,}567;\ 1{,}421^2)$</td></tr>
</table>
Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.
</td></tr>
<tr><td>Trafic routier</td><td>num</td><td>-</td><td>Selon la région administrative* et l’urbanisation*. Voir les détails après le tableau.</td></tr>
<tr><td>Démolition</td><td>disc</td><td>-1=Oui
-2=Non (réf.)</td><td>
<table>
<tr><td>1</td><td>2</td></tr>
<tr><td>0,1113</td><td>0,8887</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.
</td></tr>
<tr><td>Fréquence de loisir</td><td>num</td><td>-</td><td>Log $\mathcal{N}(0{,}076;\ 0{,}394^2)$
Les 2 paramètres de la distribution Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage.</td></tr>
<tr><td>Travaux extérieurs</td><td>disc</td><td>-1=Oui
-2=Non (réf.)</td><td>Selon le statut d’occupation* et du type de logement* (collectif (COL) versus Individuel (IND)). Voir les détails après le tableau.</td></tr>
<tr><td>Travaux intérieurs</td><td>disc</td><td>-1=Oui
-2=Non (réf.)</td><td>Selon le statut d’occupation* et la période de construction* du logement. Voir les détails après le tableau.</td></tr>
<tr><td>Sites polluants (Basias)</td><td>num</td><td>-</td><td>Selon la région administrative* et l’urbanisation*. Voir les détails après le tableau.</td></tr>
<tr><td>Sols pollués (Basol)</td><td>num</td><td>-</td><td>Selon la région administrative* et l’urbanisation*. Voir les détails après le tableau.</td></tr>
<tr><td>Pb dans l’air (Bdrep)</td><td>num</td><td>-</td><td>Selon la région administrative* et l’urbanisation*. Voir les détails après le tableau.</td></tr>
<tr><td>Tabagisme journalier</td><td>num</td><td>-</td><td>
<table>
<tr><td>0 (h)</td><td>0.5</td><td>1.5</td><td>3.5</td><td>5</td></tr>
<tr><td>0,7967</td><td>0,0610</td><td>0,0528</td><td>0,0588</td><td>0,0308</td></tr>
</table>
Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage.
</td></tr>
<tr><td>Somme XRF-détérioré</td><td>num</td><td>-</td><td>Selon la période de construction* du logement. La simulation doit prendre en compte la dépendance entre les pièces d’un même logement. Voir les détails après le tableau.</td></tr>
<tr><td>Somme XRF-état d’usage</td><td>num</td><td>-</td><td>Selon la période de construction* du logement. La simulation doit prendre en compte la dépendance entre les pièces d’un même logement. Voir les détails après le tableau.</td></tr>
</table>

Légende. * : variable dont la distribution est connue dans la population à partir du recensement INSEE 2006 ; num : numérique ; disc : discrète ; PH : Plomb-Habitat.

Détails

* indique ci-dessous les variables dont la distribution dans la population à partir du recensement INSEE 2006

Nombre de pièces et Type de pièces

Nombre de pièces investiguées	Nombre de pièces principale*							
	1	2	3	4	5	6	7	8
1	0,0351	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,5026	0,3044	0,1403	0,0161	0,0478	0,00	0,00	0,00
3	0,4624	0,1670	0,3094	0,3257	0,2822	0,1588	0,2330	0,2508
4	0,00	0,5050	0,4874	0,4037	0,3678	0,4992	0,6201	0,7492
5	0,00	0,0236	0,0629	0,2544	0,3022	0,3420	0,1469	0,00

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Type de pièce	Nombre de pièces investiguées			
	2	3	4	5
Salon	0,9826	0,4901	0,3306	0,2500
Entrée	0,00	0,0319	0,2091	0,2320
Cuisine	0,0174	0,3002	0,2818	0,2500
Salle de jeu	0,00	0,1234	0,0756	0,0927
Chambre d'un autre enfant	0,00	0,0543	0,1029	0,1753

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage, La modalité « 1 » du nombre de pièces investiguées n'apparait pas car dans cette situation le type de pièce est la chambre de l'enfant par défaut,

Trafic routier :

Region*	**Urbainisation***		**Region***	**Urbainisation***	
11	Rural	Log N(9,957; 0,798²)	52	Rural	Log N(9,187; 1,43²)
11	Urbain	Log N(12,062; 1,747²)	52	Urbain	Log N(10,067; 0,706²)
21	Rural	Log N(9,092; 0,709²)	53	Rural	Log N(9,83; 2,021²)
21	Urbain	Log N(10,367; 0,834²)	53	Urbain	Log N(10,805; 2,117²)
22	Rural	Log N(10,001; 1,048²)	54	Rural	Log N(9,591; 1,148²)
22	Urbain	Log N(11,279; 0,849²)	54	Urbain	Log N(9,932; 1,453²)
23	Rural	Log N(9,668; 1,15²)	72	Rural	Log N(8,37; 2,787²)
23	Urbain	Log N(11,093; 1,039²)	72	Urbain	Log N(10,345; 0,987²)
24	Rural	Log N(10,281; 1,256²)	73	Rural	Log N(9,867; 1,37²)
24	Urbain	Log N(9,833; 1,82²)	73	Urbain	Log N(12,294; 2,266²)

Region*	Urbainisation*		Region*	Urbainisation*	
25	Rural	Log N(9,852; 1,368²)	74	Rural	Log N(10,091; 1,8²)
25	Urbain	Log N(12,468; 1,133²)	74	Urbain	Log N(10,934; 1,518²)
26	Rural	Log N(9,847; 1,507²)	82	Rural	Log N(9,704; 1,756²)
26	Urbain	Log N(11,576; 1,496²)	82	Urbain	Log N(10,692; 1,927²)
31	Rural	Log N(11,05; 2,255²)	83	Rural	Log N(9,012; 0,895²)
31	Urbain	Log N(11,335; 0,958²)	83	Urbain	Log N(10,791; 1,285²)
41	Rural	Log N(11,116; 2,089²)	91	Rural	Log N(10,824; 0,419²)
41	Urbain	Log N(10,678; 1,492²)	91	Urbain	Log N(10,77; 0,414²)
42	Rural	Log N(10,458; 1,756²)	93	Rural	Log N(12,255; 0,463²)
42	Urbain	Log N(10,103; 0,683²)	93	Urbain	Log N(11,432; 2,165²)
43	Rural	Log N(10,519; 1,434²)	94	Rural	Log N(11,418; 0,923²)
43	Urbain	Log N(11,722; 1,372²)	94	Urbain	Log N(11,643; 1,868²)

Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage. Les codes Région sont disponibles en annexe 1.

Travaux extérieurs :

Statut d'occupation* = propriétaire

	Type de logement*	
Travaux ext,	COL	IND
1	0,00	0,0534
2	1	0,9466

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = hébergé gratuitement

	Type de logement*	
Travaux ext,	COL	IND
1	0,00	0,1731
2	1	0,8269

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = locataire HLM

	Type de logement*	
Travaux ext,	COL	IND
1	0,1576	0,0845
2	0,8424	0,9155

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = locataire privé

	Type de logement*	
Travaux ext,	COL	IND
1	0,1104	0,0747
2	0,8896	0,9253

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Travaux intérieurs :

Statut d'occupation* = propriétaire

Tvx, int,	Période de construction*			
	< 1949	1949-1974	1975-1993	≥ 1994
1	0,3282	0,2205	0,3297	0,3356
2	0,6718	0,7795	0,6703	0,6644

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = hébergé gratuitement

Tvx, int,	Période de construction *			
	< 1949	1949-1974	1975-1993	≥ 1994
1	0,0787	0,00	0,1947	0,00
2	0,9213	1,00	0,8053	1,00

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = locataire HLM

Tvx, int,	Période de construction *			
	< 1949	1949-1974	1975-1993	≥ 1994
1	0,0489	0,1121	0,2946	0,0693
2	0,9511	0,8879	0,7054	0,9307

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Statut d'occupation* = locataire privé

Tvx, int,	Période de construction *			
	< 1949	1949-1974	1975-1993	≥ 1994
1	0,2939	0,3146	0,6990	0,0564
2	0,7062	0,6854	0,3010	0,9436

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Sites polluants (Basias) :

Region*	**Urbainisation***		**Region***	**Urbainisation***	
11	Rural	Log N(0,215; 0,656²)	52	Rural	Log N(0,492; 0,514²)
11	Urbain	Log N(4,166; 1,315²)	52	Urbain	Log N(2,741; 0,865²)
21	Rural	Log N(0,159; 0,395²)	53	Rural	Log N(0,409; 0,831²)
21	Urbain	Log N(2,547; 0,594²)	53	Urbain	Log N(2,467; 1,343²)
22	Rural	Log N(0,038; 0,148²)	54	Rural	Log N(0,507; 0,859²)
22	Urbain	Log N(2,351; 1,353²)	54	Urbain	Log N(0,616; 0,64²)
23	Rural	Log N(0,317; 0,377²)	72	Rural	Log N(0,244; 0,553²)
23	Urbain	Log N(2,758; 1,851²)	72	Urbain	Log N(0,35; 0,62²)
24	Rural	Log N(1,074; 0,875²)	73	Rural	Log N(0,066; 0,279²)
24	Urbain	Log N(2,007; 1,292²)	73	Urbain	Log N(2,375; 1,28²)
25	Rural	Log N(0,002; 0,057²)	74	Rural	0
25	Urbain	Log N(1,212; 0,372²)	74	Urbain	Log N(3,506; 2,55²)
26	Rural	Log N(0,507; 0,499²)	82	Rural	0
26	Urbain	Log N(2,647; 0,646²)	82	Urbain	Log N(0,515; 1,026²)
31	Rural	0	83	Rural	0
31	Urbain	Log N(3,813; 1,215²)	83	Urbain	Log N(0,56; 0,714²)
41	Rural	Log N(0,231; 0,286²)	91	Rural	Log N(1,088; 0,353²)
41	Urbain	Log N(1,895; 0,623²)	91	Urbain	Log N(0,62; 1,093²)
42	Rural	Log N(0,159; 0,307²)	93	Rural	0
42	Urbain	Log N(1,44; 1,751²)	93	Urbain	Log N(1,24; 1,046²)
43	Rural	0	94	Rural	0

Region*	Urbainisation*		Region*	Urbainisation*	
43	Urbain	Log N(2,85; 1,261²)	94	Urbain	0

Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage. Les codes Région sont disponibles en annexe 1.

Sols pollués (Basol) :

Region*	Urbanisation*		Region*	Urbanisation*	
11	Rural	0	52	Rural	0
11	Urbain	Log N(0,44; 0,498²)	52	Urbain	0
21	Rural	0	53	Rural	0
21	Urbain	Log N(0,128; 0,2²)	53	Urbain	Log N(0,038; 0,128²)
22	Rural	0	54	Rural	0
22	Urbain	Log N(0,12; 0,286²)	54	Urbain	0
23	Rural	0	72	Rural	0
23	Urbain	Log N(0,6; 0,651²)	72	Urbain	0
24	Rural	0	73	Rural	0
24	Urbain	Log N(0,407; 0,646²)	73	Urbain	Log N(0,285; 0,344²)
25	Rural	0	74	Rural	0
25	Urbain	0	74	Urbain	0
26	Rural	0	82	Rural	0
26	Urbain	0	82	Urbain	Log N(0,08; 0,188²)
31	Rural	0	83	Rural	0
31	Urbain	Log N(0,511; 0,602²)	83	Urbain	0
41	Rural	0	91	Rural	0
41	Urbain	Log N(0,338; 0,612²)	91	Urbain	0
42	Rural	0	93	Rural	0
42	Urbain	0	93	Urbain	Log N(0,037; 0,17²)
43	Rural	0	94	Rural	0
43	Urbain	0	94	Urbain	0

Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage. Les codes Région sont disponibles en annexe 1.

Pb dans l'air (Bdrep) :

Region*	**Urbainisation***		**Region***	**Urbainisation***	
11	Rural	0	52	Rural	0
11	Urbain	Log N(0,107; 0,252²)	52	Urbain	Log N(0,205; 0,282²)
21	Rural	Log N(0,105; 0,383²)	53	Rural	0
21	Urbain	Log N(0,121; 0,518²)	53	Urbain	Log N(0,062; 0,149²)
22	Rural	0	54	Rural	0
22	Urbain	Log N(0,024; 0,1²)	54	Urbain	0
23	Rural	0	72	Rural	0
23	Urbain	Log N(0,618; 0,875²)	72	Urbain	Log N(0,029; 0,139²)
24	Rural	0	73	Rural	0
24	Urbain	0	73	Urbain	Log N(0,026; 0,167²)
25	Rural	0	74	Rural	0
25	Urbain	0	74	Urbain	Log N(0,584; 0,443²)
26	Rural	0	82	Rural	0
26	Urbain	Log N(0,045; 0,169²)	82	Urbain	Log N(0,128; 0,299²)
31	Rural	0	83	Rural	0
31	Urbain	Log N(0,23; 0,332²)	83	Urbain	0
41	Rural	0	91	Rural	0
41	Urbain	Log N(0,097; 0,214²)	91	Urbain	Log N(0,08; 0,296²)
42	Rural	0	93	Rural	0
42	Urbain	0	93	Urbain	Log N(0,014; 0,082²)
43	Rural	0	94	Rural	0
43	Urbain	Log N(0,035; 0,138²)	94	Urbain	0

Les 2 paramètres des distributions Log-Normale ont été estimés à partir des données de PH en prenant en compte les poids de sondage. Les codes Région sont disponibles en annexe 1.

Somme XRF-détérioré :

La simulation a été faite sur la distribution empirique avec pris en compte des poids de sondage sous forme d'histogramme car une distribution Log-Normale fournissait trop de faibles valeurs. Une valeur de la covariable a été simulée pour chaque pièce. La dépendance de cette covariable entre les pièces d'un même logement a été prise selon l'algorithme suivant :

- Pour la première pièce (i=1) d'un logement construit dans la période de construction B, une classe de l'histogramme est tiré selon les probabilités p dans les tables ci-dessous. Puis une valeur pour la covariable est tirée au sein de la classe de l'histogramme tirée à partir d'une loi uniforme.
- Pour la pièce $i \geq 2$:

- Si la pièce i -1 avait une valeur de la covariable hors de la première classe d'histogramme ([0; 1[mg/cm²), alors la pièce i a une probabilité q d'avoir aussi une valeur pour la covariable hors de la première classe. q était égale à 0,29 pour les logements construits jusqu'à 1974, et égale à 0 sinon. Cela correspond à $\mathbb{p}(A|B) = \mathbb{p}(A \cap B)/\mathbb{p}(B)$ où A est l'évènement « avoir au moins 2 pièces pour lesquelles la covariable a une valeur hors de la première classe » et B est l'évènement « avoir au moins une pièce pour laquelle la covariable a une valeur dans la première classe ».

On détermine alors si la pièce i a une valeur de la covariable hors de la première classe d'histogramme selon la probabilité q.

- Si la pièce i a une valeur de la covariable hors de la première classe d'histogramme, une classe est tirée selon la probabilité r montrés dans les tables ci-dessous. Puis une valeur pour la covariable est tirée au sein de la classe de l'histogramme tirée à partir d'une loi uniforme.
- Sinon la valeur pour la covariable pour la pièce i est tirée selon la même procédure que la pièce i=1.

- Table des probabilités p:

Logements construits < 1949 :

borne inf,	borne sup,	p
0	1	0,943
1	2	0,018
2	3	0,002
3	4	0,001
4	5	0,001
5	10	0,014
10	15	0,010
15	30	0,005
30	70	0,005

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1949 et 1974:

borne inf,	borne sup,	p
0	1	0,989
1	2	0,004
2	3	0,002
3	4	0,002
4	5	0,001
5	10	0,001
10	30	0,001

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1975 et 1993:

borne inf,	borne sup,	p
0	1	0,999
1	5	0,001

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits ≥ 1994: $p = 0$.

- Table des probabilités r:

Logements construits < 1949 :

borne inf,	borne sup,	r
1	2	0,316
2	3	0,041
3	4	0,018
4	5	0,013
5	10	0,250
10	15	0,182
15	30	0,095
30	70	0,086

Les probabilités ont été estimées à partir des

Logements construits entre 1949 et 1974:

borne inf,	borne sup,	r
1	2	0,358
2	3	0,174
3	4	0,165
4	5	0,09
5	10	0,120
10	30	0,093

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1975 et 1993: $r = 0$.

Logements construits ≥ 1994: $r = 0$.

données de PH en prenant en compte les poids de sondage,

Somme XRF-état d'usage :

La simulation est analogue à celle de la covariable « Somme XRF-détérioré ».

- Table des probabilités *p*:

Logements construits < 1949 :

borne inf,	borne sup,	*p*
0	1	0,858
1	2	0,020
2	3	0,007
3	4	0,014
4	5	0,024
5	10	0,030
10	15	0,019
15	30	0,015
30	70	0,012
70	150	0,002

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1949 et 1974:

borne inf,	borne sup,	*p*
0	1	0,941
1	2	0,015
2	3	0,005
3	4	0,002
4	5	0,011
5	10	0,005
10	30	0,013
30	50	0,007

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1975 et 1993:

borne inf,	borne sup,	*p*
0	1	0,998
1	5	0,002

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits ≥ 1994: $p = 0$.

- Table des probabilités *r*:

Logements construits < 1949 :

borne inf,	borne sup,	*r*
1	2	0,140
2	3	0,046
3	4	0,099
4	5	0,166
5	10	0,209
10	15	0,134
15	30	0,106
30	70	0,084
70	150	0,015

Les probabilités ont été estimées à partir des

Logements construits entre 1949 et 1974:

borne inf,	borne sup,	*r*
1	2	0,257
2	3	0,085
3	4	0,038
4	5	0,194
5	10	0,084
10	30	0,216
30	50	0,126

Les probabilités ont été estimées à partir des données de PH en prenant en compte les poids de sondage,

Logements construits entre 1975 et 1993: $r = 0$.

Logements construits ≥ 1994: $r = 0$.

données de PH en prenant en compte les
poids de sondage,

10 Valeur fixées des coefficients de régression

Valeur fixées des coefficients de régression à partir des estimations obtenues sur cas complets pour les scénarios $\mathbf{w_1}$ - $\mathbf{w_5}$.

Covariables	**Modalités**	**Coef.**	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$	$\mathbf{w_5}$	**Moy.**	**Valeur fixée**
« intercept »		β_0	-0,803	-0,181	-0,105	-0,430	-0,739	-0,451	-0,451
Emplacement du lgt	Semi-enterré	réf.							
	RDC	β_1	-0,451	-0,072	-0,050	-0,026	-0,327	-0,185	-0,185
	En étage	β_2	-0,391	-0,326	-0,106	-0,186	-0,298	-0,262	-0,262
Saison	Automne/hiver	réf.							
	Printemps/été	β_3	0,421	0,491	0,457	0,464	0,340	0,435	0,435
Lavage humide du palier	Pas de palier	réf.							
	Oui	β_4	1,867	1,589	1,233	1,485	1,684	1,572	1,572
	Non	β_5	0,209	0,226	0,123	0,140	0,151	0,170	0,170
Type de pièce	Chambre	β_6	-0,169	-0,338	-0,271	-0,291	-0,167	-0,247	-0,247
	Entrée	réf.							
	Salon	β_7	0,304	-0,043	0,075	0,166	0,350	0,170	0,170
	Cuisine	β_8	0,086	-0,098	-0,042	-0,058	0,077	-0,007	-0,007
	Salle de jeu	β_9	0,287	0,096	0,197	0,163	0,311	0,211	0,211
Fréq. Lavage humide-pièce		β_{10}	0,025	0,025	0,039	0,023	0,074	0,037	0,037
Fréq. Lavage sec-pièce		β_{11}	-0,018	-0,17	-0,111	-0,062	-0,001	-0,072	-0,072
Endroit du prélèvement	Endroit de jeu préféré	réf.							
	Centre de la pièce	β_{12}	0,011	0,076	0,066	0,016	-0,016	0,030	0,030
Log(Nombre d'activités à risque+1)		β_{13}	-0,002	-0,082	-0,081	-0,102	0,084	-0,037	0,084
Log(XRF garde-corps+1)		β_{14}	0,515	0,476	0,472	0,490	0,413	0,473	0,473
Log(Concentration en	Ne joue pas à l'ext.	réf.							
Pb-sol ext.+1)	Y joue souvent	β_{15}	0,098	0,092	0,089	0,106	0,112	0,099	0,099
	Y joue tout le tps	β_{16}	0,076	0,139	0,122	0,117	0,117	0,114	0,114
Log(Charge en	Ne joue pas à l'ext.	réf.							
Pb-poussière ext.+1)	Y joue souvent	β_{17}	0,085	0,100	0,115	0,104	0,121	0,105	0,105
	Y joue tout le tps	β_{18}	0,062	0,035	0,042	-0,016	0,138	0,052	0,110
Log(Charge en Pb-palier+1)		β_{19}	0,334	0,313	0,244	0,276	0,317	0,297	0,297
Log(Trafic routier)		β_{20}	0,010	-0,016	-0,013	-0,009	0,021	-0,001	0,020
Démolition	Non	réf.							
	Oui	β_{21}	0,342	0,511	0,401	0,422	0,317	0,399	0,399
Log(Fréquence de loisir+1)		β_{22}	0,119	0,152	0,217	0,139	0,042	0,134	0,134
Travaux extérieurs	Non	réf.							
	Oui	β_{23}	-0,123	0,005	-0,006	-0,130	-0,09	-0,069	-0,069
Travaux intérieurs	Non	réf.							
	Oui	β_{24}	0,231	0,336	0,33	0,284	0,123	0,261	0,261
Log(Basias+1)		β_{25}	0,115	0,116	0,105	0,196	0,090	0,125	0,125
Log(Basol+1)		β_{26}	0,022	-0,192	-0,017	-0,082	-0,095	-0,073	0,022
Log(Bdrep+1)		β_{27}	0,163	0,438	0,353	0,083	0,202	0,248	0,248
Log(Tabagisme journalier+1)		β_{28}	0,267	0,242	0,254	0,268	0,284	0,263	0,263
Log(Somme XRF -détériorié+1)		β_{29}	0,221	0,103	0,111	0,072	0,157	0,133	0,133
Log(Somme XRF -état d'usage+1)		β_{30}	0,023	-0,033	-0,007	0,004	0,088	0,015	0,015

11 Biais, variance et REQM par paramètre du modèle

Biais estimé de l'estimateur de chaque paramètre du modèle et pour chacun des 9 scénarios.

Coef.	**Vraie valeur**	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$	$\mathbf{w_5}$	$\mathbf{w_6}$	$\mathbf{w_7}$	$\mathbf{w_8}$	$\mathbf{w_9}$	**top3**$(\lvert\widehat{B}\rvert)$
β_0	-0,451	0,0211	-0,0145	0,0067	0,0360	0,0166	0,0171	0,0033	0,0149	-0,0324	$\mathbf{w_7}$, $\mathbf{w_3}$, $\mathbf{w_2}$
β_1	-0,185	-0,0056	0,0185	0,0094	0,0192	-0,0054	0,0156	-0,0039	-0,0059	0,0092	$\mathbf{w_7}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_2	-0,262	-0,0007	0,0122	0,0064	0,0163	0,0004	0,0122	-0,0011	0,0000	0,0099	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_3	0,435	0,0013	0,0059	0,0110	0,0075	0,0032	0,0010	0,0038	0,0037	0,0111	$\mathbf{w_6}$, $\mathbf{w_1}$, $\mathbf{w_5}$
β_4	1,572	-0,0176	-0,0201	-0,0274	-0,0352	-0,0154	-0,0232	0,0043	-0,0146	0,0041	$\mathbf{w_9}$, $\mathbf{w_7}$, $\mathbf{w_8}$
β_5	0,170	-0,0159	-0,0052	-0,0136	-0,0152	-0,0142	0,0004	-0,0089	-0,0129	-0,0052	$\mathbf{w_6}$, $\mathbf{w_2}$, $\mathbf{w_9}$
β_6	-0,247	-0,0017	0,0046	0,0022	0,0044	-0,0019	0,0062	-0,0043	-0,0019	0,0068	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$
β_7	0,170	-0,0022	0,0075	0,0053	0,0087	-0,0020	0,0133	-0,0050	-0,0020	0,0152	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_8	-0,007	0,0000	0,0060	0,0044	0,0078	-0,0001	0,0109	-0,0030	-0,0001	0,0076	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$
β_9	0,211	-0,0031	0,0077	0,0052	0,0095	-0,0024	0,0118	-0,0045	-0,0024	0,0132	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{10}	0,037	0,0043	0,0076	0,0040	0,0080	0,0029	0,0124	0,0020	0,0028	0,0044	$\mathbf{w_7}$, $\mathbf{w_8}$, $\mathbf{w_5}$
β_{11}	-0,072	0,0012	-0,0083	-0,0061	-0,0199	0,0010	-0,0213	-0,0020	0,0009	-0,0152	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{12}	0,030	-0,0003	-0,0019	-0,0011	0,0029	0,0006	0,0006	0,0035	0,0006	0,0019	$\mathbf{w_1}$, $\mathbf{w_6}$, $\mathbf{w_5}$
β_{13}	0,084	0,0032	0,0113	0,0113	0,0157	0,0034	0,0203	0,0017	0,0034	0,0087	$\mathbf{w_7}$, $\mathbf{w_1}$, $\mathbf{w_8}$
β_{14}	0,473	-0,0100	-0,0263	-0,0300	-0,0240	-0,0061	-0,0288	-0,0059	-0,0069	-0,0160	$\mathbf{w_7}$, $\mathbf{w_5}$, $\mathbf{w_8}$
β_{15}	0,105	-0,0007	0,0009	0,0015	0,0015	-0,0004	0,0016	0,0004	-0,0004	-0,0008	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_7}$
β_{16}	0,052	0,0024	0,0064	0,0059	0,0073	0,0033	0,0055	0,0056	0,0032	0,0051	$\mathbf{w_1}$, $\mathbf{w_8}$, $\mathbf{w_5}$
β_{17}	0,099	0,0008	0,0006	0,0004	-0,0003	0,0013	0,0010	0,0013	0,0013	0,0000	$\mathbf{w_9}$, $\mathbf{w_4}$, $\mathbf{w_3}$
β_{18}	0,114	-0,0008	-0,0030	-0,0019	-0,0031	-0,0002	-0,0038	0,0007	-0,0001	-0,0015	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_7}$
β_{19}	0,297	0,0004	-0,0039	-0,0034	-0,0026	0,0011	-0,0053	0,0049	0,0009	-0,0009	$\mathbf{w_1}$, $\mathbf{w_9}$, $\mathbf{w_8}$
β_{20}	0,020	-0,0008	0,0016	0,0004	-0,0013	-0,0008	0,0003	-0,0011	-0,0007	0,0011	$\mathbf{w_6}$, $\mathbf{w_3}$, $\mathbf{w_8}$
β_{21}	0,399	0,0019	0,0064	0,0065	0,0050	0,0029	0,0025	0,0026	0,0025	0,0020	$\mathbf{w_1}$, $\mathbf{w_9}$, $\mathbf{w_8}$
β_{22}	0,134	0,0048	-0,0026	0,0018	0,0012	0,0036	-0,0123	0,0083	0,0036	0,0333	$\mathbf{w_4}$, $\mathbf{w_3}$, $\mathbf{w_2}$
β_{23}	-0,069	-0,0003	0,0105	0,0048	0,0090	0,0009	0,0099	0,0064	0,0011	-0,0029	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$
β_{24}	0,261	-0,0003	-0,0033	0,0007	-0,0020	0,0003	-0,0083	-0,0005	0,0007	-0,0039	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_7}$
β_{25}	0,125	-0,0038	-0,0001	-0,0011	0,0012	-0,0046	0,0011	-0,0041	-0,0046	0,0060	$\mathbf{w_2}$, $\mathbf{w_6}$, $\mathbf{w_3}$
β_{26}	0,022	0,0074	-0,0074	-0,0019	-0,0099	0,0099	-0,0109	0,0105	0,0097	0,0126	$\mathbf{w_3}$, $\mathbf{w_2}$, $\mathbf{w_1}$
β_{27}	0,248	0,0363	0,0476	0,0545	0,0486	0,0354	0,0325	0,0480	0,0360	0,0790	$\mathbf{w_6}$, $\mathbf{w_5}$, $\mathbf{w_8}$
β_{28}	0,263	0,0021	0,0027	0,0045	0,0063	0,0013	0,0028	0,0020	0,0016	-0,0044	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_7}$
β_{29}	0,133	0,0002	0,0023	-0,0084	-0,0003	-0,0005	0,0076	-0,0011	-0,0007	0,0012	$\mathbf{w_1}$, $\mathbf{w_4}$, $\mathbf{w_5}$
β_{30}	0,015	0,0002	-0,0019	0,0014	-0,0015	0,0006	-0,0053	0,0010	0,0008	-0,0023	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$
σ_2^2	0,800	-0,0450	-0,1487	-0,1276	-0,2406	-0,0427	-0,3290	-0,1771	-0,0543	-0,4778	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$
σ_1^2	0,450	-0,0041	-0,0111	-0,0099	-0,0331	-0,0041	-0,0500	-0,0055	-0,0041	-0,0328	$\mathbf{w_1}$, $\mathbf{w_5}$, $\mathbf{w_8}$

Variance estimée de l'estimateur de chaque paramètre du modèle et pour chacun des 9 scénarios.

Coef.	**Vraie valeur**	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$	$\mathbf{w_5}$	$\mathbf{w_6}$	$\mathbf{w_7}$	$\mathbf{w_8}$	$\mathbf{w_9}$	**top3**$(\widehat{V})$
β_0	-0,451	0,2648	0,7222	0,6094	0,9019	0,2489	1,2474	0,2976	0,2502	1,1940	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_1	-0,185	0,0470	0,1226	0,1053	0,1489	0,0450	0,1944	0,0592	0,0452	0,1699	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_2	-0,262	0,0216	0,0654	0,0569	0,0834	0,0207	0,1089	0,0287	0,0209	0,0915	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_3	0,435	0,0074	0,0241	0,0219	0,0338	0,0072	0,0407	0,0100	0,0073	0,0335	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_4	1,572	0,1643	0,3945	0,3406	0,4688	0,1549	0,5998	0,1925	0,1561	0,5769	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_5	0,170	0,1266	0,2987	0,2599	0,3549	0,1220	0,4408	0,1563	0,1220	0,4532	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_6	-0,247	0,0054	0,0196	0,0144	0,0559	0,0051	0,0911	0,0063	0,0051	0,0566	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_7	0,170	0,0072	0,0257	0,0184	0,0751	0,0068	0,1265	0,0082	0,0068	0,0753	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_8	-0,007	0,0055	0,0219	0,0161	0,0616	0,0052	0,1008	0,0061	0,0052	0,0650	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_9	0,211	0,0057	0,0215	0,0164	0,0620	0,0055	0,0996	0,0070	0,0055	0,0626	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{10}	0,037	0,0028	0,0109	0,0086	0,0285	0,0027	0,0400	0,0033	0,0027	0,0266	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{11}	-0,072	0,0029	0,0100	0,0081	0,0256	0,0026	0,0370	0,0031	0,0026	0,0247	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{12}	0,030	0,0014	0,0043	0,0037	0,0120	0,0014	0,0185	0,0018	0,0014	0,0121	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{13}	0,084	0,0090	0,0246	0,0212	0,0299	0,0084	0,0382	0,0101	0,0084	0,0344	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{14}	0,473	0,0427	0,1164	0,0986	0,1538	0,0417	0,1976	0,0562	0,0421	0,1916	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{15}	0,105	0,0031	0,0075	0,0067	0,0093	0,0030	0,0121	0,0037	0,0030	0,0105	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{16}	0,052	0,0054	0,0126	0,0119	0,0147	0,0051	0,0172	0,0066	0,0051	0,0179	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{17}	0,099	0,0015	0,0034	0,0031	0,0043	0,0015	0,0058	0,0018	0,0015	0,0049	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{18}	0,114	0,0017	0,0044	0,0042	0,0058	0,0016	0,0074	0,0021	0,0016	0,0061	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{19}	0,297	0,0054	0,0149	0,0123	0,0223	0,0051	0,0258	0,0068	0,0051	0,0239	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{20}	0,020	0,0007	0,0023	0,0020	0,0031	0,0007	0,0041	0,0009	0,0007	0,0037	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{21}	0,399	0,0208	0,0573	0,0508	0,0766	0,0205	0,0983	0,0267	0,0205	0,0808	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{22}	0,134	0,0234	0,0647	0,0578	0,0834	0,0224	0,1092	0,0290	0,0226	0,0974	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{23}	-0,069	0,0325	0,1039	0,0929	0,1285	0,0309	0,1596	0,0405	0,0310	0,1373	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{24}	0,261	0,0097	0,0275	0,0253	0,0367	0,0092	0,0483	0,0114	0,0092	0,0362	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{25}	0,125	0,0014	0,0033	0,0027	0,0037	0,0014	0,0049	0,0019	0,0014	0,0066	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{26}	0,022	0,0377	0,0719	0,0654	0,0762	0,0374	0,0929	0,0574	0,0375	0,1182	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{27}	0,248	0,0608	0,1219	0,1119	0,1201	0,0582	0,1624	0,0759	0,0585	0,1889	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{28}	0,263	0,0091	0,0251	0,0207	0,0320	0,0089	0,0449	0,0116	0,0090	0,0308	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{29}	0,133	0,0066	0,0230	0,0196	0,0495	0,0063	0,0742	0,0074	0,0063	0,0501	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{30}	0,015	0,0021	0,0086	0,0076	0,0195	0,0020	0,0296	0,0025	0,0020	0,0198	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
σ_2^2	0,800	0,0035	0,0075	0,0068	0,0079	0,0034	0,0078	0,0041	0,0037	0,0038	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$
σ_1^2	0,450	0,0003	0,0010	0,0009	0,0023	0,0003	0,0033	0,0004	0,0003	0,0024	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$

RMSE estimé de l'estimateur de chaque paramètre du modèle et pour chacun des 9 scénarios.

Coef.	**Vraie valeur**	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$	$\mathbf{w_5}$	$\mathbf{w_6}$	$\mathbf{w_7}$	$\mathbf{w_8}$	$\mathbf{w_9}$	**top3**$(\widehat{REQM})$
β_0	-0,451	0,5150	0,8499	0,7807	0,9504	0,4991	1,1170	0,5456	0,5004	1,0932	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_1	-0,185	0,2170	0,3507	0,3246	0,3864	0,2123	0,4412	0,2433	0,2126	0,4123	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_2	-0,262	0,1469	0,2560	0,2385	0,2892	0,1439	0,3302	0,1693	0,1445	0,3026	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_3	0,435	0,0863	0,1554	0,1485	0,1839	0,0851	0,2017	0,0999	0,0852	0,1834	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_4	1,572	0,4057	0,6284	0,5843	0,6856	0,3939	0,7748	0,4388	0,3953	0,7595	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_5	0,170	0,3561	0,5466	0,5100	0,5959	0,3495	0,6639	0,3954	0,3495	0,6732	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_6	-0,247	0,0732	0,1403	0,1201	0,2364	0,0713	0,3018	0,0792	0,0715	0,2379	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_7	0,170	0,0851	0,1603	0,1357	0,2741	0,0826	0,3559	0,0909	0,0828	0,2748	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_8	-0,007	0,0739	0,1482	0,1269	0,2484	0,0719	0,3176	0,0779	0,0720	0,2551	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_9	0,211	0,0758	0,1469	0,1283	0,2491	0,0742	0,3158	0,0835	0,0743	0,2505	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{10}	0,037	0,0535	0,1048	0,0928	0,1690	0,0520	0,2004	0,0578	0,0521	0,1632	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{11}	-0,072	0,0534	0,1005	0,0900	0,1611	0,0510	0,1936	0,0561	0,0511	0,1578	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{12}	0,030	0,0373	0,0659	0,0609	0,1095	0,0369	0,1361	0,0431	0,0370	0,1101	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{13}	0,084	0,0950	0,1572	0,1461	0,1736	0,0917	0,1965	0,1005	0,0919	0,1857	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{14}	0,473	0,2068	0,3421	0,3155	0,3929	0,2042	0,4454	0,2371	0,2052	0,4380	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{15}	0,105	0,0554	0,0865	0,0819	0,0965	0,0547	0,1100	0,0611	0,0549	0,1025	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{16}	0,052	0,0732	0,1123	0,1094	0,1213	0,0713	0,1314	0,0817	0,0715	0,1340	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{17}	0,099	0,0391	0,0581	0,0561	0,0656	0,0384	0,0759	0,0429	0,0386	0,0699	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{18}	0,114	0,0414	0,0666	0,0647	0,0763	0,0404	0,0863	0,0459	0,0405	0,0779	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{19}	0,297	0,0734	0,1221	0,1109	0,1495	0,0711	0,1606	0,0825	0,0712	0,1545	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{20}	0,020	0,0264	0,0483	0,0450	0,0556	0,0257	0,0640	0,0297	0,0259	0,0609	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{21}	0,399	0,1444	0,2394	0,2254	0,2769	0,1431	0,3135	0,1636	0,1433	0,2842	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{22}	0,134	0,1532	0,2544	0,2404	0,2889	0,1498	0,3307	0,1706	0,1504	0,3139	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{23}	-0,069	0,1803	0,3224	0,3048	0,3586	0,1759	0,3996	0,2013	0,1760	0,3706	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{24}	0,261	0,0984	0,1659	0,1590	0,1915	0,0959	0,2199	0,1068	0,0959	0,1903	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{25}	0,125	0,0377	0,0573	0,0524	0,0609	0,0377	0,0702	0,0439	0,0376	0,0812	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
β_{26}	0,022	0,1943	0,2683	0,2558	0,2762	0,1937	0,3050	0,2398	0,1939	0,3440	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{27}	0,248	0,2493	0,3524	0,3389	0,3500	0,2437	0,4043	0,2797	0,2445	0,4417	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{28}	0,263	0,0954	0,1583	0,1441	0,1790	0,0944	0,2118	0,1077	0,0948	0,1756	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{29}	0,133	0,0810	0,1517	0,1403	0,2224	0,0794	0,2726	0,0860	0,0795	0,2238	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$
β_{30}	0,015	0,0458	0,0925	0,0872	0,1395	0,0451	0,1721	0,0501	0,0450	0,1407	$\mathbf{w_8}$, $\mathbf{w_5}$, $\mathbf{w_1}$
σ_2^2	0,800	0,0745	0,1722	0,1521	0,2565	0,0726	0,3407	0,1884	0,0818	0,4818	$\mathbf{w_5}$, $\mathbf{w_1}$, $\mathbf{w_8}$
σ_1^2	0,450	0,0181	0,0338	0,0313	0,0586	0,0178	0,0760	0,0198	0,0178	0,0586	$\mathbf{w_5}$, $\mathbf{w_8}$, $\mathbf{w_1}$

12 Articles publiés (Ne pas diffuser)

12.1 En tant que premier auteur

Historique de la réglementation relative à l'emploi de la céruse et des dérivés du plomb dans la peinture en France*

Jean-Paul Lucas[1,2]

[1] Université Paris-Est
Centre scientifique et technique du bâtiment (CSTB)
84, avenue Jean Jaurès
77447 Champs-sur-Marne cedex 02
France
<jean-paul.lucas@cstb.fr>

[2] Université de Nantes
Faculté de pharmacie
EA 4275 Biostatistique, recherche clinique et mesures subjectives en santé
1, rue Gaston Veil
BP 53508
44035 Nantes cedex 1
France

Tirés à part :
J.-P. Lucas

Article reçu le 14 mars 2011,
accepté le 3 mai 2011

Résumé. La réglementation de l'emploi des dérivés du plomb dans la peinture a régulièrement évolué depuis le XIXe siècle en France. Cette note technique propose de faire la synthèse des textes qui fondent cette réglementation. Les publications (lois, décrets et arrêtés) du *Journal Officiel de la République Française* (JORF) et différentes circulaires ont été analysées de la manière la plus exhaustive possible. Il en ressort que l'année 1949, communément considérée comme date charnière à partir de laquelle la céruse n'a plus été utilisée dans la peinture, ne peut se justifier comme telle d'après la réglementation : des mesures d'interdiction avaient en effet déjà été prises auparavant. Toutefois, la réglementation de l'emploi de dérivés du plomb dans la peinture n'émanait que du ministère du Travail jusqu'en 1993. De ce fait, les non-professionnels n'ont jamais été visés par les textes, ce qui rendait possible l'utilisation de la peinture contenant de la céruse dans les logements. La vente de ces peintures n'a été interdite qu'en 1993, alors que ces dernières étaient déjà considérées comme quasiment disparues du marché. Des chiffres précis manquent cependant sur ce point. Concernant le minium, autre composé très utilisé, la limitation de son usage sous la forme de pâte dans les travaux de peinture n'est plus en vigueur depuis 1988, mais le minium n'est vraisemblablement plus utilisé depuis le milieu des années 1990.

Mots clés : céruse ; législation ; peinture ; plomb.

Abstract

À history of French regulation of the use of white lead and other lead compounds in paints

Regulation of the use of lead compounds in paints in France has been modified regularly since the 19th century. The aim of this review is to analyse the regulations based on the most exhaustive possible compilation of laws and various types of regulatory enactments. Our reading of the regulations indicates that the generally accepted date of 1949 as the turning point when white lead was no longer used in paint is not justified. The Ministry of Health did not regulate this use until 1993; all earlier texts were issued by the Ministry of Labor and targeted occupational use of these paints. These regulations were thus never aimed at individuals painting their own homes, who were thus allowed to use lead in dwellings. The sale of white lead was not banned until 1993, although it might have disappeared from the market before then; sales data would be useful to clarify this question. Although the use of red lead in any form has been allowed in house paint since 1988, it probably disappeared in the mid-1990s due to labelling restrictions.

Key words: *lead; legislation; paint; white lead.*

* La description de la réglementation ici réalisée se base uniquement sur les textes réglementaires identifiés par l'auteur. L'auteur ne prétend pas avoir atteint l'exhaustivité en ce qui concerne cette identification.

Pour citer cet article : Lucas JP. Historique de la réglementation relative à l'emploi de la céruse et des dérivés du plomb dans la peinture en France. *Environ Risque Sante* 2011 ; 10 : 316-22. doi : 10.1684/ers.2011.0475

doi: 10.1684/ers.2011.0475

Environmental Research 116 (2012) 58–65

Contents lists available at SciVerse ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Lead contamination in French children's homes and environment ☆

Jean-Paul Lucas [a,b,*], Barbara Le Bot [c,d], Philippe Glorennec [c,d], Anne Etchevers [d,e], Philippe Bretin [e], Francis Douay [f,g], Véronique Sébille [b], Lise Bellanger [h], Corinne Mandin [a]

[a] Université Paris Est, CSTB—Centre Scientifique et Technique du Bâtiment, ESE/Santé, 84 avenue Jean Jaurès, Champs-sur-Marne, 77447 Marne-la-Vallée cedex 2, France
[b] Université de Nantes—EA 4275 Biostatistique, Recherche Clinique et Mesures Subjectives en Santé, 1 rue Gaston Veil BP 53508, 44035 Nantes Cedex 1, France
[c] EHESP—Sorbonne Paris Cité—Avenue du Professeur Léon-Bernard, CS 74312, 35043 Rennes cedex, France
[d] INSERM U1085—IRSET, 35 043 Rennes, France
[e] InVS, 12 rue du Val d'Osne, 94415 Saint-Maurice cedex, France
[f] Université Lille Nord de France, Lille, France
[g] Groupe ISA, Equipe Sols et Environnement, Laboratoire Génie Civil et géoEnvironnement (LGCgE) Lille Nord de France EA 4515, 48 boulevard Vauban, 59046 Lille Cedex, France
[h] UMR CNRS 6629 Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière BP 92208F, 44322 Nantes Cedex, France

ARTICLE INFO

Article history:
Received 8 November 2011
Received in revised form
22 February 2012
Accepted 12 April 2012
Available online 1 May 2012

Keywords:
Lead
Housing survey
Lead paint
Dust lead
Soil lead
Water lead

ABSTRACT

Lead in homes is a well-known source of childhood lead exposure, which is still of concern due to the health effects of low lead doses.

This study aims to describe lead contamination in the homes of children aged 6 months to 6 years in France (without overseas).

Between October 2008 and August 2009, 484 housing units were investigated. Lead in tap water and total and leachable lead levels from floor dust, outdoor soils and paint chips were measured. X-ray fluorescence measurements were carried out on non-metallic and metallic substrates. Nationwide results are provided.

The indoor floor dust lead (PbD) geometric mean (GM) was 8.8 μg/m^2 (0.8 μg/ft^2) and 6.8 μg/m^2 (0.6 μg/ft^2) for total and leachable lead respectively; 0.21% of homes had an indoor PbD loading above 430.5 μg/m^2 (40 μg/ft^2). The outdoor play area concentration GM was 33.5 mg/kg and 21.7 mg/kg in total and leachable lead respectively; 1.4% of concentrations were higher than or equal to 400 mg/kg. Outdoor floor PbD GM was 44.4 μg/m^2 (4.1 μg/ft^2) that was approximately 3.2 times higher than the GM of indoor PbD. Lead-based paint (LBP) was present in 25% of dwellings, LBP on only non-metallic substrates was present in 19% of homes and on metallic substrates in 10% of dwellings. The GM of lead concentrations in tap water was below 1 μg/L; 58% of concentrations were lower than 1 μg/L and 2.9% were higher than or equal to 10 μg/L. The age cut-off for homes with lead would be 1974 for paint and 1993 for indoor floor dust.

This study provides, for the first time, a look at the state of lead contamination to which children are exposed in French housing. Moreover, it provides policy makers an estimate of the number of French dwellings sheltering children where abatement should be conducted.

☆ *Funding*: This study was funded by the French Health Department (DGS) and the French Housing Department (DHUP).

* Corresponding author at: CSTB, ESE/Santé, 84 avenue Jean Jaurès, Champs-sur-Marne, 77447 Marne-la-Vallée Cedex 2, France. Fax: +33 1 64 68 88 23.
E-mail address: jean-paul.lucas@cstb.fr (J.-P. Lucas).

0013-9351/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.envres.2012.04.005

Science of the Total Environment 470–471 (2014) 768–779

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Source contributions of lead in residential floor dust and within-home variability of dust lead loading

Jean-Paul Lucas [a,b,*], Lise Bellanger [c], Yann Le Strat [d], Alain Le Tertre [d], Philippe Glorennec [e,f], Barbara Le Bot [e,f], Anne Etchevers [f], Corinne Mandin [a], Véronique Sébille [b]

[a] Université Paris Est, CSTB — Centre Scientifique et Technique du Bâtiment, 84 avenue Jean Jaurès, 77447 Marne-la-Vallée Cedex 2, France
[b] Université de Nantes, EA 4275 Biostatistique, Pharmacoépidémiologie et Mesures Subjectives en Santé, 1 rue Gaston Veil BP 53508, 44035 Nantes Cedex 1, France
[c] UMR CNRS 6629 Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière BP 92208, F-44322 Nantes Cedex, France
[d] InVS, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex, France
[e] EHESP Rennes, Sorbonne Paris Cité, Avenue du Professeur Léon-Bernard, CS 74312, 35043 Rennes Cedex, France
[f] INSERM UMR 1085 Institut de Recherche sur la Santé, l'Environnement et le Travail, Rennes, France

HIGHLIGHTS

- We estimated the contribution of lead sources to residential floor dust contamination.
- Dust lead from the landing of an apartment is the major contributor.
- Track-in of the exterior soil contaminates common area dust and interior dust.
- Exterior railings, smoking inside, demolitions, polluting sites are also contributors.
- Interior lead-based paint is no longer a contributor except for non-renovated homes.

ARTICLE INFO

Article history:
Received 9 July 2013
Received in revised form 7 October 2013
Accepted 9 October 2013
Available online xxxx

Keywords:
Lead
House dust
Exposure
Multilevel modeling

ABSTRACT

Evidence of the impact of exposure to low levels of lead on children's health is increasing. Residential floor dust is the assumed origin of lead exposure by young children. In this study, we estimate the contribution of different lead sources to household interior floor dust contamination. We also estimate the within-home variability of interior floor dust lead loadings. A multilevel model was developed based on data collected in a French survey in 2008–2009 (484 housing units, 1834 rooms). Missing data were handled by multiple imputation using chained equations. The intra-home correlation between interior floor Log dust lead loadings was approximately 0.6. Dust lead from the landing of an apartment, mostly originating outside the building, was the major contributor to interior floor dust lead. Secondary contributors included the lead-based paint on exterior railings, track-in of the exterior soil of the children's play area into the dwelling, smoking inside the home, demolition of nearby old buildings and sites of pollution in the vicinity. Interior lead-based paint contaminated interior floor dust only in old and non-renovated dwellings. To reduce interior floor dust lead levels in the general population of dwellings, common areas should be maintained, and track-in from the outside should be limited as much as possible.

* Corresponding author at: CSTB, Direction Santé Confort/Expologie OQAI, 84 avenue Jean Jaurès, Champs-sur-Marne, 77447 Marne-la-Vallée Cedex 2, France. Tel.: +33 1 64 68 88 39; fax: +33 1 64 68 88 23.
E-mail address: jean.paul.lucas@free.fr (J.-P. Lucas).

0048-9697/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2013.10.028

Journal of Applied Statistics, 2013
http://dx.doi.org/10.1080/02664763.2013.847404

Taylor & Francis
Taylor & Francis Group

Multilevel modelling of survey data: impact of the two-level weights used in the pseudolikelihood

Jean-Paul Lucas[a,b*], Véronique Sébille[b], Alain Le Tertre[c], Yann Le Strat[c] and Lise Bellanger[d]

[a]*Scientific and Technical Building Centre (CSTB), Paris Est University, Marne-la-Vallée, France;* [b]*EA4275-Sphere, University of Nantes, Nantes, France;* [c]*French Institute for Public Health Surveillance (InVS), Saint-Maurice, France;* [d]*UMR CNRS 6629 Laboratory of Mathematics Jean Leray, University of Nantes, Nantes, France*

(*Received 25 March 2013; accepted 18 September 2013*)

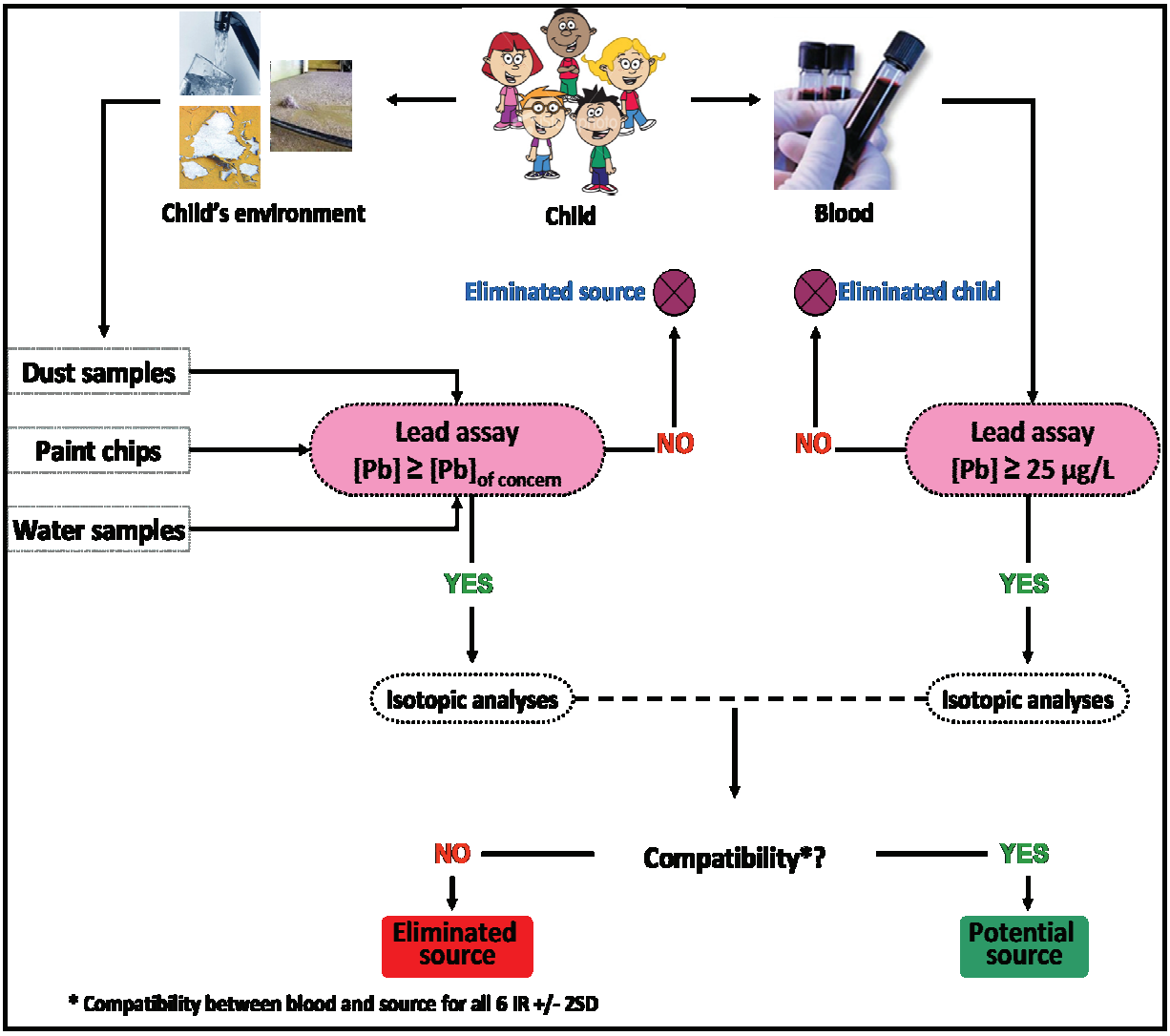
Approaches that use the pseudolikelihood to perform multilevel modelling on survey data have been presented in the literature. To avoid biased estimates due to unequal selection probabilities, conditional weights can be introduced at each level. Less-biased estimators can also be obtained in a two-level linear model if the level-1 weights are scaled. In this paper, we studied several level-2 weights that can be introduced into the pseudolikelihood when the sampling design and the hierarchical structure of the multilevel model do not match. Two-level and three-level models were studied. The present work was motivated by a study that aims to estimate the contributions of lead sources to polluting the interior floor dust of the rooms within dwellings. We performed a simulation study using the real data collected from a French survey to achieve our objective. We conclude that it is preferable to use unweighted analyses or, at the most, to use conditional level-2 weights in a two-level or a three-level model. We state some warnings and make some recommendations.

Keywords: lead exposure data; level-2 weights; multilevel model; pseudolikelihood; public database; survey data

*Corresponding author. Email: jean.paul.lucas@free.fr

12.2 En tant que co-auteur

ENVIRONMENTAL HEALTH



Glorennec P, Peyr C, Poupon J, Oulhote Y, Le Bot B: **Identifying Sources of Lead Exposure for Children, with Lead Concentrations and Isotope Ratios.** *Journal of Occupational and Environmental Hygiene* 2010, **7:** 253-260.

Identification of sources of lead exposure in French children by lead isotope analysis: a cross-sectional study

Oulhote *et al.*

Oulhote et al. Environmental Health 2011, 10:75
http://www.ehjournal.net/content/10/1/75

ENVIRONMENTAL HEALTH

RESEARCH Open Access

Identification of sources of lead exposure in French children by lead isotope analysis: a cross-sectional study

Youssef Oulhote[1,2,3*], Barbara Le Bot[1,2], Joel Poupon[4], Jean-Paul Lucas[5,6], Corinne Mandin[5], Anne Etchevers[7], Denis Zmirou-Navier[1,2,3,8] and Philippe Glorennec[1,2,9]

Abstract

Background: The amount of lead in the environment has decreased significantly in recent years, and so did exposure. However, there is no known safe exposure level and, therefore, the exposure of children to lead, although low, remains a major public health issue. With the lower levels of exposure, it is becoming more difficult to identify lead sources and new approaches may be required for preventive action. This study assessed the usefulness of lead isotope ratios for identifying sources of lead using data from a nationwide sample of French children aged from six months to six years with blood lead levels ≥25 μg/L.

Methods: Blood samples were taken from 125 children, representing about 600,000 French children; environmental samples were taken from their homes and personal information was collected. Lead isotope ratios were determined using quadrupole ICP-MS (inductively coupled plasma - mass spectrometry) and the isotopic signatures of potential sources of exposure were matched with those of blood in order to identify the most likely sources.

Results: In addition to the interpretation of lead concentrations, lead isotope ratios were potentially of use for 57% of children aged from six months to six years with blood lead level ≥ 25 μg/L (7% of overall children in France, about 332,000 children), with at least one potential source of lead and sufficiently well discriminated lead isotope ratios. Lead isotope ratios revealed a single suspected source of exposure for 32% of the subjects and were able to eliminate at least one unlikely source of exposure for 30% of the children.

Conclusions: In France, lead isotope ratios could provide valuable additional information in about a third of routine environmental investigations.

Texte non disponible

[1]EHESP - School of Public Health, Sorbonne Paris Cité, 35043 Rennes, France
Full list of author information is available at the end of the article

International Journal of Hygiene and Environmental Health 216 (2013) 743–750

Contents lists available at ScienceDirect

International Journal of Hygiene and Environmental Health

journal homepage: www.elsevier.com/locate/ijheh

Implications of different residential lead standards on children's blood lead levels in France: Predictions based on a national cross-sectional survey

Youssef Oulhote[a,b,c], Alain Le Tertre[d], Anne Etchevers[b,d], Barbara Le Bot[a,b], Jean-Paul Lucas[e,f], Corinne Mandin[e,b], Yann Le Strat[d], Bruce Lanphear[g,h], Philippe Glorennec[a,b,*]

[a] EHESP, Rennes, Sorbonne Paris Cité, France
[b] INSERM UMR1085 IRSET – Institut de Recherches sur la santé l'environnement et le travail, Rennes, France
[c] INSERM U954, Nancy University Medical School, Vandoeuvre Les Nancy, France
[d] InVS, French Institute for Public Health Surveillance, Saint Maurice, France
[e] Paris Est University – CSTB – Scientific and Technical Building Centre, Marne la Vallée, France
[f] EA 4275 Biostatistics, Clinical Research and Subjective Measures in Health, Nantes University, France
[g] Department of Pediatrics, Division of General and Community Pediatrics, Cincinnati Children's Hospital, Medical Center, Cincinnati, OH, USA
[h] Simon Fraser University, Vancouver, British Columbia, Canada

ARTICLE INFO

Article history:
Received 27 July 2012
Received in revised form 18 January 2013
Accepted 16 February 2013

Keywords:
Lead
Environmental exposure
Lead poisoning
Dust
Water
Soil

ABSTRACT

Despite the dramatic reductions in children's blood lead levels (BLLs), there is considerable evidence that low-level lead exposure is associated with intellectual deficits and behavioral problems, without apparent threshold. There are limited data, however, about the contribution of residential sources of lead to contemporary children's blood lead levels. The aim of this study is to calculate the contributions of residential sources of lead to assess the potential impact of setting new standards for lead levels in residential dust, soil and water. We enrolled 484 French children aged from 6 months to 6 years, and collected data on social, housing and individual characteristics. Lead concentrations in blood and environmental samples (water, soils, and dusts) were measured using inductively coupled plasma mass spectrometry. Data were analyzed using a multivariate generalized additive model accounting for the sampling design and the sampling weights. We found that exceedingly low concentrations of lead in dust, soil and water were significant predictors of children's BLLs, after adjustment for potential confounding variables. Lead-contaminated floor dust was the main source of lead in blood. BLLs (GM: 14 μg/L) increased by 65%, 13%, 25%, and 5% when lead content in floor dust, loose soil, hard soil and water increased from their 25th percentile to their 95th percentile, respectively. We also observed that the steepest increase in BLLs occurred at the lowest levels of lead-contaminated floor dust, which indicates that lead contamination should be kept as low as possible. Impact of different possible standards on children's BLLs was also tabulated and indicated that unless standards are set low, they will only benefit a small proportion of children who have the highest exposures.

1438-4639/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijheh.2013.02.007

13 Autres articles relatifs à l'enquête Plomb-Habitat (Ne pas diffuser)

Environment International 45 (2012) 129–134

Contents lists available at SciVerse ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

French children's exposure to metals via ingestion of indoor dust, outdoor playground dust and soil: Contamination data

Philippe Glorennec [a,b,*], Jean-Paul Lucas [c,d], Corinne Mandin [c], Barbara Le Bot [a,b]

[a] EHESP - School of Public Health, Sorbonne Paris Cité, Rennes, France
[b] Inserm U1085 - IRSET, Rennes, France
[c] Paris Est University - CSTB - Scientific and Technical Building Centre, Marne la Vallée cedex 2, France
[d] Nantes University EA 4275, Nantes, France

ARTICLE INFO

Article history:
Received 16 December 2011
Accepted 18 April 2012
Available online xxxx

Keywords:
Metal
Soil
Dust
Bioavailability
Exposure

ABSTRACT

In addition to dietary exposure, children are exposed to metals via ingestion of soils and indoor dust, contaminated by natural or anthropogenic outdoor and indoor sources. The objective of this nationwide study was to assess metal contamination of soils and dust which young French children are exposed to. A sample of 484 children (6 months to 6 years) was constituted in order to obtain representative results for young French children. In each home indoor settled dust was sampled by a wipe in up to five rooms. Outdoor playgrounds were sampled with a soil sample ring ($n = 315$) or with a wipe in case of hard surfaces ($n = 53$). As, Cd, Cr, Cu, Mn, Pb, Sb, Sr, and V were measured because of their potential health concern due to soil and dust ingestion. The samples were digested with hydrochloric acid, and afterwards aqua regia in order to determine both leachable and total metal concentrations and loadings by mass spectrometry with a quadrupole ICP-MS. In indoor settled dust most (total) loadings were below the Limit of Quantification (LOQ), except for Pb and Sr, whose median loadings were respectively 9 and 10 μg/m². The 95th percentile of loadings were 2 μg/m² for As, <0.8 for Cd, 18 for Cr, 49 for Cu, <64 for Mn, 63 for Pb, 2 for Sb, 56 for Sr, and <8 for V. Median/95th percentile of loadings in settled dust on outdoor playgrounds were 2/16, <0.8/1.3, 17/53, 49/330, 99/424, 32/393, 2/13, 86/661 and 10/37 μg/m² for As, Cd, Cr, Cu, Mn, Pb, Sb, Sr, and V respectively. In outdoor playground soil median/95th percentile of concentrations (μg/g) were 8/26, <0.65/1, 25/52, <26/53,391/956, 27/254, 0.7/4, 54/295, 23/57 for As, Cd, Cr, Cu, Mn, Pb, Sb, Sr, and V respectively. These results are comparable with those observed in other countries. Because of their representative nature, we can assess children's exposures to these metals via soil and dust and the associated risks in urban and rural environments. Ratios of leachable/total concentrations and loadings, calculated on >LOQ measurements, differed among metals. To a lesser extent, they were also affected by type of matrix, with (except for Cd) a greater leachability of dust (especially indoor) compared to soils.

* Corresponding author at: EHESP-School of Public Health, 2 avenue du Pr. L. Bernard 35043 Rennes Cédex, France. Tel. +33 299 022 680; fax: +33 299 022 675.
E-mail address: philippe.glorennec@ehesp.fr (P. Glorennec).

0160-4120/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envint.2012.04.010

L'article suivant a été soumis dans le journal *Science of the Total Environment* le 26 juillet 2013 et est au 7 octobre 2013 en cours de révision.

The exposure of children in France to metals via ingestion of tap water: contamination and exposure data from a nationwide survey.

Barbara le Bot[1, 2*], Jean-Paul Lucas[3, 4], Françoise Lacroix[1, 2], Philippe Glorennec [1, 2]

[1] EHESP, School of Public Health, Rennes, Sorbonne Paris Cité – Avenue du Professeur Léon-Bernard, CS 74312, 35043 Rennes cedex, France.
[2] Inserm, U1085-Institut de Recherche sur la Santé, l'Environnement et le Travail, Rennes, France.
[3] Paris Est University - CSTB - Scientific and Technical Building Centre, Marne la Vallée cedex 2, France
[4] Nantes University EA 4275, Nantes, France.

(*) **Corresponding author.**
E-mail address: barbara.lebot@ehesp.fr (Barbara Le Bot)
Phone: +332 99 02 29 24
Fax: +332 99 02 29 29

Abstract

29 inorganic compounds (Al, As, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cu, Fe, Gd, K, Mg, Mn, Mo, Na, Nd, Ni, Pb, Sb, Se, Sr, Tl, U, V and Zn) were measured in the tap water at the homes of children aged 6 months to 6 years in metropolitan France in 2008-2009. This nationwide housing survey (n=484) was nested in a national lead biomonitoring cross-sectional survey. Parents were asked whether or not their children consume tap water. Sampling design and sampling weights were taken into account to estimate element concentrations in tap water supplied to the homes of 3,581,991 children aged 6 months to 6 years. Median and 95th percentiles of concentrations in tap water were in µg/L: Al: <10, 48.3, As: 0.2, 2.1; B: <100, 100; Ba : 30.7, 149.4; Ca: 85 000,121,700 ; Cd:<0.5, <0.5;Ce: <0.5, <0.5; Co:<0.5, 0.8; Cr : <5,<5 ; Cu : 70,720; K: 2210, 6740; Fe:<20, 46; Mn: <5, <5; Mo: <0.5, 1.5; Na: 14500, 66800; Ni : <2,10.2;Mg: 6500, 21200; Pb :<1,5.4 ; Sb: <0.5, <0.5; Se: <1,6.7; Sr: 256.9, 1004; Tl: <0.5,<0.5; U: <0.5, 2.4;V:<1,1;Zn:53,208. Comparison of our results for harmful elements with regulatory monitoring showed less non–compliance for Pb, Ni, Cu, Sb and Ba, although it is important to note that the sample representativeness and sample methods are not the same. Of the 2,977,123 young children drinking tap water in France, some were drinking water having concentrations above the 2011 World Health Organization drinking-water quality guidelines: respectively 498 (CI95%:0-1,484) over 700 µg/L of Ba; 121,581 (CI95%:7091-236,070) over 50 mg/L of Na; 2044 (CI95%: 0-6,132) over 70 µg/L of Ni, and 78,466 (17,171-139,761) over 10 µg/L of Pb. Being representative, this tap water contamination data can be used for overall exposure assessment, in conjunction with diet and environmental (dust and soil) exposure data.

Keywords: inorganic, drinking-water, metal, metalloid, environmental exposure.

Résumé Général

Ce travail s'inscrit dans le domaine de la contamination des logements par le plomb. Il a consisté à estimer la prévalence des logements possédant un risque plomb et à identifier les déterminants de la contamination des poussières intérieures déposées au sol. Le travail s'est basé sur les données collectées dans 484 logements et 1834 pièces lors de la réalisation de l'enquête Plomb-Habitat (2008-2009). L'échantillon a été obtenu par un plan de sondage complexe à partir duquel un poids de sondage a été associé à chaque logement.

La démarche s'est organisée en 4 temps. Elle a consisté tout d'abord à valider les données avant de réaliser les analyses statistiques. Puis elle a consisté dans un second temps à estimer les niveaux en plomb dans l'eau du robinet, dans les revêtements intérieurs, dans les aires de jeu extérieures des enfants ainsi que dans la poussière intérieure déposée au sol. Dans un troisième temps les contributions des sources potentielles pouvant contaminer cette poussière ont été estimées. Ceci a été réalisé par un modèle multi-niveaux sur données d'enquête, sans pondération ; le niveau 1 était constitué des pièces et le niveau 2 constitué des logements. Dans un quatrième temps, une simulation Monte Carlo a été réalisée afin de s'assurer de la pertinence du modèle ajusté à l'étape précédente.

Ce travail a permis pour la première fois de réaliser un état de la contamination par le plomb dans les logements français. Il montre que concernant les niveaux en plomb dans l'eau du robinet, les poussières intérieures au sol, la peinture au plomb et les sols des aires de jeu extérieures, respectivement 2,5 %, 0,21 %, 24,5 % et 1,4 % des logements sont à risque dans la mesure où le seuil de référence respectif est dépassé. Ce travail met en évidence que les poussières du palier d'appartement sont le principal contributeur dans la contamination des poussières intérieures. Enfin ce travail a permis de montrer sur des données réelles d'enquête qu'une pondération non adéquate au niveau 2 d'un modèle multi-niveaux pouvait induire des estimateurs biaisés.

Il serait intéressant à partir des résultats obtenus de définir le concept de logement à risque plomb et de proposer un diagnostic technique permettant de mettre aisément en évidence le risque plomb global d'un logement. Il serait en outre intéressant d'étudier plus précisément les voies de contamination entre les différents compartiments environnementaux et de généraliser les résultats concernant la pondération à utiliser dans une modélisation multi-niveaux sur données d'enquête.

Abstract

This study belongs to the field of the lead contamination in housing. It aimed to estimate the prevalence of dwellings with lead hazards and to identify the determinants of lead contamination of interior floor dust. Data from the Plomb-Habitat survey (2008-2009) were used ; 1834 rooms within 484 housing units were investigated. The sample was drawn from a complex sampling design that provided a sampling weight for each dwelling.
Our approach had 4 steps. First, data were checked before to carry out the statistical analyses. Second, lead levels in tap water, in interior coatings, in outdoor children's play areas and also in interior floor dust were estimated. Third, the contribution of numerous sources that may contaminate the dust were jointly assessed. An unweighted multilevel model on survey data was used to estimate the source contributions ; rooms were the level-1 units and dwellings were used as level-2 units. Fourth, a Monte Carlo simulation study was carried out to make sure that our adjusted model was accurate.
We provide for the first time a look at the state of lead contamination in French housing. We showed that about lead levels in tap water, in interior floor dust, in interior lead-based paint and for outdoor play areas, lead hazard occurs in approximately 2,5 %, 0,21 %, 24,5 % and 1,4 % units respectively. Moreover we found that floor dust of the landing of an apartment was the main contributor in lead contamination of interior floor dust. Finally, based on real survey data, we showed that a not suitable weighting for level-2 in a multilevel model leads to biased estimators.
It would be interesting now to define what are homes with global lead hazards and to propose a protocole enables us to bring out easily the global lead hazard of a dwelling. Moreover it would be useful to study more precisely the pathway of lead contamination in housing between the different environmental medias and to generalize findings about the weighting at level-2 in a multilevel model on survey data.

Table des figures

Liste des tableaux

Bibliographie

[AFNOR, 2006] AFNOR (2006). *Soil quality - Pretreatment of samples for physico-chemical analysis. NF ISO 11464.* Technical report, Association Française de Normalisation, La Plaine Saint-Denis.

[AFNOR, 2008a] AFNOR (2008a). *Lead diagnosis — Chemical analysis of paints for determining the acido-soluble fraction of lead. NF X 46-031.* Technical report, Association Française de Normalisation, La Plaine Saint-Denis.

[AFNOR, 2008b] AFNOR (2008b). *Lead diagnosis — Methodology for measuring lead in floor dust. NF X 46-032.* Technical report, Association Française de Normalisation, La Plaine Saint-Denis.

[AFNOR, 2008c] AFNOR (2008c). *Lead diagnosis — Protocol for establishing the risk of exposure to lead. NF X 46-030.* Technical report, Association Française de Normalisation, La Plaine Saint-Denis.

[Ardilly, 2006] Ardilly, P. (2006). *Les Techniques de sondage.* Paris : Editions TECHNIP.

[Baron, 1997] Baron, J. (1997). La mesure du plomb au robinet de l'usager. étude des méthodes d'échantillonnage. *Techniques - Sciences - Méthodes*, 1(5), 47–54.

[Beauchemin et al., 2011] Beauchemin, S., MacLean, L. C. W., & Rasmussen, P. E. (2011). Lead speciation in indoor dust : a case study to assess old paint contribution in a canadian urban house. *Environmental Geochemistry and Health*, 33(4), 343–352.

[Beaumont & Haziza, 2012] Beaumont, J.-F. & Haziza, D. (2012). Traitement des valeurs influentes dans les enquêtes. In *7ème colloque francophone sur les sondages* Bruz : Société française de statistique.

[Binder & Roberts, 2003] Binder, D. A. & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of Survey Data* (pp. 29–48). John Wiley & Sons, Ltd.

[Bretin, 2006] Bretin, P. (2006). *Guide d'investigation environnementale des cas de saturnisme de l'enfant.* Guide méthodologique, InVS/DSE. ISBN 2-11-095899-5.

[Caravanos et al., 2006] Caravanos, J., Weiss, A. L., Blaise, M. J., & Jaeger, R. J. (2006). A survey of spatially distributed exterior dust lead loadings in new york city. *Environmental Research*, 100(2), 165–172.

[Carle, 2009] Carle, A. (2009). Fitting multilevel models in complex survey data with design weights : Recommendations. *BMC medical research methodology*, 9(1), 49.

[CDC, 2007] CDC (2007). *Interpreting and Managing Blood Lead Levels <10 µg/dL in Children and Reducing Childhood Exposures to Lead.* Technical Report Rep. 56 (RR-8), Centers for Disease Control and Prevention.

[Chantala & Suchindran, 2006] Chantala, K. & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models. In *Joint Statistical Meetings Annual Meeting* (pp. 2815–2824).

[Chatfield, 1995] Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3), 419–466.

[Chaventré et al., 2009] Chaventré, F., Kirchner, S., Bretin, P., Etchevers, A., Glorennec, P., & Le Bot, B. (2009). *Enquête nationale sur le plomb dans l'habitat 2008-2009 : étude des facteurs de risque environnementaux et comportementaux chez l'enfant de 6 mois à 6 ans. Rapport final 2008.* Présentation scientifique et protocoles ESE-SB-2009-077, CSTB.

[CITEPA, 2010] CITEPA (2010). *EMMISSIONS DANS L'AIR EN FRANCE, Métropole, Substances relatives à la contamination par les métaux lourds.* Technical report, Centre Technique Interprofessionnel d'Etude de la Pollution Atmosphérique, Paris.

[Clark et al., 2011] Clark, S., Galke, W., Succop, P., Grote, J., McLaine, P., Wilson, J., Dixon, S., Menrath, W., Roda, S., Chen, M., Bornschein, R., & Jacobs, D. (2011). Effects of HUD-supported lead hazard control interventions in housing on children's blood lead. *Environmental Research*, 111(2), 301–311.

[Clark et al., 2004] Clark, S., Menrath, W., Chen, M., Succop, P., Bornschein, R., Galke, W., & Wilson, J. (2004). The influence of exterior dust and soil lead on interior dust lead levels in housing that had undergone lead-based paint hazard control. *Journal of Occupational and Environmental Hygiene*, 1(5), 273–282.

[Clement et al., 2000] Clement, M., Seux, R., & Rabarot, S. (2000). A practical model for estimating total lead intake from drinking water. *Water Research*, 34(5), 1533–1542.

[Cook et al., 1993] Cook, M., Chappell, W. R., Hoffman, R. E., & Mangione, E. J. (1993). Assessment of blood lead levels in children living in a historic mining and smelting community. *American Journal of Epidemiology*, 137(4), 447–455.

[Davies et al., 1985] Davies, B., Elwood, P., Gallacher, J., & Ginnever, R. (1985). The relationships between heavy metals in garden soils and house dusts in an old lead mining area of north wales, great britain. *Environmental Pollution Series B, Chemical and Physical*, 9(4), 255–266.

[Davies et al., 1987] Davies, D., Watt, J., & Thornton, I. (1987). Lead levels in birmingham dusts and soils. *Science of The Total Environment*, 67(2–3), 177–185.

[Dixon et al., 2008] Dixon, S., Wilson, J., Kawecki, C., Green, R., Phoenix, J., Galke, W., Clark, S., & Breysse, J. (2008). Selecting a lead hazard control strategy based on dust lead loading and housing condition : I. methods and results. *Journal of Occupational and Environmental Hygiene*, 5(8), 530–539.

[Dixon et al., 2009] Dixon, S. L., Gaitens, J. M., Jacobs, D. E., Strauss, W., Nagaraja, J., Pivetz, T., Wilson, J. W., & Ashley, P. J. (2009). Exposure of U.S. children to residential dust lead, 1999–2004 : II. the contribution of lead-contaminated dust to children's blood lead levels. *Environ Health Perspect*, 117(3), 468–474.

[Dixon et al., 2012] Dixon, S. L., Jacobs, D. E., Wilson, J. W., Akoto, J. Y., Nevin, R., & Scott Clark, C. (2012). Window replacement and residential lead paint hazard control 12 years later. *Environmental Research*, 113(0), 14–20.

[Dixon et al., 2005a] Dixon, S. L., Wilson, J. W., Clark, C. S., Galke, W. A., Succop, P. A., & Chen, M. (2005a). Effectiveness of lead-hazard control interventions on dust lead loadings : Findings from the evaluation of the HUD lead-based paint hazard control grant program. *Environmental Research*, 98(3), 303–314.

[Dixon et al., 2005b] Dixon, S. L., Wilson, J. W., Clark, C. S., Galke, W. A., Succop, P. A., & Chen, M. (2005b). The influence of common area lead hazards and lead hazard control on dust lead loadings in multiunit buildings. *Journal of Occupational and Environmental Hygiene*, 2(12), 659–666.

[Douay et al., 2009] Douay, F., Pruvot, C., Waterlot, C., Fritsch, C., Fourrier, H., Loriette, A., Bidar, G., Grand, C., de Vaufleury, A., & Scheifler, R. (2009). Contamination of woody habitat soils around a former lead smelter in the north of france. *Science of The Total Environment*, 407(21), 5564–5577.

[Duguet et al., 1994] Duguet, J.-P., Cordonnier, J., & Brodard, E. (1994). Le plomb dans les eaux distribuées : Bilan qualitatif. détermination des zones à risque. *Techniques - Sciences - Méthodes*, 1(3), 128–130.

[European Commission et al., 1999] European Commission, Van den Hoven, J., Buijs, P., Jackson, P., Miller, S., Gardner, M., Leroy, P., Baron, J., Boireau, A., Cordonnier, J., Wagner, I., Maréeos do Monte, H., Benoliel, M., Papadopoulos, I., & Quevauviller, P. (1999). *Developing a new protocol for the monitoring of lead in drinking water, contract SMT4-CT96-2112.* Technical Report EUR 19087, European Commission, Luxembourg : Office for Official Publications of the European Communities. ISBN 92-828-6888-5.

[Euser et al., 2008] Euser, A. M., Dekker, F. W., & le Cessie, S. (2008). A practical approach to bland-altman plots and variation coefficients for log transformed variables. *Journal of clinical epidemiology*, 61(10), 978–982.

[Farfel et al., 2003] Farfel, M. R., Orlova, A. O., Lees, P. S., Rohde, C., Ashley, P. J., & Chisolm, J. J. (2003). A study of urban housing demolitions as sources of lead in ambient dust : Demolition practices and exterior dust fall. *Environmental Health Perspectives*, 111(9), 1228–1234.

[Farfel et al., 2005] Farfel, M. R., Orlova, A. O., Lees, P. S., Rohde, C., Ashley, P. J., & Julian Chisolm Jr., J. (2005). A study of urban housing demolition as a source of lead in ambient dust on sidewalks, streets, and alleys. *Environmental Research*, 99(2), 204–213.

[Farley, 1998] Farley, D. (1998). Dangers of lead still linger. *U.S. Food and Drug Administration*, (pp. 16–21).

[Fergusson et al., 1986] Fergusson, J. E., Forbes, E. A., Schroeder, R. J., & Ryan, D. E. (1986). The elemental composition and sources of house dust and street dust. *Science of The Total Environment*, 50(0), 217–221.

[Gaitens et al., 2009] Gaitens, J. M., Dixon, S. L., Jacobs, D. E., Nagaraja, J., Strauss, W., Wilson, J. W., & Ashley, P. J. (2009). Exposure of U.S. children to residential dust lead, 1999-2004 : I. housing and demographic factors. *Environmental Health Perspectives*, 117(3), 461–467.

[Geofabrik, 2008] Geofabrik (2008). European open street map. http ://freegeographytools.com/2008/european-open-street-map-osm-data-in-shapefile-format.

[Gibson, 2005] Gibson, J. L. (2005). A plea for painted railings and painted walls of rooms as the source of lead poisoning amongst queensland children. 1904. *Public Health Reports (Washington, D.C. : 1974)*, 120(3), 301–304.

[Girard, 2012] Girard, C. (2012). Estimation de la variance : quelles sont les options qui s'offrent à vous ? In *7ème colloque francophone sur les sondages* Bruz : Société française de statistique.

[Glorennec et al., 2007] Glorennec, P., Bemrah, N., Tard, A., Robin, A., Le Bot, B., & Bard, D. (2007). Probabilistic modeling of young children's overall lead exposure in france : Integrated approach for various exposure media. *Environment International*, 33(7), 937–945.

[Glorennec et al., 2005] Glorennec, P., Le Bot, B., Saramito, G., & Arcelin, C. (2005). Exposures to lead via dust ingestion of french children : a pilot study. In *15th annual conference of International Society for Exposure Analysis* Tucson Az, USA.

[Guillerme, 2002] Guillerme, A. (2002). La céruse. *Techniques et Culture*, 1(38).

[Guillerme et al., 2003] Guillerme, A., Ciriacono, S., Lanoe, C., Nègre, V., Emptoz, G., Lestel, L., Jorland, G., & Lefort, A.-C. (2003). *La céruse : usages et effets, Xe - XXe siècles.* Centre d'Histoire des Techniques CNAM.

[Harrell, 2001] Harrell, F. E. (2001). *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis.* Springer.

[Haziza, 2012] Haziza, D. (2012). La non-réponse : comment affecte-t-elle la qualité des estimations ? In *SSMI méthodologie sur les données manquantes* InVS Saint-Maurice : InVS.

[Helsel, 2012] Helsel, D. R. (2012). *Statistical methods for censored environmental data using minitab and R.* Chichester [u.a.] : Wiley-Blackwell.

[Hozhabri et al., 2004] Hozhabri, S., White, F., Rahbar, M. H., Agboatwalla, M., & Luby, S. (2004). Elevated blood lead levels among children living in a fishing community, karachi, pakistan. *Archives of Environmental Health*, 59(1), 37–41.

[Hunt et al., 2012] Hunt, A., Johnson, D., Griffith, D., & Zitoon, S. (2012). Citywide distribution of lead and other element in soils and indoor dusts in syracuse, NY. *Applied Geochemistry*, 27(5), 985–994.

[Hunt et al., 2006] Hunt, A., Johnson, D. L., & Griffith, D. A. (2006). Mass transfer of soil indoors by track-in on footwear. *Science of The Total Environment*, 370(2–3), 360–371.

[ILO, 1921] ILO (1921). White lead (painting) convention. In *C013. Presented at the General Conference of the International Labour Organisation* Geneva.

[INSEE, 2008] INSEE (2008). French population census 2006. french national institute of statistics and economic studies. http ://www.recensement.insee.fr/accesDonneesTelechargeables.action accessible au 21-11-2009.

[INSERM, 1999] INSERM (1999). *Plomb dans l'environnement : quels risques pour la santé ?* Paris : Institut National de la Santé et de la Recherche Médicale, les éditions inserm edition.

[INSERM, 2008] INSERM (2008). *Saturnisme. Quelles stratégie de dépistage chez l'enfant ?* Paris : Institut National de la Santé et de la Recherche Médicale, les éditions inserm edition.

[Jacobs et al., 2002] Jacobs, D. E., Clickner, R. P., Zhou, J. Y., Viet, S. M., Marker, D. A., Rogers, J. W., Zeldin, D. C., Broene, P., & Friedman, W. (2002). The prevalence of lead-based paint hazards in U.S. housing. *Environmental Health Perspectives*, 110(10), A599–A606.

[Jiang & Succop, 1996] Jiang, Q. & Succop, P. A. (1996). A study of the specificaton of the log-log and log-additive models for the relationship between blood lead and environmental lead. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(4), 426–434.

[Kim et al., 2006] Kim, J. K., Michael Brick, J., Fuller, W. A., & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(3), 509–521.

[Kramer, 2005] Kramer, M. H. (2005). R^2 statistics for mixed models. In *17th Annual Kansas State University Conference on Applied Statistics in Agriculture* (pp. 148–160). Manhattan, KS.

[Laidlaw & Filippelli, 2008] Laidlaw, M. A. & Filippelli, G. M. (2008). Resuspension of urban soils as a persistent source of lead poisoning in children : A review and new directions. *Applied Geochemistry*, 23(8), 2021–2039.

[Laidlaw et al., 2005] Laidlaw, M. A., Mielke, H. W., Filippelli, G. M., Johnson, D. L., & Gonzales, C. R. (2005). Seasonality and children's blood lead levels : Developing a predictive model using climatic variables and blood lead data from indianapolis, indiana, syracuse, new york, and new orleans, louisiana (USA). *Environmental Health Perspectives*, 113(6), 793–800.

[Lanphear, 2002] Lanphear, B. (2002). Environmental lead exposure during early childhood. *The Journal of Pediatrics*, 140(1), 40–47.

[Lanphear et al., 1995] Lanphear, B. P., Emond, M., Jacobs, D. E., Weitzman, M., Tanner, M., Winter, N. L., Yakir, B., & Eberly, S. (1995). A side-by-side comparison of dust collection methods for sampling lead-contaminated house dust. *Environmental Research*, 68(2), 114–123.

[Lanphear et al., 2005] Lanphear, B. P., Hornung, R., & Ho, M. (2005). Screening housing to prevent lead toxicity in children. *Public Health Reports*, 120(3), 305–310.

[Lanphear et al., 1998] Lanphear, B. P., Matte, T., Rogers, J., Clickner, R., Dietz, B., Bornschein, R., Succop, P., Mahaffey, K., Dixon, S., Galke, W., Rabinowitz,

M., Farfel, M., Rohde, C., Schwartz, J., Ashley, P., & Jacobs, D. (1998). The contribution of lead-contaminated house dust and residential soil to children's blood lead levels : A pooled analysis of 12 epidemiologic studies. *Environmental Research*, 79(1), 51–68.

[Lanphear & Roghmann, 1997] Lanphear, B. P. & Roghmann, K. J. (1997). Pathways of lead exposure in urban children. *Environmental Research*, 74(1), 67–73.

[Lanphear et al., 2003] Lanphear, B. P., Succop, P., Roda, S., & Henningsen, G. (2003). The effect of soil abatement on blood lead levels in children living near a former smelting and milling operation. *Public Health Reports*, 118(2), 83–91.

[Lanphear et al., 1996] Lanphear, B. P., Weitzman, M., Winter, N., Eberly, S., Yakir, B., Tanner, M., Emond, M., & Matte, T. (1996). Lead-contaminated house dust and urban children's blood lead levels. *American Journal of Public Health*, 86(10), 1416–1421.

[Laperche et al., 2004] Laperche, V., Dictor, M., Clozel-Leloup, B., & Baranger, P. (2004). *Guide méthodologique du plomb appliqué à la gestion des sites et des sols pollués*. Guide méthodologique BRGM/RP-52881-FR, BRGM.

[Laperche & Mossmann, 2004] Laperche, V. & Mossmann, J. R. (2004). *Protocole d'échantillonnage de sols urbains pollués par du plomb.* Technical Report BRGM/RP-52928-FR, BRGM.

[Laxen et al., 1988] Laxen, D. P. H., Lindsay, F., Raab, G. M., Hunter, R., Fell, G. S., & Fulton, M. (1988). The variability of lead in dusts within the homes of young children. *Environmental Geochemistry and Health*, 10(1), 3–9.

[Layton & Beamer, 2009] Layton, D. W. & Beamer, P. I. (2009). Migration of contaminated soil and airborne particulates to indoor dust. *Environ. Sci. Technol.*, 43(21), 8199–8205.

[Le Bot et al., 2011] Le Bot, B., Arcelin, C., Briand, E., & Glorennec, P. (2011). Sequential digestion for measuring leachable and total lead in the same sample of dust or paint chips by ICP-MS. *Journal of Environmental Science and Health, Part A*, 46(1), 63–69.

[Lestel, 2002] Lestel, L. (2002). La production de céruse en france au XIXe siècle : évolution d'une industrie dangereuse. *Techniques & Culture. Revue semestrielle d'anthropologie des techniques*, 1(38).

[Lubin et al., 2004] Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., & Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*, 112, 1691–1696.

[Lucas, 2011] Lucas, J.-P. (2011). Historique de la règlementation relative à l'emploi de la céruse et des dérivés du plomb dans la peinture en france. *Environnement, Risques & Santé*, 10(4), 316–322.

[Lucas et al., 2014] Lucas, J.-P., Bellanger, L., Le Strat, Y., Le Tertre, A., Glorennec, P., Le Bot, B., Etchevers, A., Mandin, C., & Sébille, V. (2014). Source contributions of lead in residential floor dust and within-home variability of dust lead loading. *Science of The Total Environment*, 470–471, 768–779.

[Lucas et al., 2012] Lucas, J.-P., Le Bot, B., Glorennec, P., Etchevers, A., Bretin, P., Douay, F., Sébille, V., Bellanger, L., & Mandin, C. (2012). Lead contamination in french children's homes and environment. *Environmental Research*, 116(0), 58–65.

[Lucas et al., 2013] Lucas, J.-P., Sébille, V., Le Tertre, A., Le Strat, Y., & Bellanger, L. (2013). Multilevel modelling of survey data : impact of the 2-level weights used in the pseudolikelihood. *Journal of Applied Statistics*, In Press.

[Lumley, 2004] Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(8), 1–19.

[Lumley, 2010a] Lumley, T. (2010a). *Complex Surveys*. Hoboken, NJ, USA : John Wiley & Sons, Inc.

[Lumley, 2010b] Lumley, T. (2010b). Survey : analysis of complex survey samples. r package version 3.22-4.

[Matt et al., 2011] Matt, G. E., Quintana, P. J. E., Destaillats, H., Gundel, L. A., Sleiman, M., Singer, B. C., Jacob, P., Benowitz, N., Winickoff, J. P., Rehan, V., Talbot, P., Schick, S., Samet, J., Wang, Y., Hang, B., Martins-Green, M., Pankow, J. F., & Hovell, M. F. (2011). Thirdhand tobacco smoke : Emerging evidence and arguments for a multidisciplinary research agenda. *Environmental Health Perspectives*, 119(9), 1218–1226.

[MEEDDTL, 2011] MEEDDTL (2011). Soils pollution BASOL. department of ecology, sustainable development, transport and housing. http ://basol.environnement.gouv.fr/.

[MEEDDTL & BRGM, 2011] MEEDDTL & BRGM (2011). Inventory of historic industrial sites and service activities BASIAS. department of ecology, sustainable development, transport and housing. http ://basias.brgm.fr/.

[MEEDDTL & INERIS, 2003] MEEDDTL & INERIS (2003). French registry of polluting emissions. BDREP. http ://www.irep.ecologie.gouv.fr/IREP/index.php.

[Mench & Baize, 2004] Mench, M. & Baize, D. (2004). Contamination des sols et de nos aliments d'origine végétale par les éléments en traces. *Courrier de l'environnement de l'INRA*, 1(52).

[Mielke et al., 2010] Mielke, H. W., Laidlaw, M. A., & Gonzales, C. (2010). Lead (pb) legacy from vehicle traffic in eight california urbanized areas : Continuing influence of lead dust on children's health. *Science of The Total Environment*, 408(19), 3965–3975.

[Mielke et al., 2011] Mielke, H. W., Laidlaw, M. A., & Gonzales, C. R. (2011). Estimation of leaded (pb) gasoline's continuing material and health impacts on 90 US urbanized areas. *Environment International*, 37(1), 248–257.

[Ministère de la Santé et des Solidarités, 2006] Ministère de la Santé et des Solidarités (2006). Arrêté de 25 avril 2006 relatif au constat de risque d'Exposition au plomb. JORF 26 Avril 2006, Texte 52 sur 151.

[Neyman, 1934] Neyman, J. (1934). On the two different aspects of the representative method : The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.

[Nriagu, 1990] Nriagu, J. O. (1990). The rise and fall of leaded gasoline. *Science of The Total Environment*, 92(0), 13–28.

[Pfeffermann et al., 2006] Pfeffermann, D., Moura, F. A. D. S., & Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93(4), 943–959.

[Pfeffermann et al., 1998] Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60(1), 23–40.

[Pichery et al., 2011] Pichery, C., Bellanger, M., Zmirou-Navier, D., Glorennec, P., Hartemann, P., & Grandjean, P. (2011). Childhood lead exposure in france : benefit estimation and partial cost-benefit analysis of lead hazard control. *Environmental Health*, 10(1), 44.

[Rabe-Hesketh, 2007] Rabe-Hesketh, S. (2007). *Multilevel modeling of complex survey data*. West coast stata users' group meetings 2007, Stata Users Group.

[Rabe-Hesketh & Skrondal, 2006] Rabe-Hesketh, S. & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 169(4), 805–827.

[Rabe-Hesketh et al., 2004] Rabe-Hesketh, S., Skrondal, A., & Pickels, A. (2004). *GLLAMM manual*. Technical Report Technical Report 160, Division of Biostatistics, University of California, Berkeley.

[Rabe-Hesketh et al., 2007] Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). 10 - multilevel structural equation modeling. In Sik-Yum Lee (Ed.), *Handbook of Latent Variable and Related Models* (pp. 209–227). Amsterdam : North-Holland.

[Rabinowitz et al., 1985] Rabinowitz, M., Leviton, A., Needleman, H., Bellinger, D., & Waternaux, C. (1985). Environmental correlates of infant blood lead levels in boston. *Environmental Research*, 38(1), 96–107.

[Rabito et al., 2007] Rabito, F., Iqbal, S., Shorter, C., Osman, P., Philips, P., Langlois, E., & White, L. (2007). The association between demolition activity and children's blood lead levels. *Environmental Research*, 103(3), 345–351.

[Raftery, 1998] Raftery, A. E. (1998). *Bayes Factors and BIC : Comment on Weakliem*. Technical Report 347, Department of Statistics University of Washington, Seattle, WA.

[Reissman et al., 2002] Reissman, D. B., Matte, T. D., Gurnitz, K. L., Kaufmann, R. B., & Leighton, J. (2002). Is home renovation or repair a risk factor for exposure to lead among children residing in new york city ? *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, 79(4), 502–511.

[Reiter et al., 2006] Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Survey Methodology*, 32(2), 161–168.

[Rosner, 2006] Rosner, B. A. (2006). *Fundamentals Of Biostatistics*. Cengage Learning.

[Rust et al., 1997] Rust, S. W., Burgoon, D. A., Lanphear, B. P., & Eberly, S. (1997). Log-additive versus log-linear analysis of lead-contaminated house dust and children's blood-lead levels. *Environmental Research*, 72(2), 173–184.

[Sanborn et al., 2002] Sanborn, M. D., Abelsohn, A., Campbell, M., & Weir, E. (2002). Identifying and managing adverse environmental health effects : 3. lead exposure. *Canadian Medical Association Journal*, 166(10), 1287–1292.

[Särndal et al., 1992] Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

[Schapiro & Bretin, 2006] Schapiro, E. & Bretin, P. (2006). *Sources inhabituelles d'intoxication par le plomb chez l'enfant et la femme enceinte*. Note technique, InVS/DSE.

[Sheets et al., 2001] Sheets, R. W., Kyger, J. R., Biagioni, R. N., Probst, S., Boyer, R., & Barke, K. (2001). Relationship between soil lead and airborne lead concentrations at springfield, missouri, USA. *Science of The Total Environment*, 271(1–3), 79–85.

[Skinner, 1989] Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In *Analysis of complex surveys (eds C. J. Skinner, D. Holt and T. M. F. Smith)*. Chichester : Wiley.

[StataCorp, 2011] StataCorp (2011). *Stata multiple imputation reference manual Release 12*. College Station, TX : Stata Press.

[Sterling et al., 1999] Sterling, D. A., Roegner, K. C., Lewis, R. D., Luke, D. A., Wilder, L. C., & Burchette, S. M. (1999). Evaluation of four sampling methods for determining exposure of children to lead-contaminated household dust,. *Environmental Research*, 81(2), 130–141.

[Sterne & Davey Smith, 2001] Sterne, J. A. & Davey Smith, G. (2001). Sifting the evidence-what's wrong with significance tests ? *BMJ (Clinical research ed.)*, 322(7280), 226–231.

[Sterne et al., 2009] Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research : potential and pitfalls. *BMJ (Clinical research ed.)*, 338, b2393.

[Sturges & Harrison, 1985] Sturges, W. & Harrison, R. M. (1985). An assessment of the contribution from paint flakes to the lead content of some street and household dusts. *Science of The Total Environment*, 44(3), 225–234.

[Succop et al., 1998] Succop, P., Bornschein, R., Brown, K., & Tseng, C. Y. (1998). An empirical comparison of lead exposure pathway models. *Environmental Health Perspectives*, 106(Suppl 6), 1577–1583.

[Thornton et al., 1990] Thornton, I., Davies, D. J., Watt, J. M., & Quinn, M. J. (1990). Lead exposure in young children from dust and soil in the united kingdom. *Environmental Health Perspectives*, 89, 55–60.

[Tillé, 2001] Tillé, Y. (2001). *Théorie des sondages : Echantillonnage et estimation en populations finies*. Paris : Dunod.

[Tong & Lam, 2000] Tong, S. T. & Lam, K. C. (2000). Home sweet home ? a case study of household dust contamination in hong kong. *Science of The Total Environment*, 256(2-3), 115–123.

[U.S. EPA, 2000] U.S. EPA (2000). *Lead Exposure Associated with Renovation and Remodelling Activities - Final summary report*. Technical Report EPA 747-S-00-001.

[U.S. EPA, 2001] U.S. EPA (2001). Identification of dangerous levels of lead ; final rule. 40 CFR 745. 66(4), 1206–1240.

[U.S. EPA, 2010] U.S. EPA (2010). Lead and copper rule – code of federal regulations 40 CFR part 141. (pp. Subpart I – Control of Lead and Copper).

[U.S. HUD, 1995a] U.S. HUD (1995a). *Guidelines for the Evaluation and Control of Lead-Based Paint Hazards in Housing.* Technical report, U.S. Department of Housing and Urban Development, Washington, DC.

[U.S. HUD, 1995b] U.S. HUD (1995b). *Guidelines for the Evaluation and Control of Lead-Based Paint Hazards in Housing. Chapter 5 : Risk Assessment.* Technical report, U.S. Department of Housing and Urban Development, Washington, DC.

[U.S. HUD, 2003] U.S. HUD (2003). *An Evaluation of Residual Lead Dust Following Lead Abatement Clean-up and Clearance Activities.* Technical Report Final Report, U.S. Department of Housing and Urban Development, Washington, DC.

[Wilson et al., 2007] Wilson, J., Dixon, S., Galke, W., & McLaine, P. (2007). An investigation of dust lead sampling locations and children's blood lead levels. *Journal of Exposure Science & Environmental Epidemiology*, 17(1), 2–12.

[Yates & Grundy, 1953] Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2), 253–261.

[Yiin et al., 2003] Yiin, L.-M., Lioy, P. J., & Rhoads, G. G. (2003). Impact of home carpets on childhood lead intervention study. *Environmental Research*, 92(2), 161–165.

[Yiin et al., 2000] Yiin, L. M., Rhoads, G. G., & Lioy, P. J. (2000). Seasonal influences on childhood lead exposure. *Environmental Health Perspectives*, 108(2), 177–182.

Division de la recherche et des écoles doctorales
Bureau des études doctorales & Coordination des écoles doctorales

RESUMÉ et MOTS CLÉS

Ce même document vous servira à compléter les formulaires de dépôt de thèse en BU, à insérer en couverture VERSO de votre manuscrit (obligatoire sous cette forme) et à joindre à votre imprimé de demande d'autorisation de soutenance (1 700 caractères maximum, espaces compris)

TITRE EN FRANÇAIS CONTAMINATION DES LOGEMENTS PAR LE PLOMB : PREVALENCE DES LOGEMENTS A RISQUE ET IDENTIFICATION DES DETERMINANTS DE LA CONTAMINATION

Résumé et mots-clés en français

Les niveaux en plomb (Pb) en milieu résidentiel ont été estimés pour la 1ère fois en France métropolitaine. Pour cela les outils de la théorie des sondages ont été appliqués aux données de l'enquête Plomb-Habitat (2008-2009). Un échantillon de 484 logements a été construit afin d'étudier la population de résidences principales (N = 3 581 991) abritant au moins un enfant âgé de 6 mois à 6 ans. Environ 2,9% des logements possèdent une concentration en Pb dans l'eau du robinet supérieure au seuil réglementaire (SR) de 10 µg/L ; dans 0,21% des logements et 4,1% des parties communes le SR américain de 430 µg/m² de Pb est dépassé dans les poussières intérieures ; 1,4% des sols des aires de jeu extérieures dépassent le SR américain de 300 mg/kg de Pb ; 24,5% des logements ont encore des peintures au Pb.

Le Pb des poussières est connu comme étant le principal prédicteur des plombémies infantiles. Un modèle multi-niveaux à 2 niveaux a été construit afin d'expliquer les charges en Pb des poussières des 1834 pièces (niveau 1) investiguées dans les logements (niveau 2). Aucune pondération n'a été introduite dans la méthode d'estimation (pseudo vraisemblance) utilisée pour ce type de modèle sur données d'enquête. La poussière du palier est le principal contributeur à la contamination des poussières.

Une étude de simulation a été réalisée à partir de nos données afin de comparer les différentes pondérations pour le niveau 2 d'un modèle multi-niveaux. Son résultat a permis de valider l'utilisation d'un modèle à 2 niveaux non pondéré pour expliquer les charges en Pb de la poussière. Jusqu'alors dans la littérature, seule la pondération au niveau 1 avait été étudiée pour ce type de modèle.

Mots clés : Plomb ; Logement ; Enquête ; Poussière ; Théorie des sondages ; Modélisation multi-niveaux ; Simulation Monte Carlo.

TITRE EN ANGLAIS LEAD CONTAMINATION IN HOMES: PREVALENCE OF DWELLINGS WITH LEAD HAZARDS AND IDENTIFICATION OF THE DETERMINANTS OF THE CONTAMINATION

Résumé et mots-clés en anglais

Residential lead levels were estimated for the first time in mainland France. For this, tools of the theory of survey sampling were applied to the data of the Plomb-Habitat survey (2008-2009). A sample of 484 dwelling was drawn to study the population (N = 3 581 991) of the main residences (as opposed to second home) where at least one child aged 6 months to 6 years was present. Approximately 2.9% of housing units have a lead concentration in tap water higher or equal than the regulatory threshold (RT) of 10 µg/L; in approximately 0.21% of dwellings and in 4.1% of common areas the American RT of 430 µg/m² (40 µg/ft²) was exceeded for interior floor dust lead; 1.4% of exterior play area soils exceed the American RT of 300 mg/kg of lead; 24.5% of housing units have still lead-based paint.

Lead in floor dust was pointed out as the main predictor of blood lead level in children. A multilevel model with 2 levels was fitted to explain the floor dust lead loadings of the 1834 rooms as level-1 units investigated in the homes considered as level-2 units. No weights was used in the estimation method (pseudolikelihood) used for this kind of modeling on survey data. Dust of the landing of an apartment is the main contributor to the contamination of dust by lead.

A simulation study was carried out from our data to compare the different weights for the level-2 units of a multilevel model. Its result enabled us to confirm the fitting of an unweighted model to explain the dust lead loadings. Until now, only the level-1 weights had been studied in the literature for this kind of model.

Keywords: Lead; Housing; Survey; Dust; Survey sampling; Multilevel modeling; Monte Carlo simulation.

NOM Prénom LUCAS JEAN-PAUL

Visa du Directeur de thèse