Characterization of environmental inequalities due to Polyaromatic Hydrocarbons in France : developing environmental data processing methods to spatialize exposure indicators for PAH substances

Despoina Ioannidou

le cnam

École doctorale Informatique, Télécommunications et Électronique (Paris)

Centre d'Études et de Recherche en Informatique et Communications

THÈSE DE DOCTORAT

présentée par : **Despoina IOANNIDOU**

soutenance proposée le : **25 Juin 2018**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Spécialité : **Informatique**

Characterization of environmental inequalities due to Polyaromatic Hydrocarbons in France:

Developing environmental data processing methods to spatialize exposure indicators for PAH substances.

THÈSE dirigée par

M. CAUDEVILLE Julien	*Ingénieur de Recherche, Ineris*
Mme. MALHERBE Laure	*Ingénieur de Recherche, Ineris*
M. LATOUCHE Aurélien	*Professeur des Universités, Cnam*

RAPPORTEURS

M. GUEDDA Mohammed	*Professeur des Universités, Université de Picardie*
Mme. DEGUEN Severine	*Enseignante chercheuse, EHESP*

EXAMINATEURS

Mme. CHARDON Karen	*Président du jury, Professeur des Universités, Université de Picardie*
M. ALBINET Alexandre	*Ingénieur de Recherche, Ineris*
Mme. PERDRIX Esperanza	*Enseignante chercheuse, IMT Lille Douai*

Abstract

Reducing environmental exposure inequalities has become a major focus of public health efforts in France, as evidenced by the French action plans for health and the environment. The aim of this thesis is to develop an integrated approach to characterize environmental inequalities and evaluate the spatialized exposure to PAH in France.

The data retained the monitoring quality networks reflect the actual contamination of the environmental compartments and the overall exposure of the populations. However they do not always provide an adequate spatial resolution to characterize environmental exposures, as they are usually not assembled for this specific purpose. Statistical methods are employed to process input databases (environmental concentrations in water, air and soil) in the objective of characterizing the exposure. A multimedia model interfaced with a GIS, allows the integration of environmental variables in order to yield exposure doses related to ingestion of food, water and soil and inhalation.The methodology was applied to three Polycyclic Aromatic Hydrocarbon substances, (benzo[a]pyrene, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene) in France. The results obtained permit the mapping of exposure indicators, the identifications of areas of overexposure and the characterization of environmental determinants. In the context of exposure characterization, the direct spatialization of available data from environmental measurement datasets poses a certain number of methodological questions which lead to uncertainties related to the sampling and the spatial and temporal representativeness of data. These could be reduced by acquiring additional data or by constructing predictive variables related to the spatial and temporal phenomena considered.

Data processing algorithms and calculation of exposure carried out in this work, will be integrated in the French coordinated integrated environment and health platform-PLAINE

in order to be applied on other pollutants and prioritize preventative actions.

Keywords : statistics, exposure, environment, inequalities, spatial modeling, kriging, Polyaromatic Hydrocarbons

Résumé

La réduction des inégalités d'exposition environnementale constitue un axe majeur en santé publique en France comme en témoignent les priorités des différents Plan Nationaux Santé Environnement (PNSE). L'objectif de cette thèse est de développer une approche intégrée pour la caractérisation des inégalités environnementales et l'évaluation de l'exposition spatialisée de la population aux HAP en France.

Les données produites dans le cadre des réseaux de surveillance de la qualités des milieux environnementaux sont le reflet de la contamination réelle des milieux et de l'exposition globale des populations. Toutefois, elles ne présentent généralement pas une représentativité spatiale suffisante pour caractériser finement les expositions environnementales, ces réseaux n'ayant pas été initialement conçus dans cet objectif. Des méthodes statistiques sont développées pour traiter les bases de données d'entrée (concentrations environnementales dans l'eau, l'air et le sol) et les rendre pertinentes vis à vis des objectifs définis de caractérisation de l'exposition. Un modèle multimédia d'exposition, interfacé avec un Système d'Information Géographique pour intégrer les variables environnementales, est développé pour estimer les doses d'exposition liées à l'ingestion d'aliments, d'eau de consommation, de sol et à l'inhalation de contaminants atmosphériques. La méthodologie a été appliquée pour trois Hydrocarbures Aromatiques Polycycliques (benzo[a]pyrène, benzo[ghi]pérylène et indéno[1,2,3-cd]pyrène) sur l'ensemble du territoire français. Les résultats permettent de cartographier des indicateurs d'exposition, d'identifier les zones de surexposition et de caractériser les déterminants environnementaux. Dans une logique de caractérisation de l'exposition, la spatialisation des données issues des mesures environnementales pose un certain nombre de questions méthodologiques qui confèrent aux cartes réalisées de nombreuses incertitudes et limites relatives à l'échantillonnage et aux représentativités

spatiales et temporelles des données. Celles-ci peuvent être réduites par l'acquisition de données supplémentaires et par la construction de variables prédictives des phénomènes spatiaux et temporels considérés.

Les outils de traitement statistique de données développés dans le cadre de ces travaux seront intégrés dans la plateforme PLAINE pour être déclinés sur d'autres polluants en vue de prioriser les mesures de gestion à mettre en œuvre.

Mots clés : statistiques, exposition, environnement, inégalités, modelisation spatiale, kriging, Hydrocarbure Polyaromatique

Résumé étendu en français

Introduction

Lorsque l'on parle des inégalités environnementales, on fait référence à la reconnaissance des disparités entre territoires ou populations en matière d'exposition environnementale. Dans certaines régions, les personnes sont plus susceptibles d'être exposées aux effets négatifs de la pollution de l'air, du sol ou de l'eau qu'à d'autres. C'est pourquoi, cette question a été identifiée comme une priorité des politiques publiques françaises, comme en témoigne l'émergence des plans nationaux pour la santé et l'environnement. Plus précisément, le troisième plan (2015-2019) souligne la nécessité d'établir une méthodologie pour identifier et réduire les inégalités en matière de santé environnementale. Dans ce contexte, le processus de contamination est extrêmement complexe et variable dans l'espace et dans le temps.

L'évaluation de l'exposition, qui est basée sur la surveillance environnementale à court terme et/ou la modélisation des contaminant, présent un cadre adapté pour caractériser les inégalités environnementales. La combinaison de l'évaluation de l'exposition et des données spatiales est un élément fondamental qui nécessite d'abord de surmonter différentes limites scientifiques, comme par exemple, le couplage de plusieurs bases de données pour décrire la chaîne source-effet globale. Dans le but de cartographier les inégalités environnementales, l'objectif de ce travail est d'explorer des techniques de développement d'indicateurs composites pour la représentation spatiale des risques et l'analyse des inégalités environnementales. Pour y remédier, il faut :

- la caractérisation et l'intégration de phénomènes spatiaux fonctionnant à différentes échelles spatio-temporelles;

- l'intégration et la combinaison de différents niveaux de données provenant de différents compartiments environnementaux;
- la caractérisation des principales voies d'exposition;
- la construction de scénarios d'exposition réalistes intégrant les sources passées et présentes;
- la description des phénomènes à une échelle spatiale et temporelle fine.

Les questions ci-dessus sont traitées en employant des méthodes statistiques et géostatistiques pour spatialiser les mesures de concentration des HAP en France dans les trois compartiments environnementaux (air, eau, sol).

Bibliographie

Les inégalités environnementales peuvent se rapporter à la discrimination raciale (Been 1994, Hamilton 1995) ou aux résultats environnementaux inéquitables, tels que des expositions disparates ou des impacts sanitaires / sociaux et une répartition inégale des charges environnementales (Cutter 1995, Downey 1998, Pulido 2000). et Hogan 1998). Une exposition disparate se produit lorsque la population vulnérable est plus exposée à certains polluants environnementaux qu'à d'autres (Glickman 1994, Sadd et al. 1999, Szasz et Meuser 1997), alors que les effets sur la santé disparates sont les effets négatifs sur la santé (Ross, Reynolds et Geis 2000, Downey et Van Willigen 2005).

Des modèles d'exposition multimédia peuvent être utilisés pour prendre en compte la contribution de chaque voie d'exposition dans l'estimation afin de caractériser l'exposition totale (McKone et MacLeod, 2003). L'utilisation d'un tel modèle permet d'établir un cadre quantitatif pour évaluer le transfert de substances dans l'environnement et les compartiments d'exposition, ainsi que pour analyser l'exposition de la population. Par conséquent, les résultats d'une telle évaluation peuvent constituer une base solide et fiable pour la gestion des inégalités environnementales. Les différentes échelles spatiales du modèle multimédia sont discutées en détail par McKone et MacLeod (2003). Comme l'ont souligné ces auteurs, les modèles multimédias ont évolué des modèles régionaux, aux modèles multirégionaux et modèles mondiaux. La très fine résolution ne correspond pas à

une connaissance suffisamment détaillée et ne reflète pas les variations d'exposition locales avec une précision suffisante. De plus, ces modèles de devenir et de transport n'ont pas réussi à capturer l'exposition de fond et les sources passées, car les sources d'émission historiques ne sont pas disponibles pour une période suffisamment longue. Ce problème pourrait être partiellement résolu en utilisant des réseaux de surveillance fournissant des données de bonne qualité pour la caractérisation des voies d'exposition. C'est ainsi que le modèle d'exposition multimédia MODUL'ERS a été initialement mis au point par l'INERIS (Bonnard, 2003; Bonnard et McKone, 2010) pour évaluer le transfert de contaminant de l'environnement (air, sol, eau) vers la chaîne alimentaire locale jusqu'à une exposition individuelle. Ce projet a ajouté une dimension spatiale au modèle en utilisant le SIG et a été adapté aux bases de données disponibles pour le suivi régional et national français. À l'échelle régionale et à une résolution fine, l'utilisation directe de mesures de qualité environnementales n'est pas appropriée pour évaluer les niveaux environnementaux d'un contaminant sur une zone continue, car cela entraîne une mauvaise classification des expositions. Pour tenter de réduire ce problème, plusieurs méthodes d'analyse spatiale plus sophistiquées ont été mises au point pour améliorer la représentativité des données et caractériser l'incertitude associée (Gay et Korre, 2006; Goovaerts, 2006).

La caractérisation des zones contaminées repose sur un échantillon d'observations de contaminant et de leurs niveaux. La distribution spatiale de ce contaminants doit être estimée par une technique d'interpolation ou simulée par de multiples réalisations de la zone contaminée. Les progrès réalisés dans le domaine de la géostatistique au cours des trente dernières années ont permis d'atteindre les objectifs ci-dessus. À partir du début des années 1980, l'Environmental Protection Agency des États-Unis a été l'une des premières à proposer l'utilisation de méthodes géostatistiques pour cartographier la contamination des sols (Moore et McLaughlin, 1980).

Les approches initiales comprenaient la simple cartographie bidimensionnelle des catégories de contaminant avec les approches de krigeage (Goovaerts, 2001; Van Meirvenne et Goovaerts, 2001; Cattle et al, 2002, Demougeot-Renard et De Fouquet, 2004). Les méthodes les plus évoluées comprennent aujourd'hiu l'évaluation de l'incertitude sur la classification des sols et la localisation des polluants. Ces approches géostatistiques sont

également largement utilisées dans un contexte d'évaluation des risques pour la santé afin de caractériser les sites contaminés (Glavin et Hooda, 2005; Thayer et al., 2003; Waller et Gotway, 2004).

Les solutions simples, comme l'attribution d'une valeur moyenne à une grande région, peuvent conduire à une classification erronée de l'exposition (Chen, 2006; Pennigton et al, 2005; Lepom, 2009). Des nombreuses études ont montré des disparités fortes lors de l'évaluation de la présence d'un polluant dans les zones urbaines, rurales ou fortement industrialisées. Ils ont montré, par exemple, que les personnes résidant dans les zones urbaines, à proximité des autoroutes ou des industries sont plus touchées que celles résidants dans les zones rurales. Des études épidémiologiques ont également montré le résultat aggravant de la pollution atmosphérique entre les différentes zones urbaines et non urbaines.

Les procédures d'analyse chimique ne possédant pas une précision infinie, ils existent des concentrations qui ne peuvent pas être mesurées avec précision. Un traitement approprié de ces observations ayant des niveaux de contaminant non détectables a été un problème commun. Lorsqu'une valeur approche de zéro, il existe un point où tout signal dû au contaminant ne peut pas être différencié du bruit de fond, c.-à-d. que la précision de l'instrument / de la méthode n'est pas suffisante pour discerner la présence du composé précision. On parle dans ce cas des mesures inférieures à la limite de détection. Bien que l'on puisse faire la distinction entre une «limite de détection de l'instrument» et une «limite de détection de la méthode», c'est la limite de détection de la méthode qui présente le plus d'intérêt.

Le problème de l'estimation des paramètres des distributions avec des données censurées a été largement étudié (Cohen, 1959; Gleit, 1985; Helsel et Guilliom, 1986; Helsel, 1990; Hornung et Reed, 1990; Lambert, 1991; Özkaynak et al., 1991). ; Finkelstein et Verma, 2001). L'approchela plus simple pour traiter les données censurées est la substitution des données en dessous de la limite de détection , généralement remplacée par zéro, par une fraction de la limite de détection ou par la limite de détection elle-même. Dans un bootstrap, un algorithme EM (« expectation maximisation ») a été utilisé pour effectuer l'imputation multiple. Honaker, King et Blackwell proposent un algorithme qui traite les données

des séries chronologiques, en utilisant l'algorithme EM sur des échantillons bootstrap des données d'origine pour dessiner des valeurs à partir des paramètres de données complets. Cette méthode consiste à créer plusieurs ensembles de données "complets", qui peuvent ensuite être analysés en tant que cas complets.

Dans le processus de quantification de l'exposition, il est important de tenir compte de la variabilité et de l'incertitude qui sont insérées dans le modèle. La variabilité peut être introduite à la suite de processus intrinsèquement stochastiques ou à la suite de différences entre individus d'une population, ce qui présente un intérêt particulier pour l'évaluation des risques pour la santé humaine. C'est une propriété de la nature et il ne peut pas être réduit par des recherches supplémentaires, il peut cependant être vérifié et estimé avec plus de précision. Contrairement à la variabilité, l'incertitude peut être réduite par des recherches supplémentaires, mais elle ne peut pas être vérifiée (Özkaynak et al., 2008).

Ces modèles peuvent fonctionner à différentes échelles spatio-temporelles, ce qui pose un problème lors de leur couplage dans un cadre cohérent et peut entraîner une incertitude structurelle. De plus, les incertitudes dans un composant du modèle peuvent se propager dans le reste des composants, contribuant ainsi à l'incertitude globale du résultat. Malgré l'importance de ces types de problèmes de couplage de modèles, ils n'ont pas encore été suffisamment pris en compte et font l'objet de recherches en cours. Celles-ci incluent l'identification systématique des sources de variabilité et d'incertitude, de sorte que l'analyse de l'incertitude puisse être réalisée tout au long du processus d'évaluation de l'exposition.

De plus, les données et l'avis des experts devraient être combinées pour permettre la spécification des incertitudes pour les modèles, les paramètres et les scénarios. Enfin, l'analyse de sensibilité devrait être une composante de l'analyse de l'incertitude afin d'identifier les principales sources de variabilité, d'incertitude ou les deux et de faciliter le raffinement itératif du modèle d'exposition. Diverses approches sont disponibles pour effectuer des analyses d'incertitude et de sensibilité, telles que la régression, l'analyse de variance, les arbres catégoriels et de régression, le test de sensibilité en amplitude de Fourier et la méthode Sobol où de nombreuses entrées varient simultanément.

Données et méthodologie

L'exposition implique généralement le transfert de produits chimiques d'une source à travers l'environnement (air, eau, sol, nourriture) vers une personne ou une population. Les concentrations environnementales de B[a]P ont été obtenues à partir des bases de données de surveillance disponibles. Les données superposées provenant de plusieurs sources ont été sélectionnées en fonction de leur disponibilité, de leur représentativité géographique et de leur pertinence dans le cadre de l'évaluation de l'exposition et l'évaluation des inégalités environnementales.

Dans cet travail, les données relatives aux sources de polluants, aux rejets et aux concentrations dans l'environnement servent à établir des indicateurs des expositions humaines potentielles aux polluants. Des méthodes statistiques et géostatistiques sont développées pour spatialiser les concentrations moyennes annuelles ou annuelles de B [a] P en France dans les trois compartiments environnementaux (air, eau, sol) sur une grille de référence (grille de 9 km^2).

Un modèle multimédia, développé par l'INERIS (MODUL'ERS) et fonctionnant dans un environnement SIG, est utilisé pour quantifier l'exposition humaine à des substances toxiques et pour analyser les déterminants de l'environnement.

Les données sur le compartiment du sol ont été recueillies auprès de différentes sources. Les mesures des substances HAP sont disponibles via le réseau français de surveillance des sols (RMQS). La base de données contient des concentrations de 16 substances HAP mesurées dans la couche arable (0-30 cm) sur 1710 sites d'échantillonnage sur une grille systématique de 16 x 16 km sur les 550 000 km^2 du territoire métropolitain français. Les mesures qui n'ont pas pu être mesurées avec précision sont rapportées sous forme de valeurs inférieures à la limite de détection (généralement 0.01). Le pourcentage d'observations sous la limite de détection varie pour chaque substance.

Les données qualitatives sur les sites pollués sont disponibles dans la base de données BASOL (ministère français de l'écologie et du développement durable). Cette base de données est un inventaire des sites contaminés en France, réunis pour mener des actions préventives ou curatives. Il se compose de 788 points répartis sur le territoire français.

Enfin, 14 propriétés physicochimiques du sol sont issues de la base de données INRA.

Les covariables environnementales ont été mesurées sur 500 m de profondeur: couverture de la superficie forestière et semi-naturelle, éroclimat, érosion, pente, typologie, aspect de la pente, indice topographique composé, évapotranspiration moyenne, indice d'infiltration, code de matériau parent dominant de premier niveau, productivité primaire nette maximale, capacité d'eau disponible et rugosité.

Afin d'expliquer et de cartographier la variation spatiale des substances HAP dans le sol, ainsi que de prévoir les concentrations spatiales en France, les données ci-dessus seront utilisées comme covariables dans le processus de modélisation ou comme point de départ pour construire des variables auxiliaires pertinentes.

Le krigeage du résidu est utilisé pour interpoler les mesures de concentration entre les points mesurés. Cette méthode est populaire pour prédire les concentrations spatiales de polluants dans la cartographie des sols et des contaminations, car elle permet d'intégrer des informations secondaires provenant de variables auxiliaires dans la prédiction, en exploitant les relations linéaires entre les variables. Les covariables potentielles du modèle comprennent une variable indicatrice représentant la probabilité de dépassement du seuil de détection, fonction de la distance entre les mesures et les sites pollués et d'un certain nombre de propriétés du sol pouvant affecter la distribution spatiale des HAP dans le sol. La connaissance des processus contrôlant la variation spatiale de la propriété peut être incluse dans le modèle en sélectionnant les covariables appropriées. Par exemple, nous nous attendons à ce que la variation de B[a]P dans les sols français soit influencée par le contexte géologique, les domaines d'activités différentes, telles que l'agriculture, l'industrie et les mines et le transport et le dépôt de ces éléments. Par conséquent, des covariables reflétant ces processus, à savoir une classification du matériau d'origine, une classification de l'utilisation des terres, la précipitation annuelle moyenne et l'évapotranspiration potentielle annuelle moyenne, sont introduites dans le krigeage.

Afin de modéliser la probabilité que les mesures soient en deçà de la limite de détection, la méthode du krigeage par indicateur (IK) est utilisée. IK traite des valeurs inférieures à la limite de détection (DL) et prend en compte l'incertitude locale, en appliquant un krigeage ordinaire après une transformation de variable discrète non linéaire. De cette manière, IK estime les mesures de concentration des contaminants tout en calculant la

probabilité qu'elles soient inférieures à une certaine valeur limite.

Le krigeage ordinaire est ensuite utilisé pour modéliser les valeurs transformées. Les valeurs interpolées sont calculées comme la probabilité d'être inférieure à la limite de détection, par conséquent, elles se trouvent dans l'intervalle [0, 1].

La seconde variable explicative à inclure dans le modèle est la distance entre les lieux et les sites pollués pré-identifiés. La variable est construite à l'aide de la base de données BASOL. Tout d'abord, un tampon est construit autour des sites pollués. Ensuite, la distance entre le site pollué et les emplacements mesurés tombant à l'intérieur du tampon est calculée. Différents rayons de tampons sont testés en fonction des fonctions de distance (par exemple $1/d, 1/sqrt(d), 1/d^2$, ...) pour trouver la fonction de décroissance de distance inverse optimale.

Le rayon de tampon optimal ainsi que la fonction de distance sont sélectionnés pour minimiser le critère Akaike.

Une fois les variables auxiliaires construites, le krigeage du résidu est ensuite utilisé pour interpoler les mesures de concentration. Généralement, les résidus d'une régression linéaire sont interpolés à l'aide du krigeage ordinaire. De cette façon, il est possible d'inclure des informations provenant de variables explicatives externes dans le processus du krigeage.

Une alternative à la régression linéaire est l'utilisation d'un algorithme d'apprentissage automatique, tel que la méthode de forêt aléatoire. L'approche de régression des forêts aléatoires consiste à produire des arbres de régression multiples et à les combiner pour obtenir une seule prévision consensuelle. Les forêts aléatoires sont plus adéquates pour décrire des relations non linéarité entre les variables et se sont avérées plus efficaces que la régression linéaire (Hengl et al, 2015). De plus, ils n'exigent aucune hypothèse sur la relation entre la variable d'intérêt et les variables explicatives.

Les covariables potentielles à ajouter dans le modèle incluent les deux variables auxiliaires construites, en plus des 14 covariables environnementales. L'importance des variables est évaluée dans le cadre l'algorithme des forêts aléatoires, en sélectionnant les covariables qui minimisent le taux d'erreur dit « out-of-the-bag ».

Une fois l'algorithme des forêts aléatoires calé, un variogramme est modélisé en utilisant

un modèle sphérique avec les trois paramètres standard. La performance finale est évaluée avec une validation croisée.

Les concentrations de HAP dans le compartiment de l'air sont disponibles dans la base de données nationale française sur la qualité de l'air. Les données sont collectées en France pour chaque région dans le cadre de la surveillance réglementaire. La base de données disponible se compose de 76 et 77 mesures annuelles moyennes des points de concentration pour les années 2010 et 2011, respectivement. Les stations de surveillance sont classées en fonction de leur position : urbaine (domestique), suburbaine, rurale (agriculture), trafic (mobile) et industrielle.

Des concentrations atmosphériques sur une grille de 7 x 7 km en France sont également disponibles à partir du modèle de dispersion atmosphérique eulérienne CHIMERE. Le modèle multi-échelles CHIMERE est principalement conçu pour produire des prévisions quotidiennes de l'ozone, des aérosols et d'autres polluants et effectuer des simulations à long terme (saisons ou années entières) pour des scénarios de contrôle des émissions. Il s'étend sur une gamme d'échelles spatiales allant de l'échelle régionale (plusieurs milliers de kilomètres) à l'échelle urbaine (100-200 km), avec des résolutions allant de 1 à 2 km au 100 km. Le programme national et régional de surveillance et d'évaluation des émissions (TNO) et de registre (EMEP) est utilisé pour décrire les émissions dans le modèle de dispersion eulérien pour le territoire français.

Les émissions annuelles cumulées de trois substances PAH (benzo [a] pyrène, benzo [ghi] pérylène et indéno [1,2,3-cd] pyrène) pour 2011 et 2012 provenant de l'inventaire national spatialisé sont disponibles dans un support de $0.01°x0.01°$ grille en France. Il comprend les émissions de différentes sources industrielles et non industrielles (combustion, utilisation de solvants, urbain, trafic, etc.). De plus, des données sur la population résidant sur le territoire français, ainsi que sur l'altitude, sont disponibles dans une grille de 1 x 1 km.

Le problème principal avec les mesures disponibles est le nombre restreint d'observations. La spatialisation des concentrations sur la France par le krigeage ordinaire pourrait conduire à une représentation erronée du phénomène. Cependant, il est possible d'affiner l'estimation en incluant des informations provenant d'une variable auxiliaire. Différentes approches de krigeage dans un contexte multivarié ont été mises en œuvre en France et en Europe pour

cartographier les concentrations atmosphériques : régression linéaire suivie du krigeage résiduel, krigeage à la dérive externe, cokrigeage. Le Krigeage de dérive externe (KED), qui a été utilisé et évalué pendant de nombreuses années dans le système national de surveillance et de prévision de la qualité de l'air PREV'AIR (www.prevair.org) et qui a été mis en œuvre dans des projets associant pollution atmosphérique et effets sur la santé (Markus et McBratney, 2001 ; Bengtsson et Torneman, 2009) est retenu dans cette étude.

Soit Z la variable, connue seulement à un petit ensemble de points dans la zone d'intérêt, alors KED permet de prédire la variable dans toute la zone à travers une autre variable s, qui est connue de manière exhaustive. La variable Z est modélisée comme une fonction aléatoire Z (x) et s comme une variable déterministe s (x). Les deux variables devraient être liées linéairement.

Pour une prédiction spatiale précise, la dérive linéaire (variable auxiliaire) doit être bien définie. Pour construire une variable pertinente, il faut déterminer les facteurs qui influencent les mesures de concentration. Différentes variables auxiliaires possibles ont été testées afin de déterminer le meilleur modèle. Le modèle optimal a été sélectionné comme celui qui minimise l'erreur moyenne (ME) et l'erreur quadratique moyenne (RMSE) pour chaque type de station et pour la moyenne.

La variable auxiliaire est construite de la manière suivante : Les émissions sont désagrégées dans une zone tampon de 20 km autour des emplacements de mesure des HAP. Le rayon du tampon est soigneusement sélectionné en fonction de la corrélation décroissante entre les émissions et les mesures réelles, à mesure que la distance augmente. La variable auxiliaire est alors définie en fonction de l'émission, de la fonction de décroissance de la distance inverse, de l'altitude et de la population pour l'année 2010.

Une fois le modèle défini, le variogramme est ajusté. Une validation croisée est utilisée pour évaluer le résultat final du krigeage.

Les résultats des deux années sont ensuite combinés en une seule carte spatialisée afin de servir de base au modèle d'exposition. Pour cela, une moyenne pondérée est utilisée. Les poids sont calculés pour chaque année, en fonction de la précision des prévisions individuelles.

Les mesures de concentration de substances HAP dans l'eau potable en France, retenues dans la base de données française SISE'Eaux, sont collectées dans le cadre du suivi et évaluation de la qualité des eaux distribuées en France par le ministère français de la santé.

Les concentrations sont mesurées dans des unités de distribution d'eau (UDI) qui desservent l'ensemble des communes constituant le réseau de distribution d'eau en France. Le dernier est construit de telle manière qu'une seule unité de distribution peut desservir plus d'une municipalité, et une municipalité peut être desservie par plus d'une unité, générant ainsi 39062 combinaisons. La complexité du réseau doit être abordée avant de procéder à la spatialisation afin d'obtenir des résultats précis.

La base de données comprend 8 substances HAP : anthracène, B[a]P, benzo [b] fluoranthène, benzo [k] fluoranthène, fluoranthène et indéno [1,2,3-cd] pyrène, ainsi que 2 indicateurs de concentrations cumulées de 4 et 6 substances respectivement. Les mesures ont été collectées irregulièrement au cours des années 2000 et 2012. L'échantillonnage temporel non régulier sera pris en compte dans le processus d'estimation des moyennes pluriannuelles de concentration en HAP.

Comme c'est souvent observé dans les bases de données environnementales, le taux d'observation sous la limite de détection est assez élevé. Dans cette base de données, elle varie entre 70 et 95%. Un taux aussi élevé nécessite une manipulation minutieuse, afin d'extraire le maximum d'information des mesures disponibles, sans introduire trop de biais dans les résultats finaux.

La spatialisation des observations, y compris les valeurs inférieures à la limite de détection, est effectuée en spatialisant les concentrations moyennes pluriannuelles au niveau des municipalités (unité administrative européenne NUTS 4). Trois questions doivent être abordées:

- inclure les observations sous la limite de détection, sans introduire de biais important dans le processus;
- prise en compte de l'échantillonnage irrégulier dans le temps;
- estimation de la concentration pluriannuelle moyenne par commune, en tenant compte

de la complexité du réseau.

Les valeurs sous la limite de détection sont un problème récurrent en sciences de l'environnement. Différentes approches ont été proposées, la plus simpliste étant la substitution de sous la limite de détection à une constante. Cependant, il a été prouvé que ce type de méthode introduit un biais important dans l'estimation, ainsi qu'une sous-estimation de la variance lorsque le pourcentage de données censurées à gauche dépasse largement 5%.

Une alternative est l'utilisation de l'imputation multiple. Honaker (Honaker et al, 2011), propose un algorithme de "expectation maximization" (EM) basé sur le bootstrap pour effectuer l'imputation multiple. Cette méthode consiste à créer plusieurs ensembles de données «complets», qui peuvent ensuite être analysés en tant que cas complets. L'algorithme traite les données transversales des séries chronologiques en utilisant l'algorithme EM sur des échantillons bootstrap des données d'origine pour dessiner des valeurs à partir des paramètres de données complets. De cette manière, la corrélation entre ces substances peut servir à simuler les mesures de concentration sous la limite de détection, et en particulier pour les unités de distribution où toutes les mesures d'une substance sont inférieures à la limite de détection.

Un à priori uniforme peut être utilisé, afin que le modèle tient compte du fait que les observations simulées doivent tomber dans l'intervalle [0, DL].

Une fois les ensembles de données imputés calculés, l'estimation des concentrations pour chaque unité de distribution d'eau est effectuée en tant que moyenne pondérée pluriannuelle. Les poids sont calculés par la méthode des segments d'influence (Bernard, 2006). Selon cette méthode, une médiane est tracée entre deux événements voisins, puis un poids est attribué à chaque réalisation. Ce poids est égal au nombre de jours entre les deux segments inclus, divisé par le nombre total de jours entre la date de début (1/1/2000) et la date de fin (31/12/2012). Ensuite, la concentration des substances HAP pour chaque commune est calculée comme la somme des concentrations de chaque unité desservant la commune, pondérée par la population. Enfin, les résultats sont spatialisés par le géoréférencement des concentrations de chaque municipalité sur la grille de référence de 3 3 km.

Résultats

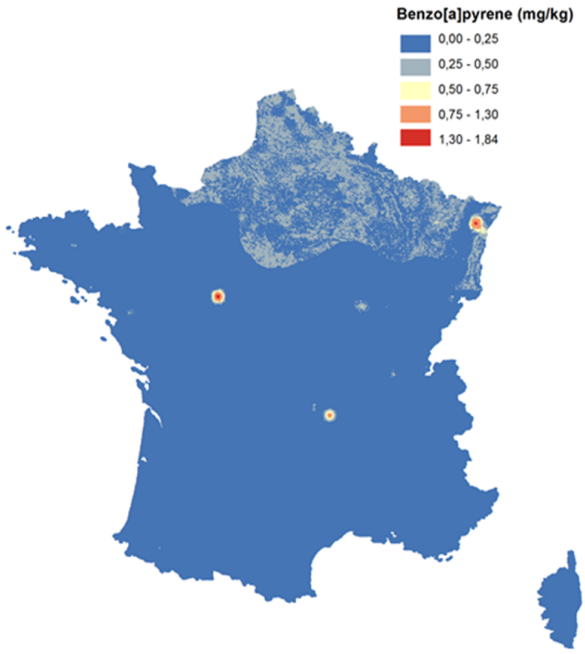


Figure 1: Spatialisation du B[a]P dans le sol (mg/kg) dans la grille 3×3 km.

Pour le sol, le modèle optimal est celui qui minimise l'erreur "Out-of-the-Bag". Le modèle des forêts aléatoires a été calibré pour s'assurer qu'il a une performance optimale. Il comprend 8 covariables: la variable indicatrice décrivant la probabilité de dépassement de la valeur seuil, la distance des sites pollués et 6 propriétés du sol incluant : couverture de la superficie forestière et semi-naturelle, rugosité, pente, aspect de la pente, moyenne évapotranspiration et éroclimat.

Des valeurs légèrement plus élevées sont observées dans le nord de la France, tandis que les valeurs les plus élevées se situent dans les points névralgiques initialement mesurés. L'impact de l'indicateur de krigeage en tant que variable auxiliaire est particulièrement évident.

Les mesures ponctuelles initiales des concentrations atmosphériques de HAP sont disponibles sur 76 sites pour l'année 2010 et sur 77 sites en 2011. Les concentrations les plus élevées mesurées correspondent principalement aux émissions industrielles. Un proxy d'émission est estimé en ajustant un modèle sur des données maîtrisées en utilisant une grille d'émission pondérée par la distance. Une matrice de voisinage est définie afin

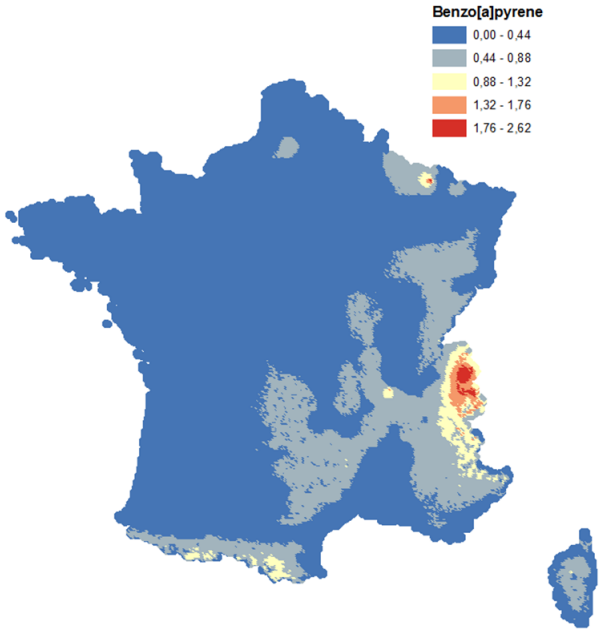


Figure 2: Spatialisation du B[a]P dans l'air (ng/m^3) dans la grille 3×3 km.

d'associer les émissions et les distances de grille à l'observation mesurée.

Les concentrations de B[a]P sont estimées sur le territoire français pour les deux années. Comme prévu, la concentration est plus élevée dans les zones de montagne (combustion du bois, dispersion atmosphérique limitée). Les résultats des deux années sont ensuite combinés en une seule carte, par une moyenne pondérée. Les points noircis correspondent aux zones montagneuses où l'incertitude due à l'altitude est trop élevée pour permettre une évaluation adéquate de la concentration.

Pour l'eau, l'estimation de B[a]P sous la limite de détection a été obtenue par la méthode d'imputation multiple. Les fortes corrélations entre les substances (85-90%) constituent une base solide pour appliquer une méthode d'imputation pour les séries chronologiques transversales (AMELIA II) (Honaker, 2013). Pour chaque unité de distribution d'eau, les corrélations importantes entre les substances et les tendances temporelles sont également prises en compte pour l'estimation des valeurs manquantes. Un minimum de 5 ensembles imputés est généralement suffisant, mais en raison du pourcentage très élevé de valeurs

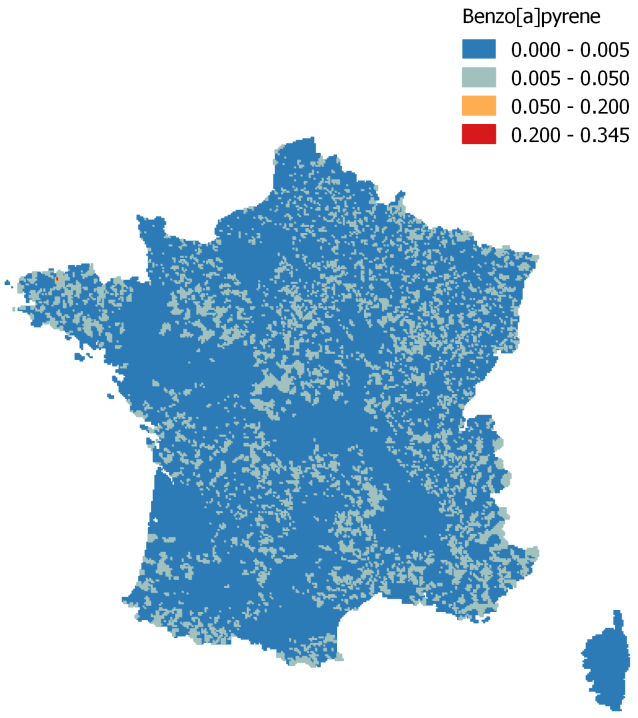


Figure 3: Spatialisation du B[a]P dans l'eau ($\mu g/l$) dans la grille 3×3 km.

inférieures à la limite de détection, 8 ensembles imputés sont obtenus à la place.

Une fois les ensembles de données imputés estimés, la concentration pondérée pluriannuelle pour chaque unité de distribution est calculée, de même que la concentration pour chaque commune, pondérée par la population desservie par chaque établissement dans chaque commune. Enfin, les concentrations sont géoréférencées sur la grille de 3 x 3 km pour être incluses dans le modèle spatial.

La base de données spatiales précédemment assemblée sert d'entrée pour le modèle d'exposition multimédia. Les résultats du modèle permettent l'analyse des déterminants de l'exposition par l'estimation des contributions des différentes voies d'exposition, des compartiments environnementaux et des transferts. Ils peuvent également être utilisés pour identifier pour quels polluants les indicateurs de risque sont les plus élevés et pour identifier les populations dans les zones potentiellement surexposées. Enfin, ils permettent de cartographier les inégalités environnementales dues à l'exposition grâce à la spatialisation des indicateurs de risque.

L'analyse de la dose journalière moyenne permet d'évaluer la variabilité des contributions pour chaque milieu d'exposition. La variance entre les valeurs modélisées de la dose journalière moyenne résulte des différents scénarios d'alimentation (facteur d'autoconsommation, type d'aliment, quantité ingérée).

La principale voie d'exposition du B[a]P pour les deux classes est l'ingestion d'aliments commerciaux qui ne comprennent pas les légumes. La concentration de B[a]P dans ces produits est considérée comme identique dans la zone d'étude et stable pour la période examinée. La deuxième contribution la plus importante provient de l'eau potable, suivie de de près par la consommation de légumes. Enfin, l'ingestion de sol a l'impact plus faible sur le DJE.

L'utilisation du SIG permet de cartographier les inégalités environnementales d'exposition, sur une résolution fine, en spatialisant l'ERI. La structure spatiale des cartes de risques reflète l'influence d'un ensemble complexe de facteurs spatiaux et environnementaux avec de grandes variations et opère à différentes échelles spatiales. L'analyse de ces structures permet de quantifier la portée et la contribution des différentes échelles de variabilité spatiale.

Discussion

La caractérisation des inégalités environnementales repose sur des hypothèses spécifiques, qui influencent et limitent le choix des données. Les données qui ne sont pas disponibles au niveau de l'agrégation de l'étude sont ventilées dans une résolution plus fine. De plus, des variables supplémentaires sont construites à partir des données disponibles pour augmenter la résolution. Enfin, un traitement supplémentaire est nécessaire pour homogénéiser toutes les données disponibles sur la même unité spatio-temporelle.

Les mesures de concentration effectuées sur des sites d'échantillonnage ne constituent pas une base solide pour spatialiser directement les concentrations de HAP dans le sol. Par conséquent, des informations supplémentaires sont utilisées pour accroître la représentativité des données De cette façon, les emplacements où les mesures ne sont pas pas connues de manière exhaustive peuvent être inclus dans le processus de spatialisation, au lieu d'être ignorés. Dans la carte finale des concentrations de B[a]P dans le sol, on peut en déduire

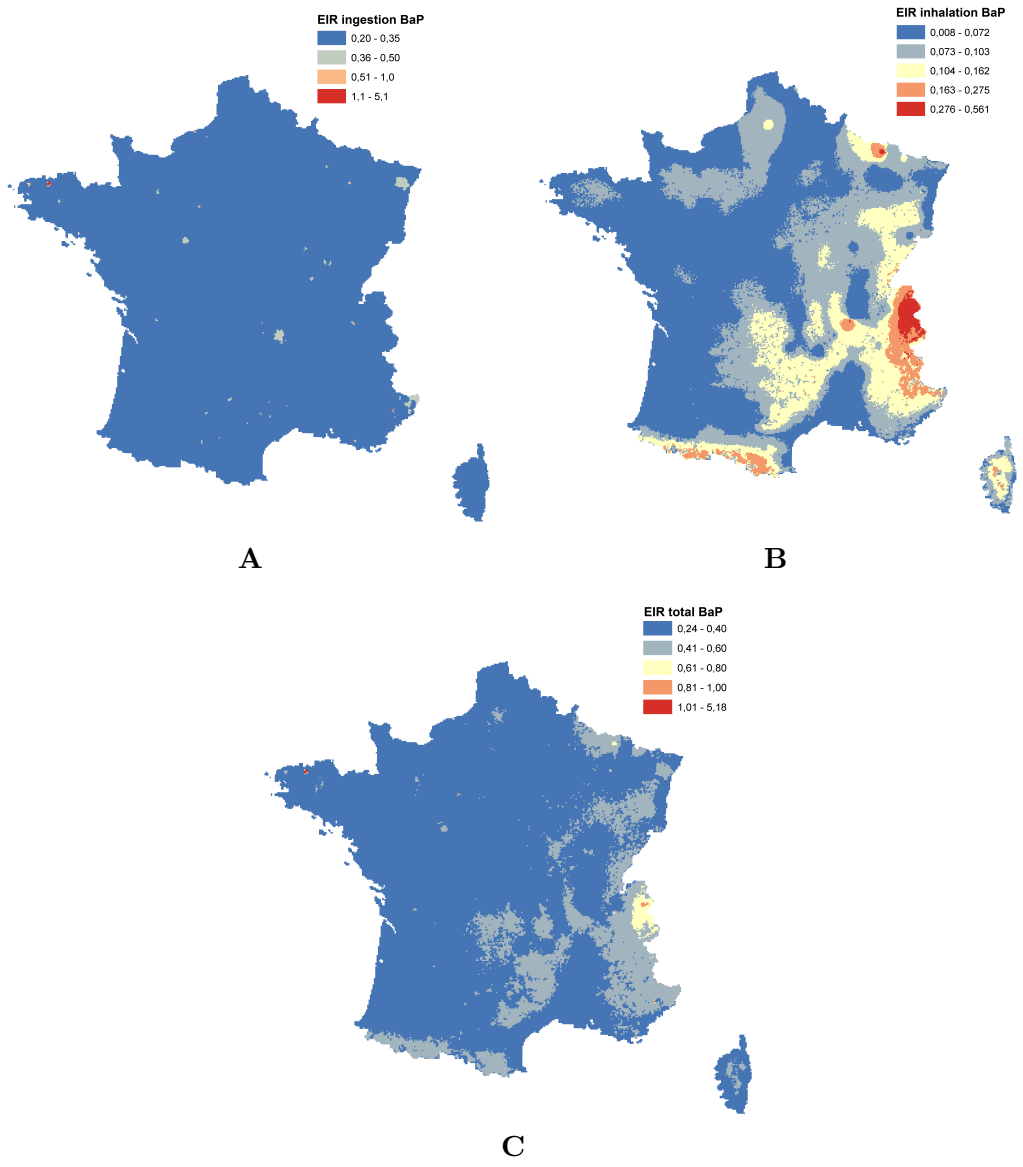


Figure 4: Cartographie des ERI pour B[a]P: (A) ERI dû à l'ingestion; (B) ERI dû à l'inhalation et (C) ERI total.

que les concentrations dans le nord de la France ont tendance à être supérieures aux concentrations observées dans le centre et le sud-ouest de la France.

Le krigeage résiduel fournit le cadre pour inclure des informations provenant des variables auxiliaires dans la prédiction, augmentant ainsi la représentativité des données et la robustesse de la prédiction. Les 16 variables sont combinées dans un modèle géostatistique pour prédire spatialement la distribution des HAP. L'inclusion de seulement 8 variables les plus significatives au lieu des 16 a amélioré la performance des indicateurs de prédiction du modèle (R^2 et cv). Ceci s'explique du fait que en incluent de variables non pertinentes peut ajouter du bruit au modèle et ne pas nécessairement améliorer la prédiction. Le fait que la couverture et la pente de la forêt en tant que facteurs prédictifs des concentrations de B[a]P soient conservés comme parmis les variables les plus importantes concorde avec la biochimie de l'HAP. Cela suppose que les HAP, en particulier en phase gazeuse, sont affectés par les dépôts secs et les précipitations - lorsqu'ils se déplacent vers un environnement plus froid et à plus haute altitude ils sont piégés par le processus de condensation froide et se déposent à la surface du sol.

Les sols enrichis en matière organique agissent comme un puits environnemental de polluant organique. Suite à la spatialisation des B[a]P, on constate des concentrations plus élevées dans le nord de la France ainsi que des points névralgiques autour de certains sites pollués. Des tendances similaires ont été observées pour d'autres substances HAP dans le nord et le nord-est de la France et correspondent à leurs niveaux les plus élevés, tandis que les concentrations sont plus faibles dans le sud-ouest et le sud-est. Les concentrations particulièrement faibles dans la zone sud-ouest pourraient résulter de la composition des sols, ce qui pourrait entraîner la lixiviation du B[a]P hors du profil.

Des études récentes ont montré que les émissions sont positivement corrélées à la distribution des HAP dans le sol. L'inclusion des émissions en tant que variable auxiliaire pourrait encore améliorer la qualité de la modélisation.

En ce qui concerne le compartiment air, deux modèles différents ont été construits pour chaque année, puis combinés en une seule carte. Le principe de distance-decay a été utilisé pour construire une variable auxiliaire à partir de l'inventaire des émissions. Il a été démontré que les HAP se concentrent dans des environnements plus froids à des

latitudes et à des altitudes plus élevées. L'altitude est donc considérée comme une variable potentiellement pertinente pour améliorer la qualité de la modélisation. En effet, pour l'année 2010 elle a été inclue dans le modèle, ce qui a augmenté sa précision. La population a été insérée dans le modèle pour limiter l'effet de l'altitude sur les zones montagneuses peu habitées. Pour l'année 2011, aucune corrélation forte n'a été observée. Cette différence entre les deux années pourrait s'expliquer par les différences de conditions climatologiques.

Puisque le krigeage de dérive externe est une méthode largement influencée par les observations initiales, l'incertitude relative dépend du caractère parcimonieux du réseau de mesure. La qualité des cartes de concentration d'air repose principalement sur la qualité des données d'entrée utilisées, à savoir : la qualité des observations, les résultats modélisés et les informations auxiliaires introduites dans la méthode d'interpolation spatiale.

Un modèle plus raffiné pourrait mieux répondre aux différences sous-jacentes dans l'exposition. Dans ce contexte, différents modèles de dispersion de distance pourraient être adaptés aux deux cas. Les conditions climatiques soumises à la saisonnalité ou à d'autres spécificités ont montré une influence sur l'exposition aux HAP.

Les mesures de concentration disponibles des substances HAP pour l'eau potable ont été prises dans le réseau de distribution d'eau en France. Reconstruire les concentrations sous la limite de détection n'est pas anodin, surtout lorsque le taux d'observations est élevé. La méthode d'imputation choisie a donné des résultats encourageants, car le nombre d'itérations pour les substances manquantes est assez similaire pour chaque ensemble, ce qui indique la stabilité de l'algorithme. Pour tester davantage ses performances, la méthode a été appliquée à un deuxième jeu de données qui a été artificiellement imputé. Les résultats obtenus ont montré que la méthode d'imputation multiple surpassait la méthode de substitution à valeur constante.

Pour les districts où aucune mesure n'était disponible, les concentrations étaient considérées comme étant 0.Cependant, cela pourrait avoir un impact sur la distorsion de la voie d'exposition, en sous-estimant la contribution de l'eau potable. L'incertitude associée au processus d'imputation est considérée sous forme de variance entre les ensembles de données imputés.

B[a]p se trouve généralement à des concentrations plus élevées que les autres substances HAP et est couramment utilisé comme indicatrice de contamination par les HAP en raison de sa valeur toxicologique de référence supérieure et contribue à environ 50% du potentiel cancérigène total du groupe HAP. Bien que les HAP non métabolisés puissent avoir des effets toxiques, la toxicité de leurs métabolites réactifs constitue une préoccupation majeure. Une limitation de la méthode employée ici est qu'elle ne permet pas de capturer la toxicité potentielle due à la formation de ces métabolites réactifs des substances HAP.

L'estimation des concentrations dans les milieux d'exposition à partir des compartiments environnementaux nécessite le couplage de certains compartiments, selon le choix des équations, afin de simplifier le calcul des transferts entre les matrices. Dans cette modélisation, seuls les transferts verticaux sont pris en compte. En particulier, le transport de polluants d'un maillage à un autre n'est pas pris en compte. Le transport atmosphérique est intégré à partir de la modélisation des concentrations.

Le compartiment du sol est couplé au dépôt atmosphérique, mais le reste des compartiments est considéré comme indépendant et statique sur la période considérée. Les milieux d'exposition sont estimés à partir du couplage d'un ou deux compartiments environnementaux. Certaines simplifications dans l'approche de modélisation pourraient introduire un biais dans les estimations. Les transferts ne sont estimés que dans un sens, d'une matrice à l'autre, sans tenir compte des échanges de retour. Cela a été montré pour conduire à une surestimation des transferts. De plus, même si le découplage des compartiments environnementaux facilite la mise en œuvre et simplifie le système d'équations et les hypothèses initiales, il limite également la dimension dynamique du modèle, i.e. l'estimation de l'évolution temporelle des concentrations dans l'environnement. Dans le cas des cancérogènes, pour lesquels les risques sont caractérisés à partir des moyennes sur la période pertinente des concentrations dans les compartiments environnementaux, la caractérisation de l'échelle temporelle est nécessaire. Cette approche permet d'améliorer la résolution et la représentativité des données.

Ces méthodes nécessitent la mobilisation d'un grand nombre de données et de traitements, rendant difficile la caractérisation des incertitudes.

En résumé, l'approche multimédia permet d'inclure les concentrations des différents

compartiments environnementaux et de leurs interconnexions. L'évaluation de l'exposition a fourni un cadre méthodologique et la définition d'une mesure intégrant toutes les expositions des principaux médias à la construction de différents indicateurs basés sur les quotients de risque. Enfin, le SIG a permis l'interopérabilité de différentes variables d'intérêt pour l'évaluation des expositions.

Les résultats finaux ont montré que la principale voie d'exposition de B[a]P est l'inhalation. En ce qui concerne l'ingestion, les produits commerciaux ont la plus forte contribution à la dose totale d'exposition, l'eau potable et les légumes ayant des contributions similaires. Le sol semble avoir la plus faible contribution pour les deux groupes d'âge. En examinant de plus près les habitudes de consommation de deux zones d'étude, on observe des différences dans les contributions locales et nationales. Malgré ces différences, les contributions individuelles de chaque classe de légumes restent proportionnelles aux deux cas. Certaines des surestimations peuvent être expliquées par les données utilisées pour construire les coefficients de transfert. Aucune remise en suspension n'a été prise en compte pour ces estimations.

Les modèles d'exposition spatiale prennent en compte la source du polluant et de l'environnement pour déterminer la force et la propagation de l'exposition. Ces modèles ont besoin de données d'entrée précises, aussi géographiques que liées à la source, car l'incertitude des données d'entrée peut avoir un effet important sur les prévisions. Des efforts sont faits pour sélectionner des ensembles complets, exactes et actualisés de données pour inclure dans le processus d'évaluation de l'exposition. Cependant dans ce processus, les hypothèses et les décisions d'analyse des données ont été effectuées lorsque les données sont rares ou indisponibles. Ces hypothèses et décisions entraînent des incertitudes et des limites dans les conclusions à tirer.

Dans cette travail, le principal intérêt était de caractériser la représentativité de la base de données spatialisée, servant comme entrée du modèle. Différentes limitations ont été imposées par la qualité des bases de données disponibles pour chaque compartiment environnemental. Pour le compartiment aquatique, malgré la pluralité des données disponibles, la majorité des valeurs étaient inférieures à la limite de détection, rendant difficile la spatialisation directe des mesures. Pour y remédier, la méthode d'imputation

multiple choisie prendre en compte l'incertitude et permet ainsi de générer des valeurs plausibles pour chaque point de données non observé.

De cette manière, l'incertitude peut être rapportée sous forme de variance entre les ensembles de données imputés. De plus, en raison de la complexité du réseau d'eau en France, l'évaluation de la population desservie par chaque unité de distribution implique une hypothèse supplémentaire dans le processus de modélisation. En ce qui concerne le compartiment air, la spatialisation du nombre limité et restreint d'observations pourrait conduire à une classification erronée de l'exposition. L'utilisation de propriétés physico-chimiques du sol telles que la topographie permet de prendre en compte l'hétérogénéité régionale qui peut influencer les concentrations de HAP dans le sol.

Une perspective de ce travail serait de permettre le couplage de l'approche spatiale avec la modélisation multimédia, ce qui permettrait de caractériser l'incertitude totale. Un premier examen consisterait à analyser les propagations d'erreur selon les différentes étapes du traitement et de la modélisation des données.

Conclusion

L'évaluation de l'exposition en tant qu'outil d'identification et de caractérisation des inégalités environnementales implique des exigences et des contraintes supplémentaires qui ont une influence directe sur le choix des données et de la modélisation. Une évaluation complète de l'exposition afin de cartographier les inégalités environnementales à une échelle fine nécessite l'intégration de données provenant de différentes sources et opérant souvent à différentes échelles spatio-temporelles. Le choix des paramètres habituellement utilisés dans l'évaluation du risque d'exposition pourrait générer des distorsions entre les contributions et l'exposition ou les voies d'administration, ce qui empêcherait la comparaison des contributions entre celles-ci et l'analyse des déterminants de l'exposition. Une révision du modèle d'exposition a donc été effectuée pour remplacer les principaux paramètres par des paramètres plus réalistes, tels que les facteurs de transfert.

La présente étude a nécessité un travail préparatoire important en termes d'acquisition de données, de traitement des données et de mise en œuvre d'outils de modélisation. L'évaluation de l'exposition dans le contexte de la cartographie des inégalités environnemen-

tales implique le développement des approches intégrées capables de prendre en compte toutes les principales voies d'exposition, l'intégration des sources passées et présentes, locales et mondiales, ainsi que la construction de scénarios réalistes.

La méthodologie a été appliquée pour trois Hydrocarbures Aromatiques Polycycliques (benzo[a]pyrène, benzo[ghi]pérylène et indéno[1,2,3-cd]pyrène) sur l'ensemble du territoire français. Les résultats permettent de cartographier des indicateurs d'exposition, d'identifier les zones de surexposition et de caractériser les déterminants environnementaux. Dans une logique de caractérisation de l'exposition, la spatialisation des données issues des mesures environnementales pose un certain nombre de questions méthodologiques qui confèrent aux cartes réalisées de nombreuses incertitudes et limites relatives à l'échantillonnage et aux représentativités spatiales et temporelles des données. Celles-ci peuvent être réduites par l'acquisition de données supplémentaires et par la construction de variables prédictives des phénomènes spatiaux et temporels considérés.

Les outils de traitement statistique de données développés dans le cadre de ces travaux seront intégrés dans la plateforme PLAINE pour être déclinés sur d'autres polluants en vue de prioriser les mesures de gestion à mettre en œuvre.

Contents

List of Tables

List of Figures

Definitions

Administration route	Route through which a substance enters the body. There are three different routes of administration: inhalation, ingestion and skin contact that can be differentiated according to the transfer medium involved: - inhalation of a substance in gaseous state; - direct ingestion of soil, food (plants grown on site, animals raised on the site), contaminated water; - cutaneous absorption by contact with contaminated soil, dust and/or water (bath, shower, nautical activities, ...).
Aggregated population data	Population data available at the local level (such as district) or a large territory (such as region or country)and aggregated according to one or several variables. The aggregation variable can be a geographic region, or other variables such as socio-economic status or a given period of time.
Average Daily Dose	Dose (internal or external) of a substance received by the body, relative to the weight of the individual and the number of days of exposure (in the case of a non-carcinogenic substance) or to the number of days of the whole life (in the case of a carcinogenic substance).

Biomarker	A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.
Chronic Exposure	Persistent, continuous or discontinuous exposure, occurring over a long period of time, ranging from a year to a lifetime.
Environmental indicator	Variable used to assess the state of the environment (such as concentration of pollutant).
Environmental media	Environmental media refers to air, groundwater, surface water, soils, sediment.
Exposome	Comprehensive and integrated view of the history of exposure to biological, chemical and physical agents during lifetime including the prenatal period (Wild, 2005).
Exposure concentration	Concentration of a chemical agent in the medium at the point of contact with an individual.
Exposure media	The media in contact with the target studied. In the case of human populations, this may include: indoor air, tap water, topsoil layer, but also food.
Exposure route	Pathway of a substance from the source to a target. An exposure route includes a source, a point of exposure and a route of administration. If the point of exposure differs from the source, there is also a propagation mechanism and an intermediate compartment where the pollutant is transported.

External Dose	Quantity of a substance in contact with the body's barriers (intestinal walls, pulmonary alveoli, skin). It is usually expressed as the mass of substance per unit of body weight and per unit of time.
Geographic Information System (GIS)	A computer system for organizing, analyzing, combining and presenting geographically located information from a variety of sources.
Geographic analysis resolution	Level of geographic detail used for the analysis (example: mesh, IRIS, district, canton ...).
Internal Dose	Quantity of a substance in the body. It is estimated by measuring substances and their metabolites in the blood, urine, saliva or tissue.
Kriging	Linear interpolation method in geostatistics ensuring minimum variance.
Spatial Analysis	The process of examining the locations, attributes, and relationships of features in spatial data through overlay and other analytical techniques in order to address a question or gain useful knowledge.
Specific study population	Population that can be grouped according to a criterion (age, sex...) or with a particular vulnerability or risk (elderly, children...).
Study period	The period of time during which the data is studied.
Toxicological Reference Value (TRV)	Values to establish a relationship between a dose and an effect (dose-threshold effect) or a dose and a probability of occurrence of an effect (effect without threshold). TRVs are specific for an effect (usually critical), a duration of exposure (acute or chronic) and a route of exposure (oral or respiratory).

Transfer	Migration of dissolved or non-dissolved substances in one or more media (such as through or on the surface of a soil, water, air and human activities, or by soil organisms).

General Introduction

Environmental inequalities refer to the recognition of disparities between territories or populations concerning environmental exposure. In some areas, people are more likely to be exposed to the negative effects of air, soil or water pollution. This issue has been identified as a priority for French policies, as evidenced by the emergence of the French national plans for health and environment. Specifically, the third plan (2015-2019) highlights the need to establish a methodology to identify and reduce environmental health inequalities.

With the increase in geographical data, environmental quality data are often available and could be used to determine exposure disparities. However, at a regional scale and fine resolution of exposure outcome prerequisite, environmental monitoring networks are not sufficient to characterize the multidimensionality of the exposure concept. In an attempt to increase representativeness of spatial exposure assessment approaches, spatial exposure indicators could be built using additional available databases and theoretical framework approaches to conjugate statistical and deterministic models.

Human exposure to the substances released to the environment depends on the population's contact with water, air, soil and contaminated food and concentrations of substances in these environmental compartments. The multiplicity of exposure routes, the need to take into account long-term exposures, and the difficulty to obtain direct measurement of individual human exposures contribute to a significant need for modeling in this area. As a result, multimedia models are used to assess exposure because they provide an appropriate framework for understanding the interactions between chemicals and the environment.

There must be a tight coupling of exposure model, with geospatial analysis and spatial statistic techniques to facilitate the characterization of environmental inequalities and enable integrated exposure assessment of population living on territories. In this way, the fundamental prerequisite is combining contaminant source and environmental concentration database, exposure model and population behavior with spatial approaches. That implies the need to a) define interest variables and indicators that could be built to associate and

describe the global source-exposure-effect chain and b) develop adapted methods in order to improve spatial data representativeness and resolution.

The main objectives of the present PhD research project are:

- to integrate and process various levels of data from different sources;
- to develop statistical methods for the spatial representation of exposure indicator and the analysis of environmental inequalities;
- to build a multimedia environmental exposure approach suitable to substances of interest and able to represent realistic exposure pathways;
- to characterize spatial phenomena that operate at different spatio-temporal scales;
- to detect environmental inequalities and identify the major sources of uncertainty by characterizing data representativeness and transfer contributions.

The above issues are addressed by employing statistical and geostatistical methods to spatialize concentration measurements of Polycyclic Aromatic Hydrocarbons (PAH) in France in the three environmental compartments (air, water, soil).

For feasibility reasons three substances are selected: Benzo[a]pyrene, Benzo[ghi]perylene and Indeno[1,2,3-c,d]pyrene. All three substances have been identified by the United States Environmental Protection Agency (US EPA) as priority PAH, due to their carcinogenicity or genotoxicity. The geographic area considered is the totality of the metropolitan French region. Data from the three environmental compartments (water, air, soil) are available from different sources, and will be spatialized with respect to their specificities.

The first chapter introduces the context in which this thesis is realized. In a first part, we discuss the connection between the environment and health in public policies in France, Europe and in the world. In the second part, we present the overview of environmental geographical inequalities, the methodological framework to characterize them, the concept of exposome and the substances of interest.

The second chapter presents the different scientific issues of the approaches used to characterize environmental inequalities. This chapter deals with the different types approaches developed to detect areas or potential overexposed populations. Examples of exposure assessment framework developed to characterize environmental inequalities examples are presented. Spatial analysis techniques and multimedia exposure modeling are described. This part also presents, in the framework of the characterization of the

environmental inequalities: the Geographic Information Systems (GIS), as a computer system designed to support the capture, management, manipulation, analysis, modeling and display of spatial referenced data that provide a mathematical framework for manipulating spatially-referenced data.

The third chapter addresses the methodological choices to address the constraints of the exposure assessment framework in the context of environmental inequalities. Available database, equations, transfers and exposure pathways considered in the model are presented. Furthermore, the spatial data workflow employed to construct the model's input is explained. The issue of data representativeness impact on the exposure assessment is also addressed here.

The fourth chapter presents the results of the work. In the first section the results of the spatialization methodology employed for each environmental compartment are presented. The spatialized dataset then serves as an input for the multimedia model. The modeling results are presented in the second section.

The fifth chapter discusses the final results of the thesis. The principal developments and processes applied on the available data in the context of environmental inequalities characterization are discussed. The limitations and results concerning spatial and temporal support as well as data homogeneity and resolution are presented. The coupling approaches put in place to combine the different data are also detailed: spatial data, environmental compartments, exposure pathways. Finally, exposure assessment results are analyzed and compared to other studies and uncertainties are discussed in order to provide efficient exposure map to prioritize public action.

In the sixth chapter a conclusion is presented by recalling the main methods and tools developed, the results obtained, their limits and the potential for improvement. The different developments and possible uses of this study are defined in the perspectives.

Chapter 1

Introduction and Context

In this section, we will present the context in which this thesis is realized. In a first part, we will discuss the connection between the environment and health in public policies in France, Europe and in the world. In the second part, we will present the overview of environmental geographical inequalities, the methodological framework to characterize them, the concept of exposome and the substances of interest.

1.1 Public policies on the spatial and environmental inequalities

1.1.1 In France

The national plans for health and environment (PNSE)

PNSE 1

The analysis of the linkage between health and environment has become a major preoccupation for French public health as evidenced by the emerging national plans for health and environment (Plan National Santé Environment). The first plan included 45 actions that the French government decided to set in motion for the years 2004-2008, in order to reduce the impact of the environmental degradation to French people's health. It is part of a common European effort that deputed in the early 2000 and is supported by the World Health Organization (WHO) and local authorities.

The first PNSE underlined the necessity to develop new methods to link the sanitary

and wellness data with environmental factors. More specifically, the plan called for an improvement in knowledge and deeper understanding of the sanitary impact of the various environmental factors, while introducing approaches allowing the public authorities to manage efficiently the health environmental policies. Moreover, in order to assess the population's exposure to the various pollutants and other risk factors as well as to measure the impact of the different environmental compartments to human health, it suggested the closer surveillance of emissions with the use of available tools (such as information systems,..). Finally, in the plan is advised the evolution and customization of hypothesis when modeling data from different sources (epidemiology, monitoring of the quality of the environment, application of the regulation,...) and studying their potential relations, as the information systems used so far are subjected to constraints and restrictions.

The plan was gradually declined in a regional level, in a multidisciplinary and transdisciplinary manner, seeking to better take local issues into account.

PNSE 2

The second national plan for health and environment (PNSE 2) adopted in 2009 for the years 2009-2013, resumes the commitments initial issued in the first plan. It explicitly draws up on the proposals for close cooperation of concerned parties, the public and local authorities, associations and experts. Two greater axes are identified in the context of PNSE 2:

1. Reducing the exposures responsible for the high impact pathologies in human health, such as respiratory conditions and cancers.

2. Reducing the environmental inequalities.

The need to identify and manage the geographical zones for which an overexposure to toxic substances is observed is also highlighted. Each region drew up a Regional Environment and Health Action Plan (PRSE) to implement the main objectives of the PNSE according to its own specific needs. Different regions in France - as Lorraine - have included the environmental health inequalities reduction in their plan, and need assessment to guide priorities for voluntary action. Here environmental health inequalities are seen as the unequal geographic distribution of exposure.

PNSE 3

After more than 10 years of actions aimed to the prevention of environmental health risks the third plan (PNSE 3, 2015-2019), consolidates the progress already achieved but also proposes a new approach to environmental health, both robust and more connected to the territories while integrating the development of new scientific concepts like that of the exposome. The recently emerged term of exposome is used to describe these complex exposures, taking into account all sources, routes, and - when possible - the interactions of pollutants, that are likely to contribute to the health alteration of individuals. That way, human health can be considered to depend on two main components: the genome and the exposome.

The challenge is to tackle the particularly complex health and environmental problems. Meeting this challenge requires mobilizing all public power through the various policies (energy, planning, urban planning, transport, industry, research, agriculture, etc.). Implementation of the PNSE 3 requires action in collective (including pollution reduction) and individual prevention. Among the policies concerned, a major effort in the field of research is needed to improve our knowledge of the impact of the environment and, in the broad sense of the term, on human health. This action is essential to assess the fraction attributable to the environmental factor for certain pathologies.

A total of 107 actions are prescribed to address the health priorities, exposure knowledge environmental health and information and communication issues. 4 actions are specifically dedicated to the environmental inequality thematic:

- Action no 38: develop and disseminate, via a common platform, reference methodologies at the national level for the characterization of locally disparate environmental inequalities, taking into account the vulnerable situations of the populations

- Action no 34: identify and analyze the methods of constructing spatialized and integrated exposure indicators

- Action no 39: using tools for analyzing environmental inequalities to cross exposure models and population data (biomonitoring, epidemiological data, social and health vulnerabilities)

- Action no 40: implement in the context of the PRSE multi-exposure studies in several territories, based on methodological references.

1.1.2 In Europe

Environment and Health Strategy

The Environment and Health Strategy was adopted by the European Commission in June 2003 in order to encourage effective policy making regarding environment and health issues. The three objectives identified for this strategy were:

- Reduce the disease burden caused by environmental factors in the EU.
- Identify and prevent new health threats caused by environmental factors.
- Reinforce the EU capacity for policy-making in this area.

The first cycle, also referred to as the SCALE (Science, Children, Awareness, Legal instrument, Evaluation) initiative, was set in motion for the years 2004-2010 with the aim of studying and understanding the complex interactions between health and environment so that the impact of environmental factors to human health can potentially be reduced. The first cycle targeted the link between the environment and childhood respiratory diseases such as asthma and allergies; neurodevelopmental disorders; childhood cancer and disruption of the endocrine system (glands which secrete hormones). A total of 13 actions were proposed, covering the following three thematics:

- Preparation of environmental health indicators.
- Development of methodological approaches to analyze interactions between environment and health.
- Promote the development of a tool permitting the detection and treatment of emerging risks in environment and health.

This action plan constitutes the contribution of the European Commission to the 4rth Ministerial conference on the subject of environmental health, organized in Budapest by the WHO in June 2004.

Second Program of Community Action in the Field of Health 2008-2013

The second program intended to complement, support and add value to the policies of the Member States and contribute to increase solidarity and prosperity in the European

Union by protecting and promoting human health and safety and by improving public health.

The overarching objectives were:

- Improving citizens' health security by proactively preparing against diseases and health threats from various sources (chemical, physical, biological) while promoting patients' safety through relative actions (healthcare, scientific advice and risk assessment).
- Promoting health and reducing environmental health inequalities in an effort to increase healthy life years and aging.
- Generating and disseminating health information and knowledge.

Specifically, the actions related to studying the potential effects of environmental risks as wells as the analysis of the various environmental factors to human health contribute, when possible, to the prevention of major diseases and the reduction of their morbidity.

EU Health Program 2014-2020

Continuing Union's action in the field of health, the third program was adapted in March 2014. New challenges are being addressed in this program in order to help the EU countries respond to economic and demographic challenges facing their health systems and enable citizens to stay healthy for longer. The four principal goals are:

- Promoting health and preventing diseases.
- Protecting citizens from cross-border health threats.
- Enhancing health systems' capacity building.
- Facilitating citizens' access to better healthcare.

Once again it is particularly noted the need of actions in the field of communicable diseases and health threats specifically caused by biological and chemical exposure.

World Health Organization (Europe section)

Environment and health is an important working axe for World Health Organization. In 2017, 27 years are completed since European countries initiated the first ever process to

eliminate the most significant environmental threats to human health. Progress towards this goal is driven by a series of ministerial conferences held every five years and coordinated by WHO/Europe.

The uneven distribution of population exposure to environmental conditions and the likely diseases that result is strongly linked to a range of socio-demographic determinants. Environmental health interventions to address these inequities need to be based on an assessment of their magnitude and the identification of population groups most at risk or vulnerable to the environmental hazards. However, the data on this subject are not always available. To fill this gap, and to follow up on the commitments made at the Fifth Ministerial Conference on Environment and Health in 2010, the WHO Regional Office for Europe carried out an evaluation in 2012 the scale of inequality in environmental health in the European region, based on a set of 14 indicators of inequality [226].

1.1.3 World: Environmental Public Health Tracking Program

Integrated environment and health surveillance systems have recently been developed at international level. The Californian Policy Research Centre (CPRC), WHO-Europe, the Institut National de Santé Publique du Quebec and the CDCs (US Centers for Disease Control and Prevention), have produced comprehensive reports on the strategy to implement an for environment and health monitoring.

In January 2001, the US Environmental Health Commission called for the creation of a coordinated public health system to monitor and combat the effects of environmental degradation in order to further reduce health damage to people in the United States. In response, the US Congress has allocated funding to the CDC to develop the National Environmental Public Health Tracking Program (EPHT).

The aim of this program is to establish a national health and environmental monitoring network and provide the necessary information to support actions that improve the health of the population. This network is based on a wide range of stakeholders: federal, state and local health actors, environmental organizations, non-governmental organizations (NGOs), schools of public health, etc. It is based on systematic collection, integration, analysis, interpretation and dissemination of environmental data within the network to identify areas and populations most likely to be impacted. It also aims to provide information to public health managers and researchers to examine the possible relationships between health and the environment. Information can thus be used to drive public health policies in different administrative levels of management. For this program, spatial approaches and GIS have

been identified as the preferred tools.

1.2 Scientific context

1.2.1 Environmental Inequalities

The term "Environmental Inequalities" introduced in France in early 2000, is closely intertwined to the term of "Environmental Justice", initially appeared in United States, and eventually adopted worldwide. The concept of "Environmental (in)Justice" emerged in the early 1980s in order to describe the fact that certain communities or human groups were disproportionately more susceptible to be subjected to higher levels of environmental risks than other segments of society.

The unfair environmental burdens and evidence of racial and economic injustices led to a grassroots civil rights campaign [28]. In early 1990, the concept was taken up by philosophers, and later by sociologists, geographers, economists and politicians, leading to the globalization of the Environmental Justice Movement. Theory and practice of environmental justice necessarily includes distributive conceptions of justice, but also embraces notions of justice based in recognition, participation and capabilities [178]. The US EPA - Office of Environmental Justice, later defined Environmental Justice as: "the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. It will be achieved when everyone enjoys the same degree of protection from environmental and health hazards and equal access to the decision-making process to have a healthy environment in which to live, learn, and work."

In this context, environmental inequality is defined as the different relationships that individuals, social groups and populations have with the environment. Different issues, such as access to natural resources (water, raw materials, etc.), intersect not only at an individual level, but also at the national level (developed/developing countries). This also applies to exposure to risk (odor nuisances, industrial pollution, etc). Naturally, inequality also exists between social or regional groups in relation to environmental policies. A key distinction is made here between inequalities in exposure and sensitivity. Environmental inequalities among individuals and groups indeed depend on a combination of exposure (socio-economic context, geographical context, behaviors, etc.) and sensitivity (age, health, etc.).

It is not always easy to evaluate the framework in which environmental inequalities emerge, relying on static predictors. A potential predictor is exposure during work as studies have shown that a quarter of environmental qualities observed are linked to work related exposures [56; 80]. Moreover, the geographical area of residence seems to be an important factor, as related cartography has shown that zones where heavy industry and transport are accumulated are potentially more susceptible to over-exposure. Walker [76] has determined that people in the most deprived 10% of areas in England experience the worst air quality, and 41% higher concentrations of nitrogen dioxide from transport and industry than the average. A similar study conducted in France has also found exposure risk to be much higher for French cities that comprise a "sensible urban area" than those who don't. A clear cumulative pattern of environmental and social inequalities can be identified here, as poor social conditions make people more vulnerable to risk, while exposure to risk can further affect their health and well-being.

A coherent conceptual framework is needed to understand and tackle environmental inequalities and the related health effects. A general challenge of all work on environmental justice is that more empirical research is needed, in particular into the interactions between the different determinants and geographical levels, requiring longitudinal environmental, health, and socio-demographic data. This could further improve our understanding of environmental inequalities and the related health effects and offer new opportunities for policy action [99].

In the context of the PNSE, environmental inequality term expresses the idea that populations are not equal in the face of pollution, nuisances and environmental risks. This inequality operates at different scales (global, regional, local) and could not be apprehended by the study of a single medium, but by the integration of varied contamination pathways: air, water, soil and food. In addition, environmental risk factor issues are multiple. Here, environmental inequalities are specifically seen as the unequal geographic distribution of exposure.

Environmental regulations generally set limits for individual pollutants in air, water, soil, food, and other sources. That kind of approach even though it has been shown to be effective, it fails to account for multiple pathways and different sources [141]. Despite the importance of cumulative impacts is evident, measuring and quantifying them is quite challenging. Quantitative assessment of cumulative risk is impractical or impossible in many real-world situations because data on interactions among environmental stressors are unavailable, information on place- and population specific exposures is lacking, and

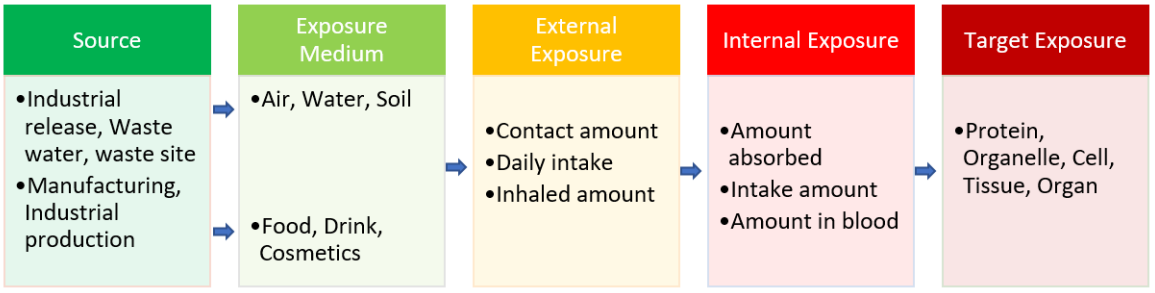


Figure 1.1: Source-to-exposure continuum diagram

validated models relating exposure to effect for multiple chemicals and combinations of chemicals do not exist [181].

Cumulative risk assessment is defined as a science-policy tool for organizing and analyzing relevant scientific information to examine, characterize, and quantify, as much as possible, the combined adverse effects on human health from exposure to a combination of environmental stressors [31]. The ultimate goal of cumulative risk assessment is to provide answers to decision-relevant questions based on organized scientific analysis; even if the answers, at least for the time being, are inexact and uncertain [181]. Cumulative risk assessment, therefore, involves the quantitative evaluation of risks to health from multiple families of substances that could differ drastically since available data, behavior and transfer processes vary.

1.2.2 The integrated exposure assessment framework

In general, exposure occurs when an agent or risk factor (chemical substance, material, pollutant,..) enters an organism. Naturally, environmental exposure doesn't always result in adverse effects, so one priority is to sort out these adverse environmental health effects related to the exposure. The term source-to-exposure continuum refers to the linkage, connecting the source of the exposure with the target exposure (Figure 1.1). The study of the processes that take place at the interface between the environment contaminants of interest and the organisms being considered, is called exposure assessment. As public concern and awareness about adverse health effects of exposure to environmental contaminants more and more interest is drawn to exposure assessment, as evidenced by the ever-growing number of related studies.

In order to assess the environmental exposure, the potentially important factors include: agents (biological, chemical, physical, single agent, multiple agents, mixtures), sources

(anthropogenic/non-anthropogenic, area/point, stationary/mobile, indoor/outdoor), transport medium (air, water, soil, dust, food, product/item), exposure pathways (eating contaminated food, breathing contaminated workplace air touching residential surface,..), exposure concentration, exposure routes (inhalation, dermal contact, ingestion, multiple routes), exposure duration and frequency, time frame and geographic scope (site/source specific, local, regional, national, international, global).

Environmental exposure usually evolves subtle effects that are harder to quantify and to assess. Even where actual exposure-related measurements exist, assumptions or inferences will still be required, therefore the exposure assessment will be based on a number of assumptions with varying degrees of uncertainty. The decision analysis literature has focused on the importance of explicitly incorporating and quantifying scientific uncertainty in risk assessments [142; 57]. The construction of scientifically sound exposure assessments and the analysis of uncertainty go hand in hand, so that the results have integrity or that significant gaps exist in available information that can make decision-making a tenuous process.

Many different approaches can be used for quantifying environmental exposures: direct methods (measuring, monitoring or biomonitoring) or indirect methods, involving exposure estimations from measurements and existing data, like environmental monitoring, questionnaires and exposure models. In many case-study approaches, direct methods cannot be applied, so, it is necessary for risk managers and decision makers to apply flexible tools which can assess environmental and human exposures in realistic way considering complex scenarios.

Aggregate exposure, i.e. the quantitative exposure assessment to a single agent from all potential exposure pathways and the related exposure routes, poses specific questions that need to be addressed:

- Identification of contamination sources;
- Estimation of the different environmental media contamination;
- Identification of exposure mechanisms (pathways and relevant routes);
- Internal dose in target tissue(s) based on temporal variation of exposure and contribution of exposure routes;
- Identification of vulnerable populations or specific susceptible groups (e.g. infants);

- Identification of contribution of sources to exposure, or possible exposure patterns when biological indices of exposure (biomarkers) are measured (reverse modelling);

- Direct evaluation of available biomonitoring data to toxicological/legislative thresholds (Biomonitoring Equivalents).

Refined aggregate exposure assessment is data-intensive, requiring detailed information at every step of the source-to-dose pathway. Integrated exposure assessment requires methodologies to allow calculating the aggregate exposure systematically; and computational research tools to disaggregate the exposure into the different contributing sources. Based on the needs described above, research objectives are to bring together all available information within a coherent methodological framework for assessing the source-to-dose continuum for the entire life cycle of substances covering an extensive chemical space.

1.2.3 The exposome concept

Exposure science is the study of the nature of contact between physical, chemical, or biologic stressors and humans or other ecosystem elements, with the objective to identify and understand fundamental, shared mechanisms and common biological pathways (e.g., inflammation, methylation, oxidative stress, and other epigenetic changes) underlying a broad range of complex diseases. The study of the various exposures has direct implications in the actions taken to improve individual and community health interventions [93; 204; 169; 9].

In this context, exposome was introduced in 2005 by cancer epidemiologist CP Wild [227], to define the totality of human environmental exposures (non-genetic) from conception, through lifetime as a conceptual framework for understanding the environmental context of health outcomes. Three overlapping domains constitute the exposome:

- The general external environment (urban environment, education, climate factors, social capital, stress,..);

- The specific external environment with specific contaminants (radiation, infections, tobacco, alcohol, diet, physical activity,...);

- The internal environment to include internal biological factors (metabolic factors, hormones, gut microflora, inflammation, oxidative stress).

The first two domains are the "eco-exposome" (external environmental stressors) and the last one is the "endo-exposome" (inward effects arising from the external). The assessment of source-exposure-disease relationships have been mainly studied in the endogenous level, concentrating on the effects of inward exposure on humans [9], while characterization of the "eco-exposome" has been more challenging [227].

Data and information emerging from an invigorated and expanding field of exposure science can be integrated in the exposome conceptual framework that provides the necessary linkages between source and internal exposure, and helps to identify and compare relationships between differential levels at critical life stages, personal health outcomes, and health disparities at a population level across space, place, and time [95]. This framework could be a layered structure that describes the elements of exposure pathways (Figure 1.1), the relationship between those elements, and how data describing the elements is stored and utilized for selected outputs, such as exposure assessment, exposure prediction or public health decision making [197].

The need for risk manager to identify population at-risk in the context of substantial data deficiencies that hinder evaluation of cumulative health risks brings the operational declination of the concept at the territorial scale. The characterization of the territorialized exposome implies the development of dynamic, multidimensional, longitudinal approaches, and information systems that require the adoption of transdisciplinary methods of data analysis. For example, integrated approaches could bring together all information necessary for assessing the source-to-human-dose continuum using Geographic Information System, multimedia exposure and toxicokinetic model.

This way, the hypotheses about pathways through which exogenous and endogenous exposures result in adverse personal health outcomes and population level health disparities can be generated and tested. Finally, it enables the identification of vulnerable individuals and communities at risk in order to target public health interventions [95; 197].

1.2.4 Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) refer to a group of over 100 different chemicals that are formed during the incomplete combustion of organic material such as coal, oil and gas, garbage, or other substances like tobacco or charbroiled meat is described. PAHs are usually found widely as a mixture containing two or more of these compounds, such as soot. The main pathway PAH substances enter the human body is through ingestion of food, either due to their presence in uncooked food or due to the cooking process. Other

important pathways include inhalation of ambient air and drinking water. In France, the main source of PAHs is residential activities, accounting for 58% of the total emissions, followed by transport (26%), industry (10%) and agricultural activities (6%).

Polycyclic Aromatic Hydrocarbons (PAHs) are a class of complex organic chemicals of increasing concern for their occurrence in the environment and effects. Depending on the atmospheric conditions, PAHs can be found far away from their emission sources, resulting to a widespread distribution in a continental scale and then accumulate in the food chain [74]. Therefore, this health risk posed by PAHs requires a need to control through air management by implementing policies to reduce ther emissions [74].

Studies about multiple human exposure to PAH substances are linking the pollutants to chronic illnesses. Moreover, evidence suggests that exposure to PAH mixtures can induce carcinogenicity and aggravate pre-existing conditions. It is therefore crucial to identify and monitor populations at risk of exposure. Furthermore, the toxicity among the PAH substances varies. Specifically, benzo[a]pyrene is characterized as a carcinogenic agent for humans, benzo[a] anthracene and dibenzo[a,h]anthracene as probably carcinogenic, while benzo[b]fluoranthene, benzo[k]fluoranthene, benzo[ghi]perylene and indeno[1,2,3cd]pyrene as possibly carcinogens (International Agency for Research on Cancer). In this case study, the considered substances are benzo[a]pyrene, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene.

Assessing human exposure to PAHs by integrating the multiple ways of exposure can therefore provide a significant contribution help identify the susceptible populations and potentially influence administrative policies, by providing solid knowledge bases.

Chapter 2

Literature review and objectives

The second chapter introduces the different scientific issues of the approaches used to characterize environmental inequalities. This chapter deals with the different types approaches developed to detect areas or potential overexposed populations. Examples of exposure assessment framework developed to characterize environmental inequalities examples are presented. Spatial analysis techniques and multimedia exposure modeling are described. This part also presents, in the framework of the characterization of the environmental inequalities: the Geographic Information Systems (GIS), as a computer system designed to support the capture, management, manipulation, analysis, modeling and display of spatial referenced data that provide a mathematical framework for manipulating spatially-referenced data.

2.1 Characterizing environmental inequalities

Environmental injustice can refer either to the racially discriminatory intent [13; 75] or to the inequitable environmental outcomes, such as disparate exposure or health/social impacts and uneven distribution of environmental burdens [44; 49; 166; 189]. Disparate exposure occurs when vulnerable populations are more highly exposed to environmental pollutants than others [67; 174; 193], while disparate health impacts are the negative health effects associated with this exposure [172; 109].

While studying the relationship between exposure, proximity, and health it should also be recognized the distribution of health outcomes across social groups as a function of at least three factors, exposure to health risks, biological susceptibility, and access to health care and other health-enhancing resources, which are often not distributed equally

across social groups [140; 228]. Therefore, groups that are equally exposed or proximate to environmental hazards may not bear the same health burden from exposure or proximity.

The majority of environmental studies to characterize potential environmental injustice has been performed in the USA but they have influenced other countries as well (Australia [113], Canada [90], South Africa [133]). Numerous studies published over the last decade on the thematic of environmental inequalities in Europe, proves the shifting interest on the thematic: UK [22; 221; 212], France, Italy, Germany, Netherlands [97; 19; 23; 206]. The common denominator of the aforementioned studies is the identification of the less affluent populations (minorities and low-income groups) as the populations most likely to be exposed to environmental risks such air pollutants, hazardous waste facilities, contaminated foodstuffs and pesticides [199; 235; 193; 160; 19; 127].

In the French context, environmental indicators could be used to characterize environmental inequalities when they are specifically seen as the unequal geographic distribution of exposure.

2.2 The exposure assessment framework as a method to quantify environmental inequalities

The notion of environment is broad and complex; it includes a variety of risk factors related to exterior, interior and professional environment, the different ways of life, the dietary habits, etc. Due to this complexity, the study of exposures has emerged as a new scientific discipline [155].

Different definitions of exposure are given by various organisms (AIHA, OECD, ATSDR, USEPA,...), however a general definition would be the contact made between a chemical, physical or biological agent with an organism, capable to elicit a biological response. An organism can be an individual or a population. The exposure dose is the amount of the agent actually deposited within the body. Accordingly, exposure assessment is the process of estimating or measuring the magnitude, frequency and duration of exposure to an agent along with the number of characteristics of the population exposed. Ideally, it describes the sources, pathways, routes, and the uncertainties in the assessment.

In this sense, it describes the potential exposures by constructing various scenarios and hypotheses, based on age, occupation, way of life, etc. In epidemiological studies, exposure assessment is used to relate the exposures to their health effects. In the context of public

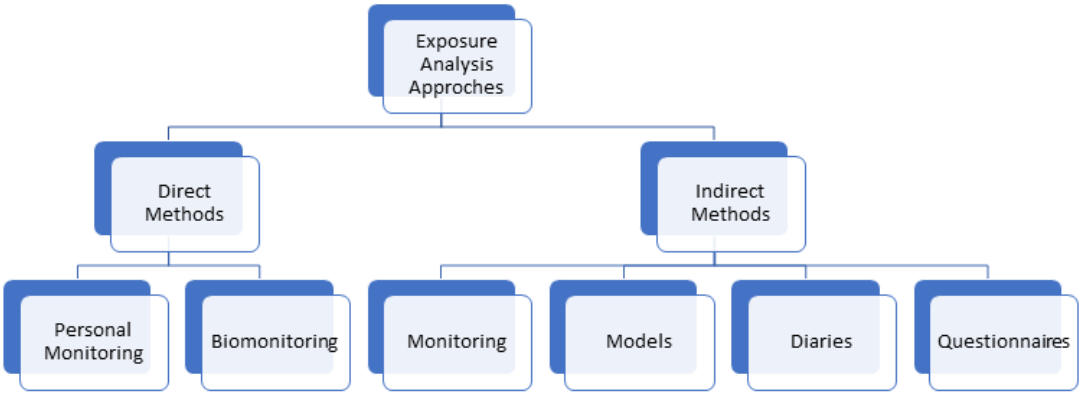


Figure 2.1: Schema of approaches used to determine personal exposure to pollutants (adapted from Risk Assessment and Toxicology Steering committee, 1999)

policies, exposure assessment can be used to drive and manage them as well as evaluating their efficacy. That ways, it provides an adequate framework to characterize environmental inequalities when exposures are accumulated or aggregated.

2.2.1 Principles

Chemical risk assessment could be defined as "the evaluation of the potential for adverse effect on human health or ecosystem from exposures to chemicals" (RATSC 1999). The environmental risk assessment is intended to describe the relationship between exposure rate to chemicals and toxic human health effects (Human Health Risk Assessment, HHRA) or ecological effects (Ecological Risk Assessment, ERA) from chemical contacts [150; 203]. It relies on the calculation of the internal or external dose. The potential (or external) dose is the amount of a chemical contained in material ingested, air breathed or a bulk material applied to skin. The applied dose refers to the amount of chemical in contact with a primary absorption boundaries (e.g. skin, lungs, gastrointestinal tract) and available for absorption. The absorbed (or internal) dose is finally the amount of a chemical penetrating across an absorption barrier or exchange boundary via a physical or biological process.

The process of chemical risk assessment can be divided into four steps: the hazard identification; the exposure assessment; the dose-response characterization and the risk characterization (National Research Council, 1989; RATSC, 1999). The exposure assessment step is very important and consists in quantifying the levels of chemicals to which environmental and human targets are exposed, in terms of magnitude, duration and frequency (RATSC, 1999). Many different approaches can be used for quantifying human

exposures (Figure 2.1). Direct methods include measurements of exposure taken at the point of contact or at the moment it occurs (monitoring or biomonitoring).

The exposure can be measured in different ways (Figure 2.1): at the point of contact (the outer boundary of the body) while it is taking place, measuring both exposure concentration and time of contact and integrating them (point-of-contact measurement). Indirect methods involve exposure estimations from measurements and existing data, like environmental monitoring, questionnaires and exposure models [149].

This evaluation process implies the analysis of emissions, contaminants' concentrations, identifications of exposure pathways, populations exposed and exposure doses [200]. By exposure factors, we refer to the aspects related to human behavior and activities (for example how much time is spent commuting to work each day), and contact rates (for example how much drinking-water is ingested per day). Exposure factors that are be used to quantify exposure and dose have been summarized by US EPA [200; 201] and are dependent to physiological factors.

A classical exposure assessment approach uses the concentration measurements of chemicals or pollutants in the different environmental compartments (air, soil, water) combining them with the population habits and behaviors as well as the physiological factors.

2.2.2 The daily exposure dose

The daily dose of exposure refers to the internal or external dose of a substance received by an organism in relation to the individual's weight and the number of days of exposure (in the case of a non-carcinogenic substance) and number of days during the whole lifetime (in the case of a carcinogen).

To evaluate the human exposure to different pollutants or chemicals, the total daily exposure dose is calculated. It is estimated by considering all the possible ways an individual can come into contact with a hazardous substance (exposure routes and pathways): air, water, soil, diet. By determine the relative magnitude of the contribution of each of the routes of exposure to the dose, this approach paves the way for the further elaboration of measures likely to contribute most effectively to the protection of human health.

For the daily exposure intake to be precise, the data describing the body weight, the volume of air inhaled, the quantities of ailments, water and soil ingested and if possible, to integrate the habits and general lifestyle of potentially exposed populations.

2.2.3 Modeling the exposure

In order to estimate exposure of individuals, population groups or even entire populations, exposure models can be developed. Usually, the exposure is estimated through the model as a continuous variable that spans over time from minutes to the entire lifetime. Consequently, the modeled output consists in descriptional statistics (means, medians), parameters of the distribution (standard deviation, quantiles, ranges) or even complete probability density distributions. In terms of variety, exposure models can differ widely in complexity, approach, inputs or outputs.

Exposure models are important tools for indirect exposure assessments. An exposure model is "a logical or empirical construct which allows estimation of individual or population exposure parameters from available input data" [223]. In risk assessment and risk management, models are used as a method to explain the relationships between the emissions and concentrations as well as to predict the results of risk management measures. They are usually used where direct measurements of exposure or biological monitoring data are not available or where these techniques are not applicable for the exposure assessment conditions. Predicting potential exposures for future releases or contact events is also possible by quantifying the human exposure through these models (RATSC 1999).

The degree of complexity of an exposure model can vary according to what is required by the assessment. The different exposure routes and pathways considered and included in a model depend on the complexity and the specific aim of the case study. So, a wide variety of exposure models are currently employed all over the world, characterized by different evaluating phases. Specific models have been developed to meet the requirements set for chemical exposure assessment by responsible authorities and agencies (e.g. ECETOC TRA, Stoffenmanager, Risk of Derm, ConsExpo are models developed in the context of REACH project). The existing exposure models, taking into account human exposure models and ecological risk assessment, can be categorized according to the following types of exposure source: environmental, dietary, consumer product, occupational, or aggregate and cumulative. Aggregate exposure models consider multiple exposure pathways, while cumulative models consider multiple chemicals.

Exposure to chemicals can be evaluated through the use of direct measurements in the environment (monitoring methods). It may seem that direct measurements give more reliable results than model estimations. However, even measured exposure concentrations can have a considerable uncertainty, due to temporal and spatial variations. Therefore, when carrying out an exposure assessment it may be very useful to compare and integrate

the estimated and measured concentrations in order to select the "right" data for use in the exposure characterization phase (Commission of the European Communities, 1994).

2.2.4 Spatially explicit exposure models

Spatially resolved multimedia fate and multi pathway exposure models facilitates estimation of environmental concentration distributions, related levels of contaminants in different sources, and the fraction of a chemical release that will be taken in by the entire human population (the intake dose) at the regional or local scale.

When spatial resolution of computations is low, usually variations in environmental characteristics tend to average out, and adoption of roughly selected representative or characteristic values allows the depiction of the correct orders of magnitude of outputs [119]. Research has been starting to cope with spatially explicit models of fate and transport with increasing resolution, and now a few models with resolution from a few tens of km up to 1 km are available for calculations at the continental scale [162; 146]. However, in any case the computational effort associated with this modeling strategy is generally quite high and limits routine applications when a large number of chemicals need to be evaluated. If a finer spatial resolution is advisable, media specific, specialized models. Specialized models such as GREAT-ER [17] and GEMCO (Anonymous, 2003) solve the problem with resolution, but lose the tight link between multiple pathways of fate and transport which were originally considered as the basis for the box-type models.

It has been shown [124] that in most cases the use of single-media models is satisfactory for the description of the fate and transport of chemicals, as feedback transport of chemicals across media is very low.

A GIS-based modeling platform developed by INERIS for quantifying human exposure to chemical substances (PLAINE: environmental inequalities analysis platform; [33]) aims to spatialize an environmental indicator related to human health using risk assessment methods and mapping environmental disparities at a fine resolution. This indicator integrates soil, water, air, food, demographic and behavioral georeferenced data to construct population exposure doses and associated risks. The project models trace elements (nickel, cadmium and lead) within a region of France. A stochastic multimedia exposure model was initially developed by INERIS [21] to assess the transfer of contaminants from the environment (air, soil, water) through the local food chain to individual exposure. This project as added a spatial dimension to the model using GIS and has been adapted to the French regional and national available databases. The Spatialized Risk Indicator (SRI) was constructed to

compare the risk contributions of different routes of exposure and to produce maps of the risk potential. The indicator is based on exposure modeling and integrates risks via the inhalation and ingestion routes of exposure. Spatial statistic methods are not integrated in PLAINE that is not able to propagate uncertainties in the global calculation chain

2.3 Spatio-temporal representativeness and modeling

Environmental decisions often rely upon observational data or model estimates. For instance, evaluation of human health or ecological risks often includes information on pollutant emission rates, environmental concentrations, exposures, and exposure/dose response data. Whether measured or modeled, each of these elements of information has certain underlying limitations. In addition to basic accuracy and precision issues, spatial and temporal representativeness of data and their applicability to different population or receptor groups of interest are important concerns. Understanding the various sources of variability and uncertainty in exposure and risk information is relevant to many types of environmental decision-making, including setting standards, determining emissions controls, and mitigating exposures to pollutants. Moreover, critical gaps in knowledge can be determined based on evaluation of uncertainties, so that priorities for collecting new data can be specified to improve confidence in future assessments [156].

Incorporating variability and uncertainty surrounding the estimates of human exposures is key to making sound decisions and maximizing the benefits attained from such decisions.

2.3.1 Data and representativeness

Environmental data are used to characterize possible exposures and are therefore one of the four primary data inputs for the public health assessment process, along with exposure data, health effects data, and community concerns. The quality and usability of all environmental data, should be assessed before employing them in the health or risk assessment processes., as many factors can bias environmental sampling results (EPA 1992a). Ideally, direct measures of exposure (e.g., biomarkers or personal monitoring data) for all key stressors related to a common health effect, throughout the critical time-period of exposure, and in the population of interest would be available [55]. Exclusive use of biomarker data in exposure assessment co characterize geographical environmental inequalities is currently not practicable when considering a large number of diverse chemicals due to analytical and resource limitations [201] specifically when the assessment should cover a

large territory. Environmental quality data are often available at a fine administrative or resolution level, and enable the building of environmental indicators on a regional scale. The definition of indicators for the identification and characterization of environmental inequalities depends on the reutilization of this type of data, which is very diverse by nature with regard to its initial intended objectives. Typically, the data available in a region of interest characterize levels of contamination at very specific locations, over a given spatial support (i.e. the support on which the data is measured such as point, surface or volume), and for very specific time frames. Determining how representative those measured levels of contamination are of other locations or time frames is not always an easy task (ATSDR, 2005). A variety of definitions is given for data representativeness and it does also reflect the variety of objectives covered under the term of spatial representativeness. Different definitions can be required to suit different purposes: model calibration and validation, detection of spatio-temporal outliers, design of monitoring networks, exposure assessment, statistical evaluations, etc. Representativeness according to Nappo et al. (1982) [148] is defined as the extent to which a set of measurements taken in a space-time domain reflects the actual conditions in the same or different space-time domain taken on a scale appropriate for a specific application. Larssen et al. (1999) [101] define the area of representativeness as the area in which the concentration does not differ from the concentration measured at the station by more than a specified amount. Another definition is given by Spangl et al. (2007) [179], referring this time to a representative monitoring station if the characteristic of the differences between concentrations over a specified time period at the station and at the location is less than a certain threshold value. Specifically, when assessing the human exposure to the various chemical substances, spatio-temporal representativeness refers to the capacity of the available measurements to adequately describe and characterize the risk to human health.

When risk cannot be quantified from available measurements in any meaningful or reliable way due to lack of representative data or missing source contributions, more sophisticated methods of spatial analysis can be applied to include additional information and take benefit from spatial and inter-variable correlation. To characterize the different scales of local, regional and global variability of the phenomena studied, the analysis of spatial data structures through geostatistical tools (variogram, autocorrelation analysis) is often employed.

2.3.2 Spatial statistics methods

Geostatistical and Bayesian methods are often employed in the context of environmental data modeling to answer to different data processing needs such as for spatialization purposes.

For interpolation purposes, the most common approaches include geostatistical methods (ordinary kriging, simple kriging, cokriging) [173; 91; 11; 10; 205; 60], geometric methods (inverse distance weighting (IDW), local polynomials) and statistical methods (linear regression model) [167; 40; 194]. Geostatistics is a rapidly developing branch of applied statistics with a huge set of methods and models for the analysis, processing, and representation of spatially distributed data. The use of geostatistical method, makes it possible to improve the reliability and quality of decisions based on spatially distributed data [137]. Unlike heuristic traditional interpolating approaches, which inevitably generate errors when generalizing the values measured by measurement stations, geostatistical methods provide a robust alternative when characterizing the exposure in the three environmental compartments.

With those methods, information from auxiliary variables can be included to compensate for a limited number of data and avoid exposure media distortion [135].

Recently, hybrid techniques that combine different approaches such as regression kriging have thus received increasing attention [131]. In regression kriging the residuals from a previously fitted regression model are interpolated using simple (known mean) or ordinary kriging (unknown mean). That way secondary information from an auxiliary variable can be included to improve the performance of the algorithm. Zhu [110] and Yao [61] compared the performance of ordinary kriging and regression kriging for soil properties and concluded that the second performed better when strong relationship existed between the initial and the auxiliary variables. Kriging with external drift is an alternative to regression kriging. Even though the two methods are the same in principal, in the case of kriging with external drift, the estimation is obtained in the same way as with ordinary kriging, with the difference that the covariance matrix of residuals is extended with the auxiliary linear predictors [171]. However, the drift and residuals can also be estimated separately and then summed. Goovaerts (1997) [68] uses the term kriging with an external drift to refer to a family of interpolators, and refers to regression kriging as simple kriging with varying local means. Minasny and McBratney (2007) [131] simply call this technique Empirical Best Linear Unbiased Predictor.

Despite the effectiveness of the interpolation algorithms, there are different factors that can affect their performance. Li and Heap (2011) [108] have reviewed the main reasons that the performance of the spatial modeling cannot be optimal. These include: sampling density [185; 217; 134; 79; 139], sample spatial distribution [18], sample clustering [185; 102; 159], surface type [46; 188; 220; 159], data variance [18; 128; 177], data normality [145; 217; 43; 151], quality of secondary information [2; 18; 128; 68; 104; 69; 132; 214; 165; 81], stratification [27; 220], and grid size or resolution [81]. Interactions among different factors may also exist [159]. The sources of errors in spatially continuous data and factors affecting the reliability of spatially continuous data have been discussed by Burrough and McDonnell (1998) [134]. However, there are no consistent findings about how these factors affect the performance of the spatial interpolators.

Bayesian hierarchical modeling is also often employed in spatial modeling [89; 170; 182; 157]. It has been shown in some cases to respond better to the common issues of spatial modeling (parameter estimation and inference, model specification and prediction), than the conventional frequentist approach [8; 106; 216]. An advantage consists in the solid foundation it provides as the uncertainties and existing knowledge of parameters are taken into account. In contrast to the frequentist modeling where parameters are fixed and unknown [116; 39], in Bayesian modeling, the parameters follow prior distributions. Moreover, Bayesian modeling framework allows the combination of experts knowledge, in the form of priors with observed data, which is particularly helpful in the case of limited or censored data.

Despite its advantages, Bayesian modeling requires integrations that are usually not tractable in closed form and therefore must be approximated numerically which can be computationally challenging [8]. Markov chain Monte Carlo integration methods such as such as the Metropolis-Hastings algorithm [78; 138] and the Gibbs sampler [63; 64] have been incorporated into available software programs (like WinBUGS [117], Package R2WinBUGS'for R (v.2.1 2015), MATBUGS for Matlab,etc.).

A collection of methods using GIS have been developed recently in order to integrate long range model predictions into exposure estimates [26; 3]. Regressions using Land Use Regression (LUR) have received particular attention [38; 24; 83]. Although these methods have been validated on continental scales and at fine resolutions [12], they have mainly been used in a local urban context. Dasymetric methods were developed a few years ago to allow aggregation of spatial data on relevant subsets. Auxiliary variables such as topography, satellite imagery, or land use have been used to redistribute population data on finer spatial

units [15].

In general, all the above approaches can be used to construct or process variables according to the objectives of the study and the type of data available through the development of spatial analysis techniques or the use of covariates [207]. The objective of the study should be considered when selecting the methodological approach.

The analysis of the uncertainties and variability of each of the links in the modeling chain allows the characterization of the sources of errors, the improvement of the interpretability of the results and the assignment of confidence levels relative to the results obtained. In the framework of the association of the spatial approach with the modeling of the exposure, a first examination consists in an analysis of the propagation of the errors during the various processing methods of the data used, the parametric uncertainties and modeling carried out.

2.3.3 Selection of spatio-temporal scale and resolution

Selection of scale is one of the most important factors in creating and analyzing data for exposure assessment and epidemiology. According to available data and the objectives of the study the environmental spatial scales can vary.

The different spatial scales encountered in an exposure assessment study could include [152]:

- Cartographic scale: Traditional map scale ratio relates the size of a feature on the ground to the size of a feature on the map. This is the scale normally listed on a road map. Scale selection results in the amount of detail including roads, water bodies, and land use patterns.

- Geographic extent: Refers to the size of the study area. For example, a study can be regional scale or global scale. The extent of the study area and/or its subsets can affect the analysis results (e.g., different results might be obtained when looking at cancer incidence in one state or province versus nationwide).

- Spatial resolution: Refers to the grain, or smallest unit that is distinguishable.

- Operational scale: Refers to the scale at which the process of interest occurs. For example, contaminant transport may occur at a small or large scale. Processes can be resolution dependent, that is, they can be detected at one scale but not another.

Homogeneity and heterogeneity of spatial data are affected by scale, and the scale chosen may affect the ability of the study to detect a relationship between the environmental exposure and the health outcome. This issue is similar to the modifiable areal unit problem, a term introduced by Openshaw and Taylor (1979) [195] that has long been recognized as an issue in the analysis of aggregated data such as disease incidence rates and census enumeration [229; 186].

A common spatial resolution needs to be defined for the data analysis and cartography of exposure. Transformations are employed in order to homogenize the data in the same spatial scale. Such methods include aggregating/disaggregating data in order to obtain assigned values for each spatio-temporal unit selected for the study. The spatial scale selected needs to reflect the local variations of the phenomenon studied. Moreover, it should minimize the need for spatial transformations of the data, which can be an important source of errors. Selecting a non-adapted spatial scale resolution can lead to the misrepresentation of the phenomenon variability and its extreme cases [48], and can affect the computation time.

The selection of the spatial scale in the spatial analysis of health and environment play an important role in the geographical studies [195; 147]. A number of methods have been proposed to address this issue [154; 62; 7; 53]. Often, the scale change has been achieved with the development of kriging techniques [70; 71].

In the context of exposure evaluation, the temporal variability of the contaminants studied, often poses additional constraints [98], as a single measurement (1 day or 1 week) hardly captures the central tendency of a longer period (1 year). As a result, the associated measurement error makes difficult the evaluation of health risks arising from environmental contaminants [184]. A large part of the temporal variability can be attributed to identifiable sources, microenvironments and activities. The human activity is the one that affects the different mechanisms (distribution, resuspension, degradation, translocation, removal [103; 129; 51], and determines the contact with the resulting concentrations. Longitudinal investigations (i.e. a research design that involves repeated observations of the same variables over short or long periods of time) of the activities that influence changes in contaminant levels over time are scarce but are necessary to increase our understanding of the role they might play in the risk of adverse health outcomes [51]. Recent studies have been integrating the space-time budgets and the spatio-temporal mobility of individuals in the studies [51; 196; 115].

2.3.4 Variability and uncertainty

By the term variability in exposure and risk assessment, we often refer to the differences over time, space or individuals of a group, and is a property of the system that is being modeled. Variability can be introduced as the result of inherently stochastic processes or as the result of differences between individuals of a population (spatial variation of concentrations to which individuals are exposed), which is of particular interest in human health risk assessment. It is a property of nature and it cannot be reduced by additional research, it can however be verified and estimated with greater accuracy.

Uncertainty, on the other hand, refers to the lack of information of an actual value of a quantity or the relationships among quantities. Unlike variability uncertainty may be reduced by additional research, but it cannot be verified. Some sources of uncertainty include: scenario uncertainty, model uncertainty, and data/input or parameter uncertainty [156].

Scenario uncertainty arises when the specifications of the exposure scenario that needs to be evaluated, is either incorrect or incomplete. Model uncertainty is a result of the mathematical model's limitations due to simplifying assumptions, exclusion of relevant processes, misspecifications of the model boundary conditions or due to the use of models developed for different purposes. Data uncertainty usually stems from random or systematic errors in the quantification process, or sampling errors (such as non-representative sampling), the use of modeled data instead of directly relevant or lack of foundation to characterize the input.

When conducting an integrated assessment, often is necessary to work with multiple models. Then, each model represents a different component of the scenario (emissions, exposure, dose and effects). These models can operate in different spatio-temporal scales, which poses a challenge when coupling them in a coherent framework and can result in structural uncertainty. Moreover, uncertainties in one model component can propagate through the rest of the components, contributing this way to the overall uncertainty of the output. Despite the importance of these types of model linkage issues, they have not yet been adequately addressed and are the subject of ongoing research.

Another way to report the uncertainties can be in the form of intervals. In this regard, the use of ranges of values for a particular metric of decision-making relevance (e.g., range of uncertainty associated with a particular estimate of risk) may be adequate, as long as the bounds of the range are properly placed with a probabilistic context (e.g., 95%

confidence intervals, interquartile range). When the associated uncertainty is too high for the results to be trustworthy, it should be clearly reflected in the presented results. This way, decision-makers are ought to understand whether their decisions are likely to be improved by waiting for additional information on critical factors influencing these decisions, or regarding the need to consider an adaptive strategy based on iterative reassessment [156].

2.3.5 Data representativeness in environmental compartments

In Soil

Characterization of contaminated soil areas is based on a finite sample of observations of contaminants and their levels. In spatial assessment, the contaminant spatial distribution between the sample points needs to be estimated by an interpolation technique, or simulated by multiple realizations of the contaminated area. These methods have an advantage over the heuristic ones since they are able to capture the complex spatial distribution of the contaminants and allow the integration of relevant secondary information, like physicochemical properties, as well as proposing ways to compute the uncertainty associated with the estimates.

Advances in the domain of geostatistics in the last thirty years have allowed to tackle the above objectives. Geostatistics are now used in mapping contaminant levels, describing the declination of contaminated areas and quantifying local and global uncertainties of pollutant grades and pollutant volumes on real case studies. Starting from the early 1980s, the US Environmental Protection Agency was one of the first to propose the use of geostatistical methods to map soil contamination. Initial approaches included the simple two-dimensional contaminant grades mapping with kriging approaches [187]. More evolved methods include the uncertainty assessment on the soil classification and the pollutants' localization [72; 73; 32; 47]. Indicator kriging which makes no assumption on the underlying distribution of the sample data and allows the explicit estimation of local uncertainty is one of the most frequently-used methods as well as disjunctive kriging and stochastic simulations [70; 104]. Geostatistical approaches are also widely used in a health risk assessment context as a way to characterize contaminated sites [66; 100].

The spatial distribution of PAH substances varies with scale [230; 208]. PAH soil concentrations are shown to decrease along with urban-suburban-rural gradient [215; 232; 213]. They are highly affected by land use, as higher concentrations are observed close to industrial areas and main roads. In a fine scale, the complex patterns of their distributions

which are influenced by multiple factors becomes more evident. Soil physicochemical properties have also been shown to affect the distribution of PAH in soil [161]. Moreover, due to their properties, PAH substances are characterized by their high durability in the environment, accumulating that way in the soil and degrade with difficulty. Their decomposition depends to a large extent on soil properties [122], and the scope of their harmful impact on organisms is directly related to soil type [122].

Often, quantitative data are reported as a single point (e.g., average concentration of a chemical in soil) or represent only a moment in time (e.g., concentrations in emissions from an incinerator). Rarely, however, do these single data points represent the full range of actual conditions. For example, an average soil concentration does not indicate the highest or lowest detected value. An emissions sample from an incinerator does not represent changing conditions, such as increasing capacity or varying waste stream composition. The use of geostatistical models and auxiliary variables are widely used to overcome these limitations.

In Air

Accurate and spatially high resolved maps of PAH levels in the ambient air are essential in the assessment of individual exposures to these substances. Anthropogenic sources of total PAHs in ambient air emissions are greater than those that come from natural events such as forest fires and volcanic eruptions. Apart from localized risk at or near the source of emission, PAHs can be dispersed regionally and intercontinentally through atmospheric long-range transport.

Identifying the development of models for assessing air pollution exposure within cities has been identified as a priority for future research [11; 84]. Studies assessing the exposure of PAH substances, often disregard the aspect of spatial distribution of pollutants, using directly environmental quality data to evaluate the health risk associated to the exposure [162; 34; 105]. Nevertheless, spatial distribution variability is an important aspect when calculating the exposure dose. Oversimplifications, such as assigning an average value to a large region, can lead to misestimation [162; 41] and consequently misclassification of the exposure [163].

Multiple studies have shown strong disparities when assessing the presence of a pollutant in urban, rural or highly industrialized areas. They have shown that people residing in urban zones, close to highways or industry are more affected than those in rural areas [92; 190]. Epidemiological studies have also shown the aggravating outcome in human health of air pollution between the different urban and non-urban zones [202; 35]. In the

associated bibliography, the necessity for spatially assessing the associated risk consists in new methods to address the recurring thematic.

When spatially assessing the human exposure to air pollutants from monitoring networks, stochastic methods such as kriging are preferred. An issue commonly reported is the availability of data. Some databases include only a limited number of observations (monitoring stations) and therefore it is not possible to assess the population's exposure adequately. External drift kriging is then widely used in air quality modeling, in order to include auxiliary information in the model.

In Water

Left censored data have been an ubiquitous problem in environmental and occupational health sciences. They are the result of limitations of chemical analysis procedures, and thus small concentrations cannot be precisely measured. Appropriate handling of these observations reported to have non-detectable levels of contaminants has been common problem [86].

Even with technical advances in the areas of field sampling, sample processing protocols, and laboratory instrumentation, the desire to identify and measure contaminants down to extraordinarily low levels means that there will always be a threshold below which a value cannot be accurately quantified [65]. As a value approaches zero there is a point where any signal due to the contaminant cannot be differentiated from the background noise, i.e. the precision of the instrument/method is not sufficient to discern the presence of the compound and quantify it to an acceptable level of accuracy. While one can differentiate between an "instrument limit of detection" and "method limit of detection" it is the method limit of detection that is of main interest.

The problem of estimating the parameters of distributions with censored data has been extensively studied, for example [37; 86; 65; 58]. The most simplistic and easy approach to handling left censored data is the substitution of under the detection limit measurement with a constant, usually replaced with zero, with some fraction of the detection limit (usually either $\frac{1}{2}$ or $\frac{1}{\sqrt{2}}$) or with the detection limit itself. However, this kind of method has been proven to introduce great bias in the estimation, as well as underestimating the variance when the percentage of left-censored data is well over 5%. Extrapolation strategies use regression or probability plotting techniques to calculate the mean and standard deviation based on the regression line of the observed, above the detection limit values and their rank scores.

An alternative to the above methods is the use of multiple imputation. The idea behind it, is instead of filling in a single value for each missing value, replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiple imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. Usually in environmental science, the method of model based imputation is suggested. Model-based imputation consists in generating randomly observations below the detection limit, using the detected sample values to predict the distribution, and then proceed with analyzing the values as complete cases. Such a method would not work in our data, as we would have to adjust a distribution for each distribution facility, in order to impute the values, however there are numerous facilities for which most or all the measurements are beneath the detection limit, making it unreliable to fit a distribution.

In a bootstrap based expectation-maximization (EM) algorithm was employed to perform the multiple imputation [85]. This method consists in creating multiple "complete" data sets, that can then be analyzed as complete cases. Honaker, King and Blackwell propose an algorithm that treats time series-cross sectional data, by using the EM algorithm on bootstrap samples of the original data to draw values from the complete data parameters. This way, the correlation between the three substances can serve to the simulation of the concentration measurements under the detection limit, and in particular for the distribution units were all the measurements of a certain substance are under the detection limit.

Various tools are available for performing (multiple) imputation. As a guide for the interested reader we list some procedures available in R, but not exhaustively: Amelia II, array, Impute, cat (for categorical-variable datasets with missing values), EMV, impute, mi and mice. Tools are also available within other statistical packages, such as ICE in STATA, the SAS PROC MI, Missing Data Library, and NORM for S-Plus and SOLAS. MI can also be performed with MLwiN, and recently with SPSS (version 20, 2012). Horton and Kleinman have applied imputation with Amelia II, Hmisc, mice, and other commercial packages (i.e. ICE/Stata, Iveware, LogXact, SAS, and S-Plus) and found similar parameter estimates for all analysis, as well as considerable reduction of the standard error estimates when compared to the complete case estimators.

2.4 Thesis problematic

2.4.1 Research questions

World Health Organization in a recent report (2012) has identified environmental inequalities as a priority issue in need to be addressed by the national governments in Europe. It is highlighted that populations from socioeconomic disadvantaged groups are often more likely to be overexposed to environmental pollution and more vulnerable to their resulting health impacts. In this report, it is also specified that reducing these inequalities, requires identifying and characterizing exposure as well as social factors in order to interpret how they accumulate across a territory and prioritize interventions. As the health status of a population is the result of complex interactions between several social, territorial and environmental factors, all related information needs to be studied in order to assess it. Therefore, the development of methods allowing the characterization of environmental health inequalities, is a prerequisite for the implementation of public health actions aimed at the protection of populations.

In France, the emergence of the third plan for health and environment prioritizes the elaboration of a methodology to integrate the exposome concept in an effort to link environmental inequality indicators with exposure and spatial data (PNSE3, action 38). At a regional level, this implies the analysis in order to identify and manage the areas at risk of overexposure where they are suspected to generate a potential increasing risk to human health.

Nevertheless, constructing methods and tools to help orient public policies in order to reduce territorialized environmental health inequalities requires the evaluation of phenomena not always easy to apprehend, as evidenced by the emergence of the concept of exposome and the reliability and representativeness of available information that usually demand statistical treatment.

The previously conducted bibliography has shown that population census, environmental quality monitoring and regional disease and mortality mapping that have been conducted independently of one another, depend on the specific needs, constraints and objectives of each study's individual purpose. In addition, those studies have revealed important spatial disparities in the quality of environmental media.

Furthermore, studies showed that no matter the source (natural or anthropogenic) the contaminants can be transferred and transported between the different environmental compartments and humans can come in contact with them through the various media

(drinking water, air inhalation,..). This plurality of exposure routes during the lifetime along with the difficulty to assess the exposure for each individual and the need to carry out prospective studies, contribute to the strong need for modeling in this domain.

The construction of indicators to identify and characterize the territorialized environmental inequalities depends on the availability of data. The available databases are often assembled for diverse objectives, and often re-processed using statistical methods. As a result, they often lack a common spatial support and therefore prior spatial analysis in order to homogenize them or increase their resolution is required. In addition, the temporal support also differs between the available data (punctual measurements, year averages, etc.) which also requires additional treatment.

Spatialization and crossing of available databases poses a number of methodological difficulties and can introduce uncertainties in the cartography process carried out. For this reason, different methods and techniques are employed to specifically treat environmental databases in order to take benefit from all available information and reduce the uncertainties in the final map.

With regard to the above remarks, and in order to address the issue of characterizing environmental exposure inequalities to different chemical substances, a number of research questions arise:

- Which approach can be developed to combine different levels of information relating to spatio-temporal resolution, degrees of analysis of the conceptual framework of exposure and knowledge of the links between health and environment?

- What types of variables can be constructed to characterize exposure at a fine spatial resolution?

- What statistical/mathematical methods can be employed to process the different types of data?

- What environmental determinants control the PAH exposure variabilities at the local and national scale? What uncertainties arise from the data processing and modeling?

- How can the methodological approach of characterizing and evaluating environmental inequalities be improved in the framework of the evaluation of exposure?

2.4.2 Thesis' objectives

The objective of this work is to explore techniques for the development of composite indicators for the spatial representation of risks and the analysis of environmental inequalities, by answering the above research questions. To respond to the general objective, it is required:

- the characterization and integration of spatial phenomena that operate in different spatio-temporal scales;
- the integration and combination of various levels of data from different environmental compartments;
- the characterization of the principal exposure pathways;
- the construction of realistic exposure scenarios integrating past and present sources;
- the description of the phenomena at a fine spatial and temporal scale.

The availability of data on the geographic area of interest for the pollutants evaluated is an essential prerequisite. In order to construct the exposure maps from spatialized databases in the context of evaluating health risks, the development of methods is required to treat and harmonize the available data, with respect to their specificities (missing values, limited number of observations, etc) in the same resolution and support. Ad-hoc methodologies are used to align the available data to the same pixels.

Spatial analysis and statistical methods are employed to process and assemble the databases for the purpose of the study, using R and GIS. The exposure model MODUL'ERS is employed [21] to calculate the exposure doses as well as the spatialized risk indicators, associated with the ingestion of food products, water and soil and air inhalation. The model's input is georeferenced environmental databases (air, water, soil, foodstuffs), and the output obtained is the Relative Spatialized Indicators, that corresponds to the population's exposure for a period of 30 years, considering that pollution practices remain the same and no environmental restoration has been employed.

The family of pollutants selected for this study is the Polycyclic Aromatic Hydrocarbons (PAH). For feasibility reasons three substances are selected: Benzo[a]pyrene, Benzo[ghi]perylene and Indeno[1,2,3-c,d]pyrene. All three substances have been identified by the US EPA, the US Agency for Toxic Substances and Disease Registry (ATSDR), and

the European Food Safety Authority (EFSA) as priority PAH, due to their carcinogenicity or genotoxicity and/or ability to be monitored. The geographic area considered is the totality of the metropolitan French region. Data from the three environmental compartments (water, air, soil) are available from different sources, and are spatialized with respect to their specificities.

Chapter 3

Materials and Methodology

3.1 Basic concepts and tools

In this section, the principal tools employed for the modeling, the analysis and the cartography of population exposure to chemical substances are presented.

3.1.1 Multimedia exposure models

In an effort to assess the health risks associated with exposure to environmental factors, detailed indicators need to be built. This requires the integration of a big amount of data in order to describe the phenomena with accurate precision, especially for long term exposures. Obtaining data to quantify the exposure is not always possible, mainly due to lack of associated data.

Environmental models are important tools in understanding the behavior of contaminants in complex environments. Environmental models can be divided into single-medium models, used to simulate chemical transport with a focus on a single medium, and multimedia fate models, used to calculate chemical behavior in various environmental compartments. Examples of single-medium models include ISC (Lakes Environmental Software Inc, 2006), CALPUFF (TRC Companies Inc, 2006), and UAM [198] for air dispersion, QUAL2E [158] and WASP [210] for water quality. All of these models can be used to analyze and predict the temporal and spatial distributions of chemical substances; however, they provide no information on multi-compartmental relations.

The evaluation of multimedia exposure allows the estimation of exposure through the multiple exposure and absorption pathways. This at a time requires the collection of an

important amount of data and the development and integration of algorithms for each exposure pathway, in an order to combine them all in a single indicator. A cumulative multi-pathway exposure assessment for humans relates contaminant concentrations in multiple environmental media to concentrations in the media with which a human population has contact (personal air, tap water, foods, household dusts soils, etc.). The potential for harm is assessed either as the average daily intake or uptake rate or as time-averaged contact concentration.

Multimedia exposure models are able to include all the relative information about partitioning, inter-media transfer properties of the chemical substances of interest and reaction along with information about exposed humans.

Multimedia models can be divided in two types:

- Models that combine multiple computation models, each one of them allowing the calculation of the pollutant transfers within a compartment, where the output data of each compartment is used as input for the adjacent compartment, accounting that way for the inter-compartment transfers. That way, the pollutant's mass is not conserved between the environmental compartments and the secondary transfers are not accounted for completely.
- Models that conserve the pollutant's mass amongst the environmental compartments. That ways, the pollutants are considered to be distributed in an environment divided in compartments.

Different ways to classify models have been proposed. A common ground appears to be developing around the following classification scheme: 1) physical or empirical; and 2) deterministic or stochastic (probabilistic). However, alternative classifications may be considered as well.

The physical model uses processes, physicochemical characteristics and mass relationships based on balance principles in order to predict exposures. That way, the model uses a mathematical approach in an attempt to reproduce the physical and chemical reality, relevant to the exposure of interest. This type of model normally produces single point estimates of outputs by employing single point values for input variables, however it can also produce distributions of output variables by sampling from distributions of input variables (i.e. the probabilistic approach).

Physical exposure models are built on laws of physics, chemistry and lifestyle (behavioral) data and factors potentially influencing exposures that is to say real-world exposure

phenomena that are represented by equations. Results can be calculated even when there are no measurements of the output variables. Therefore, they can be applied in situations where no measured exposure data are available or where such measurements are even impossible, as in the assessment of exposures in the past and predicting exposures for future scenarios.

On the other hand, an empirical model is a mathematical representation of the relationship between input and output variables based on historic measurements. Empirical models are widely applied when descriptive data analysis is needed. To identify and explore possibly correlating phenomena statistical methods are employed. In order to reveal possible causal mechanisms behind these correlations, further studies are required using conceptual methods. Once the causal mechanism is known, the process equations can be analytically written and then physical models developed to describe these relationships.

3.1.2 The Geographic Information System

The term geographic information system (GIS) generally refers to any information system that integrates, stores, edits, analyzes, shares, and displays geographic information. It has numerous applications in different domains such as: environmental analysis, epidemiology, public health, urban planning or telecom and network services.

The GIS first emerged in the early 1960s and the mid-1970s as a new discipline, mainly developed and used from national agencies (surveillance, defense). From the mid-1970s to early 1980s GIS saw the adoption of technologies by national agencies that led to a focus on the development of best practice. By 1982 until the late 1980s GIS developed and exploited by the commercial marketplace whilst the final phase since the late 1980s has seen a focus on ways of improving the usability of technology by making facilities more user-centric. Nowadays due to the recent technological advancements (powerful processors, data storage,..) as well as data availability, have made GIS ubiquitous. GIS is entering an era of open source software, where a big shift is noted and the users build their own GIS software in an open, collaborative way.

GIS provides the framework for building complex models by combining primitive analysis functions. Even though the different systems may vary as to the complexity and the specific functions they provide, there is a standard set of primitive analytical functions that are accessible to the user. These include:

- Retrieval, re-classification, and generalization;

- Topological overlay techniques;
- Neighborhood operations;
- Connectivity functions.

Especially when dealing with real-world problems and applications, the selection of the functions to use should be made in the context of a complete analysis strategy, in such way to lead to high quality of information produced [5].

Two types of data format are readable by GIS:

- Raster: The entire area of the map is subdivided into a grid of tiny cells. A value is stored in each of these cells to represent the nature of whatever is present at the corresponding location on the ground. Raster data can be thought of as a matrix of values. The major use of raster data involves storing map information as digital images, in which the cell values relate to the pixel colours of the image. To reproduce the image the computer reads each of these cell values one by one and applies them to the pixels on the screen.
- Vector: The features are recorded one by one, with shape being defined by the numerical values of the pairs of xy coordinates (shape can be points, lines or polygons). Vector data can be thought of as a list of values. In vector data, the position and shape of a feature is captured as a series of four pairs of numerical coordinates. To reproduce the feature in a GIS the computer reads these values and draws a line linking the coordinate positions. The vector version can also store additional context information about these features.

Finally, a system of coordinates (spherical or projected) needs to be assigned in order to provide a framework for defining real-world locations. The data organizing is done in layers, each layer containing homogenous information. The use of GIS in spatial statistics for linking exposure and health data in the context of epidemiologic analysis is a growing field of research [53; 180; 52].

3.1.3 Spatial analysis

Spatial analysis is used to address the distributional relations of geographic features on a defined space. It is used to tackle fundamental issues in: the definition of its subjects

of study, the construction of the analytic operations to be used, the use of computers for analysis, the limitations and particularities of the analyses which are known, and the presentation of analytic results.

Common errors often arise in spatial analysis, due to the mathematics of space, to the particular ways data are presented spatially or to the tools which are available. These problems represent a challenge in spatial analysis because of the power of maps as media of presentation. When results are presented as maps, the presentation combines spatial data with analytical results which may be subject to uncertainty, leading to an impression that analytical results are more accurate than they really are.

Spatial analysis can be summarized in three main concepts:

- The cartographical modeling, allowing spatial representation and manipulation of data with different spatial operations such as: spatial query, transformations, intersections, network analysis, location analysis (spatial optimization), etc..

- Mathematical modeling of geographic features taking into account their spatial dependencies. Extension of traditional statistics into a spatial setting or quantification of the spatial relationships between observations of different locations for estimation of unknown locations.

- Spatial modeling which involves constructing models to predict spatial outcomes and focuses on spatial statistical modeling to study the spatial variability.

Environmental spatial analysis includes the theory, methods, and technologies associated with proper handling and use of spatial data for the analysis and management of environmental problems and processes. Core technologies include geographic information systems (GIS), remote sensing, and quantitative spatial analysis. Methods and models for evaluating environmental patterns and processes exploit developments in statistical, cognitive, and information sciences.

The volume of digital environmental spatial data is growing rapidly and requires innovative approaches to make sense of it efficiently. The data come from a variety of sources, including new remote sensing technologies, survey research, automated monitoring technologies coupled with global positioning systems (GPS), and digitization of analog data collections. Efforts to synthesize these data must be coupled with rigorous understanding of how uncertainty affects the quality of derived information.

The complexity of the environmental problems we face requires understanding and fusion of spatial and other data from many sources, in many formats, and from multiple disciplinary perspectives.

3.1.3.1 Geoprocessing

In the modeling process it is often required to integrate data from different sources, which often requires dealing with different spatial supports and different spatial resolutions. To address that, geometric transformations of data are employed, in order to obtain all relevant data in a common reference system.

3.1.3.2 Spatial interpolation

Spatial data interpolation consists in predicting values in a continuous area or points of choice, from a limited number of sample data points. It can be used to predict unknown values for any geographic point data. It is based on the assumption that spatially distributed objects are spatially correlated; that is to say, features that are close together tend to have similar characteristics. Interpolation methods can be global, where a single function is mapped across the whole region or local, where an algorithm is applied repeatedly to a small portion of the total set of points each time. With respect to the modeling, interpolation methods fall in two categories: deterministic, where the interpolators model a data point at a particular position using geometric and mathematical properties, or stochastic where geostatistical models are employed to model the probability of a data point being at a particular position.

In general, the deterministic approach is employed and preferred when the objective is to construct an exact model describing the phenomena, by formalizing the physical laws with mathematical equations. All available data (quantitative and qualitative) are taken into account and are used to calibrate the model. However, when data is incomplete, sparse or missing, a probabilistic model is employed to describe the phenomena using random variables. Geostatistical models are employed in that case as they also consider the spatial positioning and autocorrelation of realizations.

The principles of geostatistics are described in a large body of literature e.g. [30; 43; 94; 185; 87; 130; 219; 36]. Generally, geostatistical techniques rely on statistical models that are based on a random function (or a random variable) theory to model the uncertainty associated with spatial estimation and simulation. The covariance function and variogram

model are then fitted to the spatial structure that characterizes the observations. A common method to evaluate the performance of the geostatistical models is cross-validation. Cross validation consists in removing one or more data locations and predict their associated data using the data at the rest of the locations. In this way, the predicted value is compared to the observed value and useful information is obtained about the quality of the model (for example, the variogram parameters and the searching neighborhood).

3.1.4 Statistical analysis and modeling

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data. Statistical modeling specifically is a simplified, mathematically-formalized way to approximate reality (i.e. what generates the data) and optionally to make predictions from this approximation. The purpose of statistics is to develop and apply methodology for extracting useful knowledge from both experiments and data. In addition to its fundamental role in data analysis, statistical reasoning is also extremely useful in data collection (design of experiments and surveys) and also in guiding proper inferences [59; 88].

Inferential statistics involves the identification of a suitable probability model. The model is then fitted to the data to obtain an optimal estimation of the model's parameters. The model then undergoes evaluation by testing either predictions or hypotheses of the model. Models based on a unique sample of data can be used to infer generalities about features of the whole population.

Statistical modeling in environmental science is employed in order to improve the scientific understanding of processes in the environment to predict how they behave and then use random variables to model those components of the process that we do not yet understand. This also allows to make predictions from data with associated measures of uncertainty, which can potentially help the scientist, policymaker and environmental manager, because it allows them to explore and understand the risk associated with different decisions that they have to make about processes with uncertain outcomes.

The use of classical data modeling approaches along with geostatistics methods can offer a complete framework and help overcome the shortcomings of both approaches.

The development of computer technology since the 1950s has led to the creation of many very useful statistical software packages for analyzing data. Many diverse statistical software packages are currently available that offer a wide variety of capabilities. Language-

based packages such as R [168] allow the user to develop their own statistical macros and functions in addition to the comprehensive range of statistical routines available.

Various packages are available for reading, visualizing and analyzing spatial data. In this case, the focus is on geographical data, where observations are identified as geographical locations. Some useful packages that were used for in this work include: sp, rgdal and raster for reading/writing and handling spatial data, gstat and geoR for geostatistics and mapview for visualization of spatial objects.

The estimation of exposure due to the main sources is essential for assessing the magnitude of the exposure associated risk to the different substances. For this, it is necessary to integrate the transport and transfer scenarios between the environmental compartments and exposure media as well as approaches that take into account the different spatial and temporal scales. The choice of data to describe the environmental compartments focuses on measured data, which increases the need for georeferenced data. The processing methodology is presented in Chapter 3. The results include homogeneous spatialized environmental databases, on the same grid of reference and are presented in Chapter 4.

3.1.5 General approach

No statewide data are available to provide direct information on exposure. Exposure generally involves transfer of chemicals from a source through the environment (air, water, soil, food) to an individual or population. Environmental concentrations of B[a]P are obtained from available monitoring databases. Overlaying data from several sources are selected according to the availability, spatial coverage and their relevancy in the framework of exposure evaluation and environmental inequalities.

For the purposes of the study, data related to pollutant sources, releases, and environmental concentrations are used to build indicators of potential human exposures to pollutant. Statistical and geostatistical methods are developed to spatialize annual or annual average concentrations of B[a]P in France in the three environmental compartments (air, water, soil) on a referent grid (9 km^2 grid).

A multimedia model, developed by INERIS (MODUL'ERS), operating in GIS environment is employed for quantifying human exposure to toxic substances and analyze environmental determinants. An illustration of the different steps of this process is found in Figure 3.1.

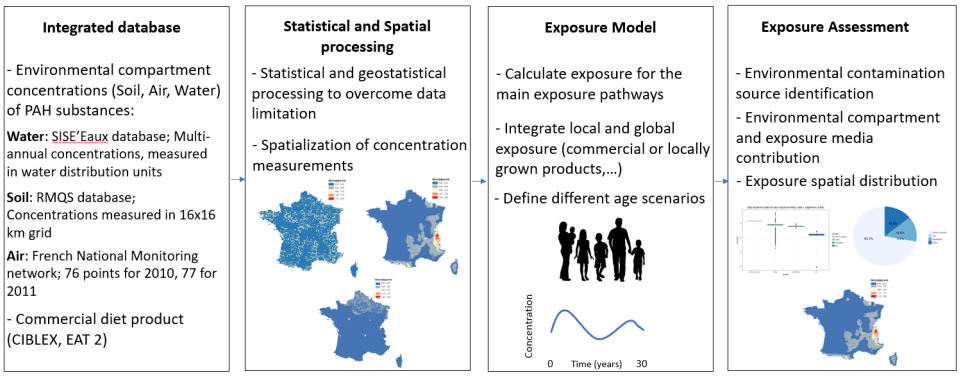


Figure 3.1: Illustration of the exposure assessment process.

3.2 Data selection

On the base of the analysis of the bibliography [123] the following conclusions may be drawn:

- PAH are found in natural objects;
- PAH pollution increases in regions with intensive economic activity;
- the relative content of considered PAH is higher on particles than in the gaseous and dissolved phase in the atmosphere, in precipitation and in water basins;
- soils, bottom sediments, plants and sea organisms accumulate PAH in a considerable quantity.

Then, PAH environmental inequalities could not be apprehended by the study of a single medium, but by the integration of varied contamination pathways: air, water, soil and food. That requires:

- to take in account major exposure pathways;
- to describe phenomena at a fine resolution;
- to integrate actual and past sources on local and global scales;
- to build realistic scenario that integrate background exposure.

The estimation of environmental exposure requires knowledge of the concentrations of environmental compartments to which population is exposed. These concentrations can be

Variable	Original data resolution and support	Source
Atmospheric concentration	76 (2010) and 77 (2011) measurement points	National air quality database (2010 and 2011 data).
Atmospheric emission	Grid $0.01° \times 0.01°$	Annual emissions of the National Spatialized Inventory (2007 and 2012).
Atmospheric dispersion concentrations	$7km \times 7km$	CHIMERE (Eulerian atmospheric dispersion model)
Altitude	Centroids of French municipalities	National inventory.
Drinking water concentration	Surface (municipalities)	Multi-annual measurements of SISE'Eaux database (years 2000-2012).
Population	Grid 1×1 km	INSEE.
Topsoil concentration	Grid 16×16 km	RMQS [208] for year 2010.
Polluted sites	788 points	BASOL (National database of polluted or potential polluted soils).
14 Soil properties	Grid 3×3 km	Database provided by INRA.

Table 3.1: Available data and support.

measured or modeled. A wide range of environmental data might potentially be mobilized for integrated assessment.

A data inventory is assembled for PAH characterizations to the following main themes: soil, air, drinking water quality, atmospheric emissions/concentrations, polluted sites, and exposure factors (Table 3.1). In order to make the task manageable, attention has initially been focused on data available at the national scale which are gathered on a routine basis. Nevertheless, the geographic coverage or extent, for example, is inevitably ambiguous. Indeed, most environmental data, by their very nature, are samples and do not provide complete area coverage. That is why approaching nation coverage is possible if different datasets are collated and combined. Here, data sources have generally been included when they are considered to represent a potential basis for assessing exposures across populations at the national level, either directly or by interpolation.

From this inventory, the database selection and study design definition is guided to reach the best compromise between data representativeness and method robustness, consistent with the objectives of the study.

Polycyclic Aromatic Hydrocarbons refer to a group of over 100 different chemicals that

are formed during the incomplete combustion of organic material. Nevertheless, most of the monitoring network are only following the 16 priority PAHs. Benzo[a]pyrene is characterized as a carcinogenic agent for humans, benzo[a] anthracene and dibenzo[a,h]anthracene as probably carcinogenic, while benzo[b]fluoranthene, benzo[k]fluoranthene, benzo[ghi]perylene and indeno[1,2,3cd]pyrene as possibly carcinogens (International Agency for Research on Cancer, [225]). Due to toxicity and availability in each considered database (limited by atmospheric modeling substance availability), the considered substances are benzo[a]pyrene, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene.

3.2.1 Integration and processing of different sources of data

A share-of-population census, monitoring, modeling of environmental quality data production were independently conducted of each other according to specific needs and constraints. This feature implies that the different data types from different sources and databases cannot directly be represented under a common denominator, namely, their spatial location or distribution. This representation is achieved by depicting the different data types as layers and superposing those layers in the same geographical reference grid. The problem of linking data sets derived from incompatible spatial frameworks (for example, linking point and pixel-based environmental data) has attracted a considerable amount of interest. A referent grid of 3 km^2 is generated for the study, and all the spatial variables are discretized on this grid. The grid selection allows the study of the exposures in a fine resolution while remaining computationally efficient.

To overcome data limitations, mathematical and statistical models are constructed in order to quantify exposure. Algorithms are developed using spatial analysis and statistical methods to build and discretize interest variables from different supports and resolutions on the 3 km2 regular grid within France.

3.3 Multimedia modeling methodological approach

In order to adequately describe and characterize the human exposure to PAH, it is important to consider the principal pathways. Ideally, all routes of exposure and all environmental media should be assessed to determine their relative contribution to the overall exposure associated with environmental contamination. Likely media of environmental exposure include air, water, food, and soil. In this study, the two main exposure pathways of ingestion and inhalation are taken into account excluding dermal pathway that is not

relevant for PAH substances.

Human exposure modeling is based on the measuring or estimating the concentrations of the substance of interest in each exposure medium, and the time individuals or populations spend in contact with each such medium. Ingestion behavior directly affect the magnitude of exposure to substances present in foodstuffs.

Models have been developed for the multimedia transfer of pollutants, and the uptake by the food chain, leading to estimates of lifetime average daily dose for various scenarios of human category (location and age class). The number of processes modeled is large and the uncertainty in the calculated results, particularly for some of the more-complex pathways, is correspondingly large. Such models as INTEGRA [175] or MerlinExpo [191] are not considering spatial aspects and are specifically not adapted to characterize environmental inequalities.

Implemented exposure model are quantitative constructs that estimate the relationships between measurable events. They are entered into the model as input variables. These inputs causally lead to other events, which are the outputs or result variables of the model. The relationships between the inputs and outputs are described in the model using algorithms, equations and intermediate variables. The modeling is an iterative process used to adapt the data and objectives.

MODUL'ERS, a multimedia model, operating in a GIS environment for quantifying human exposure to toxic substances was developed in INERIS [21]. After analyzing a number of available computational models [20], MODUL'ERS was developed in order to fill in the gap due to the lack of a sufficiently adapted model in the context of evaluating the exposure risk. Therefore, INERIS has developed its own tools from the equations available in the literature to carry out its studies.

In this work, the equations of the model equations were adapted according to data availability and representativeness, once the data inventory was assembled. The spatial dimension is inserted by the coupling of the model with a GIS tool.

The model computes the exposure doses and the risk indicators from multiple exposure pathways and environmental media for the specific human receptor. It consists of a deterministic model core. The adapted equations take into account the local and chronic character of the exposure. The model input is a uniform spatial database, constructed for each compartment, that contains data from different sources.

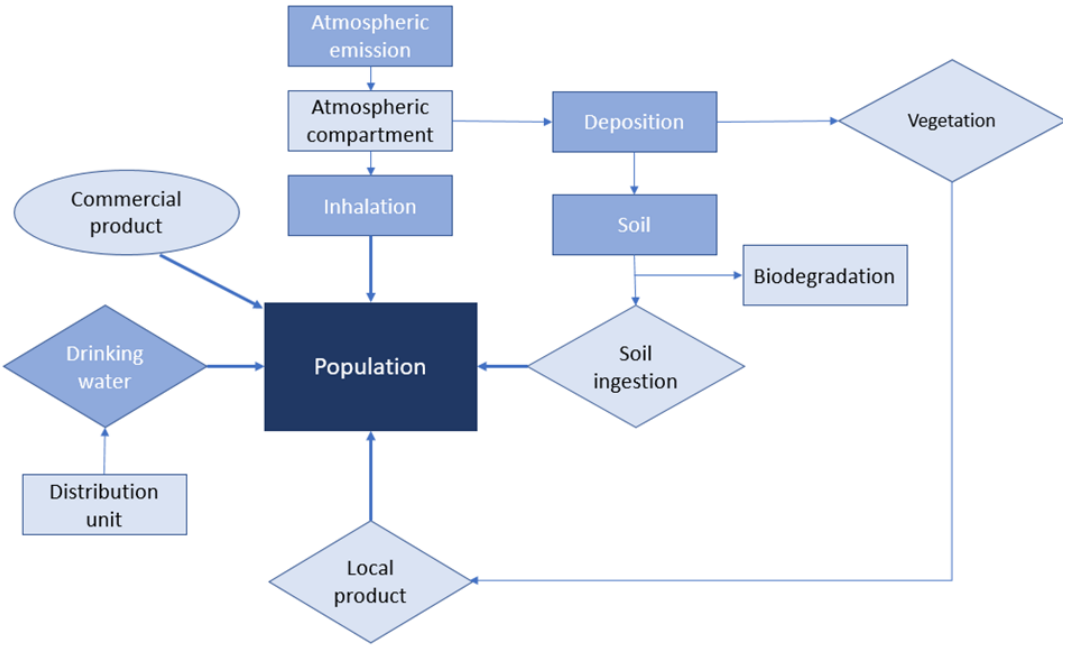


Figure 3.2: Conceptual diagram of exposure pathways and transfers taken into account in the model

3.3.1 Selection of exposure pathways

To assess the exposure dose, predicted concentrations of the substance of interest from local exposure media and national data for commercial foodstuffs are integrated in the model. The pathways of exposure considered here are:

- inhalation of local air;
- ingestion of locally produced vegetables;
- ingestion of commercial product;
- ingestion of soil;
- ingestion of drinking water.

The conceptual diagram (Figure 3.2) shows the environmental compartments, the different exposure pathways and the transfers between them. To characterize the population's exposure to the external environment, only the main routes of exposure (ingestion and inhalation of outdoor air) are considered in the present work.

Dermal contamination is considered negligible compared to other exposure routes. In this work only the aspects related to external exposure are considered.

The Eulerian atmospheric dispersion model (CHIMERE) is used to assess the atmospheric deposition on soil. The pollutant concentration in soil is used to estimate the soil ingestion pathway and transfer in vegetation and animal products. Soil concentration calculations are used to account for substances loss by biodegradation mechanism. Deposition is used to take pollutant soil inputs into account.

Soil concentrations based on soil depth are used to calculate exposure via several pathways:

- the ingestion of local plants contaminated by root uptake;
- the direct ingestion of soil by humans.

The chemical concentration in produce is calculated by summing the contributions of three mechanisms: root uptake, the direct deposition of particles and the indirect deposition of particles due to resuspended particulate matter from the soil. We defined 2 classes to characterize plant media: Root vegetables and Other vegetables (fruit, fruiting, leafy vegetables), The root uptake was calculated using a plant-soil bioconcentration factor, which is defined for each substances and plant class.

The PAH contributions to the diet from commercially produced foodstuffs are calculated using data from the second French Total Diet Survey (EAT2, ANSES, 2011). Except from soil, data measured or modeled around 2010 are used and assumed constant for the entire exposure duration.

3.3.2 Lifetime average daily dose

An Average daily dose (ADD) can be calculated by averaging the dose rate over body weight and an averaging time, provided the dosing pattern is known so the integral can be solved. to have such data for human exposure and intake over extended periods of time, so some simplifying assumptions are commonly used [199]. For effects such as cancer, where the biological response is usually described in terms of lifetime probabilities, even though exposure does not occur over the entire lifetime, doses are often presented as lifetime average daily doses (LADDs). Doses may be expressed in several different ways. The dose per unit of time is the dose rate, which has units of mass/time (e.g., mg/day). Because intake and uptake can vary, dose rate is not necessarily constant. An average dose rate

over a period of time is a useful number for many risk assessments.

The use of LADD as metric of exposure allows the combination of the relative exposure, including the totality of exposure pathways of the ingestion routes (water, soil, diet). This approach also allows the integration of exposures related to background noise of food produced outside the study area. In the calculation, it is also considered the characterization of substances as potential carcinogens. Taking the above into account the daily exposure dose is calculated in the following way:

$$LADD_{i,m} = \frac{C_{i,m} \cdot IR \cdot ED}{BW \cdot LT}$$

where:

- $C_{i,m}$: Mean concentration of a substance (mg/kg) for the substance i for the grid cell m
- IR: Ingestion rate (kg/day)
- ED: Total duration of exposure (here 30 years)
- BW: Body weight (mean weight of population during the exposure period (kg)
- LT: Duration of lifetime (70 years)

The different scenarios of exposure concerning the age class, localization, dietary habits, allow to construct concrete examples in order to quantify the population's exposure to the various pollutants.

3.3.3 Excess of Individual Risk

Cancer slope factors (Inhalation Unit Risk and Oral slope factor) are used to estimate the risk of cancer associated with exposure to a carcinogenic or potentially carcinogenic substance. A slope factor is an upper bound, approximating a 95% confidence limit, on the increased cancer risk from a lifetime exposure to an agent by ingestion or inhalation.

The following approach aims to aggregate the various exposure pathways in a unique, integrated indicator. The daily exposure dose for ingestion, as well as the atmospheric concentrations are compared to the toxicological reference value, usually used for the evaluation of the sanitary risks, in order to obtain the Excess of Individual Risk (EIR) for

the carcinogenic potential.

Increased cancer risk from inhalation exposure

Concerning the effects without a threshold for the exposure through inhalation, the risk is formulated as the EIR, for each one of the substances and for each grid, in the following way:

$$EIR_{inh,i,m} = \frac{CI_{i,m} \cdot IUR_{inh,i} \cdot ED}{AT}$$

where:

- $CI_{i,m}$: Inhalation mean concentration ($\mu g/m^3$) for a substance$i$ for the grid cell m
- $IUR_{inh,i}$: Inhalation Unit Risk $(\mu g/m^3)^{-1}$ for a substance$i$
- ED: Total duration of exposure, here considered 30 years
- AT: Time period over which the dose is averaged. For carcinogenic substance, the duration considered is 70 years

Increased cancer risk from oral exposure

For the ingestion, the risk of the substances without a threshold is formulated as the Excess of Individual Risk in the following way:

$$EIR_{ing,i,m} = \frac{LADD_{i,m} \cdot OSF_{ing,i} \cdot ED}{AT}$$

where:

- $LADD_{i,m}$: Lifetime average daily dose $mg \cdot kg^{-1} \cdot d^{-1}$ of a substance$i$ for the grid cell m
- $OSF_{ing,i}$: Oral slope factor $(mg \cdot kg^{-1} \cdot d^{-1})^{-1}$ for the ingestion pathway for a substance i
- ED: Total duration of exposure, here considered 30 years
- AT: Time period over which the dose is averaged. For carcinogenic substance, the duration considered is 70 years

Finally, the total EIR for a substance i is given as:

$$EIR_{total} = EIR_{inh} + EIR_{ing}$$

3.3.4 Description of the modeling platform

3.3.4.1 Data platform

The construction of exposure indicators for the population requires the integration of national database. The modeling platform with a GIS interface is set in place in order to calculate the exposures due to ingestion and inhalation (Figure 3.2). This platform integrates the available data on the grid of reference after the individual processing of each element of the source-effect chain (Figure 1.1). The relevant data include:

- Spatialized atmospheric emission data and polluted soil localization;
- Spatialized environmental concentration data for each compartment (air, water, soil);
- Environmental spatial layer (topography, erosion,...);
- Population data, describing population density and dietary habits of homogeneous population subgroups of reference;
- The concentrations in the exposure media;
- Exposure doses from multimedia modeling.

The platform is constructed and parameterized by a GIS which allows the management, the analysis, the modeling and illustration of spatially referenced data.

The structure of the model allows to highlight areas with high level of emission, important contamination of environmental compartments or potential overexposure in a fine scale in France. The environmental data used to assemble the spatial database are used as an input for the multimedia exposure model in order to estimate exposure doses, by estimating the exposure in the considered media. For this work ArcGIS software version 10.1 is employed.

3.3.4.2 Computational model

The model core consists in solving the transfer equations. The concentrations for every substance considered need to be defined in each environmental compartment. The exposure scenarios' parameters and the description of degradation phenomena have to also be defined for all areas of interest.

The results correspond to simulations over a 30-year period. During this time, it is considered that there has not been any change in polluting practices and no policy for environmental restoration has been set. Contamination in air and water is considered constant, while for the soil compartment the degradation and the deposits affect the concentrations during the duration of exposure.

The output of the model contains:

- The concentrations of the substances in the local exposure media;
- The intake dose for every population class and each exposure route;
- The EIR and spatialized risk indicators.

The values are given for each grid cell and they are directly importable in a GIS.

Simulations run independently for every spatial entity defined in the spatial databases. The results are exported in a spreadsheet covering each spatial element, easily identifiable by an identification code denoting the common spatial unit.

The model output allows to identify the areas at risk of potential overexposure, vulnerable population classes, the local or generalized character of the exposure, the pollutants and the exposure media whose contribution to the calculation risk is more important.

3.3.4.3 Setting the model parameters

The model's parameters concerning the dietary habits in France, are mainly retained from the CIBLEX database.

The CIBLEX database is constructed in order to provide descriptive parameters of the behavior of the French metropolitan population for a given spatio-temporal scale in an effort to offer the experts in charge of risk assessments, a common framework in the definition of reference groups on a specific spatial and temporal reference. Depending on the type of data, two spatial division systems are proposed, related either to the administrative description of

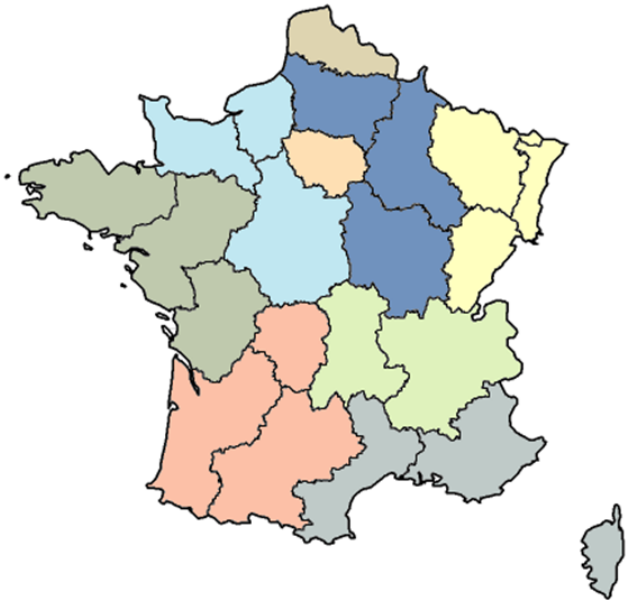


Figure 3.3: Subdivision of French territory into 9 economic zones for spatial planning (ZEAT)

the territory (region, department, municipality) or to its statistical description (ZEAT: area of study and spatial planning, region and fragment of agricultural region). The division in ZEATs is presented in Figure 3.3.

The database makes it possible to describe socio-demographic profiles (age classes, morpho-physiology) and behavioral profiles (food consumption, space-time budget). The different diets are presented based primarily on the age of the individuals concerned. CIBLEX (ADEME, 2003) provides, according to various criteria, an amount of food ingested daily for food categories. Food consumption for the ages 0 to 3 are based on various dietary suggestions from recent studies in France, in the national level. Dietary information for the ages 3 to 60 are derived from the Individual and National Food Consumption Study (INCA 1999: [209]) carried out by AFSSA. This survey focuses on the food consumption of 2491 individuals, representative of the French population. To describe the ingestion, two levels of geographic aggregation of data are proposed: a national level and a level corresponding to the ZEATs defined by the INSEE (the ZEAT of Parisian Basin has been divided into two due to its size to Parisian Basin east and Parisian Basin west). The food consumption values -presented in CIBLEX- do not represent an average for the entire population, but only for consumers, that is to say those who actually consumed at least one of the products during the study. The consumption quantities of the different aliments make it possible to integrate more realistic dietary scenarios.

Concerning the construction of age subgroups, two age groups are identified to describe more realistic exposure scenarios. Based on data from the CIBLEX database on individuals' food consumption, these groups are:

- Children and teenagers 2-17 years old;
- Adults 17-70 years old.

These categories determine the associated data that correspond to:

- Water and food consumption;
- Body weight.

The food behavior provided by the CIBLEX study for each ZEAT (Figure 3.3), is described for the defined age classes. This breakdown makes it possible to integrate sub-regional disparities in food consumption according to homogeneous zones in terms of diet.

To integrate commercial and local food production in our exposure scenario, we use the production rate of locally grown food as a variable in the generic equation to calculate the chemical intake. Those rates characterize the portion of food consumed that is supposed to come from the selected area of exposure. The remaining part is called non-local consumption. For this purpose, for each of the food categories, a percentage of local production is defined. This percentage is calculated from data from the INSEE survey [118]. It corresponds to the difference between the total consumption of food per person and the consumption given excluding consumption.

The auto consumption factor used in this study to characterize the consumption of homegrown products versus commercial ones was taken from the work of Caudeville [33].

3.4 Soil spatial data processing

3.4.1 Introduction

Deposition of PAH near emission sources such as traffic, production of creosote, cement, asphalt, and petrochemical industry can display complex spatial patterns depending on prevailing weather conditions (e.g. wind direction, wet and dry deposition), the local topography and the characteristics of the source.

Soil can become a reservoir for non-degraded PAHs, and interactions with some soil constituents can favor accumulation. Therefore, soil is a key environmental compartment regulating their fate and recording evidence of the contamination process [208]. Failure to accurately capture this complexity for environmental inequality characterization can lead to misidentifying populations at risk [227] from exposure to environmental intake concentrations due to inhalation of surrounding air, unintended ingestion of soil, and cultivated crops or other vegetation. The information available for areas impacted by source typically soil measurements could be complemented by data about the source and its environment, such as polluted site localization or soil properties [71].

Often, the datasets of PAH concentrations in soil tend to be small such that, even with the help of advanced geostatistical methods, they do not suffice for a detailed modeling of the spatial distribution of contaminants.

3.4.2 Available Data

Soil compartment data were collected from different sources. Measurements of PAH substances are available through the French Soil Monitoring Network (RMQS) [208], for the year 2010 3.4. The database contains concentrations of 16 PAH substances, measured in the topsoil (0-30 cm) in 1710 sample locations on a systematic 16 x 16 km regular grid across the 550.000 km^2 of French metropolitan territory. Trace amount measurements that were not able to be precisely measured are reported as values under the limit of detection (usually 0.01 $mg \cdot kg^{-1}$). The percentage of observations under the detection limit varies for each substance.

Localization of polluted sites in France are available from the BASOL database (French Ministry in charge of the environment) and are used for the construction of an auxiliary variable. This database constitutes an inventory of the contaminated sites, to orientate preventive or curative actions of the administrative authorities and it contains information on substances present in the contaminated sites. It consists of 788 points spread in the French territory that permits to localized potential PAH contaminated sites.

Finally, the environmental covariates derivedfrom the INRA database include 14 exhaustive covariates in relation to the soil forming factors available in 500 m resolution: coverage of forest area and semi-natural environment, ecoclimate, erosion, slope, typology, slope aspect, compound topographic index, mean evapotranspiration, water infiltration index, meadow and forest type vegetation, first level dominant parent material code, maximum net primary productivity, available water capacity and roughness [144].

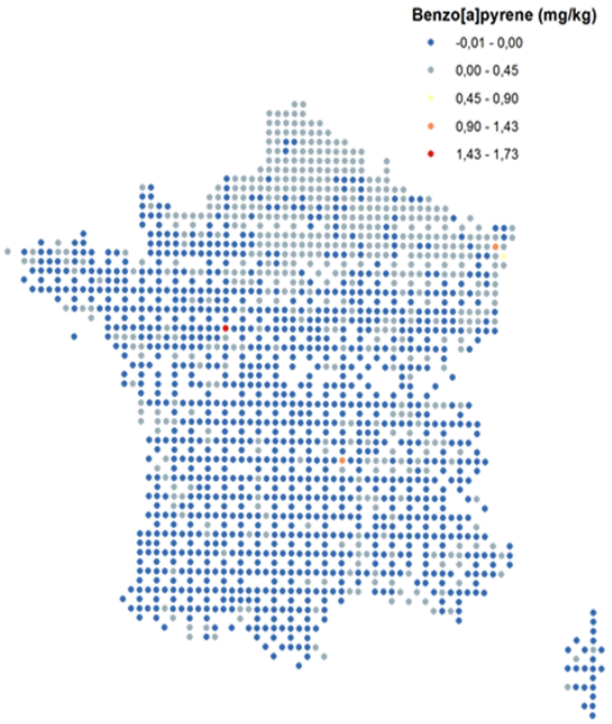


Figure 3.4: Topsoil available measurements for Benzo[a]pyrene in France, on a 16-km grid. The negative values correspond to the data under the detection limit

In order to explain and map the spatial variation of PAH substances in soil, as well as predict the spatial concentrations in France, the above data will be used as covariates in the modeling process or as a starting point to build pertinent auxiliary variables.

3.4.3 Exploratory analysis

The initially available data for Benzo[a]pyrene on the 16-km regular grid for the French territory are presented in Figure 3.4. The negative values correspond to the data under the detection limit. The negative values correspond to the values under the detection limit.

The values under the detection limit vary in the available data. For the three PAH substances of interest (benzo[a]pyrene, benzo[ghi]perylene and indeno [1,2,3-cd]pyrene). For benzo[a]pyrene, the percentage of values under the detection limit is 67.8%. The distribution of the available values is presented in (Figure 3.4).The negative concentrations again correspond to the values under the limit of detection, while the majority of the observed concentrations falls in the [0, 0.01] $mg \cdot kg^{-1}$ interval.

The 14 available soil properties, available for the French territory that will be examined

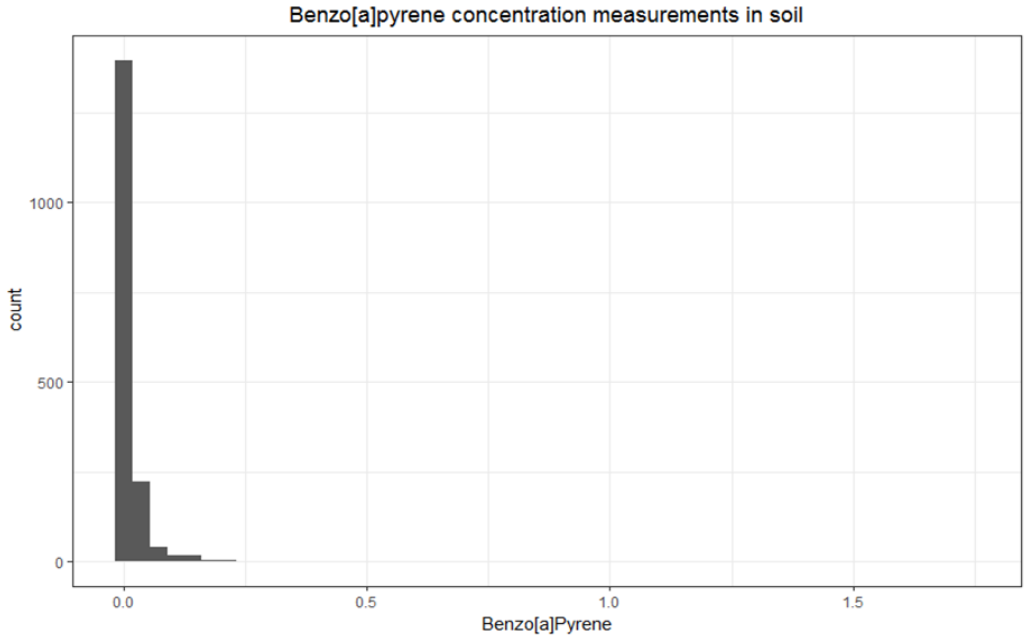


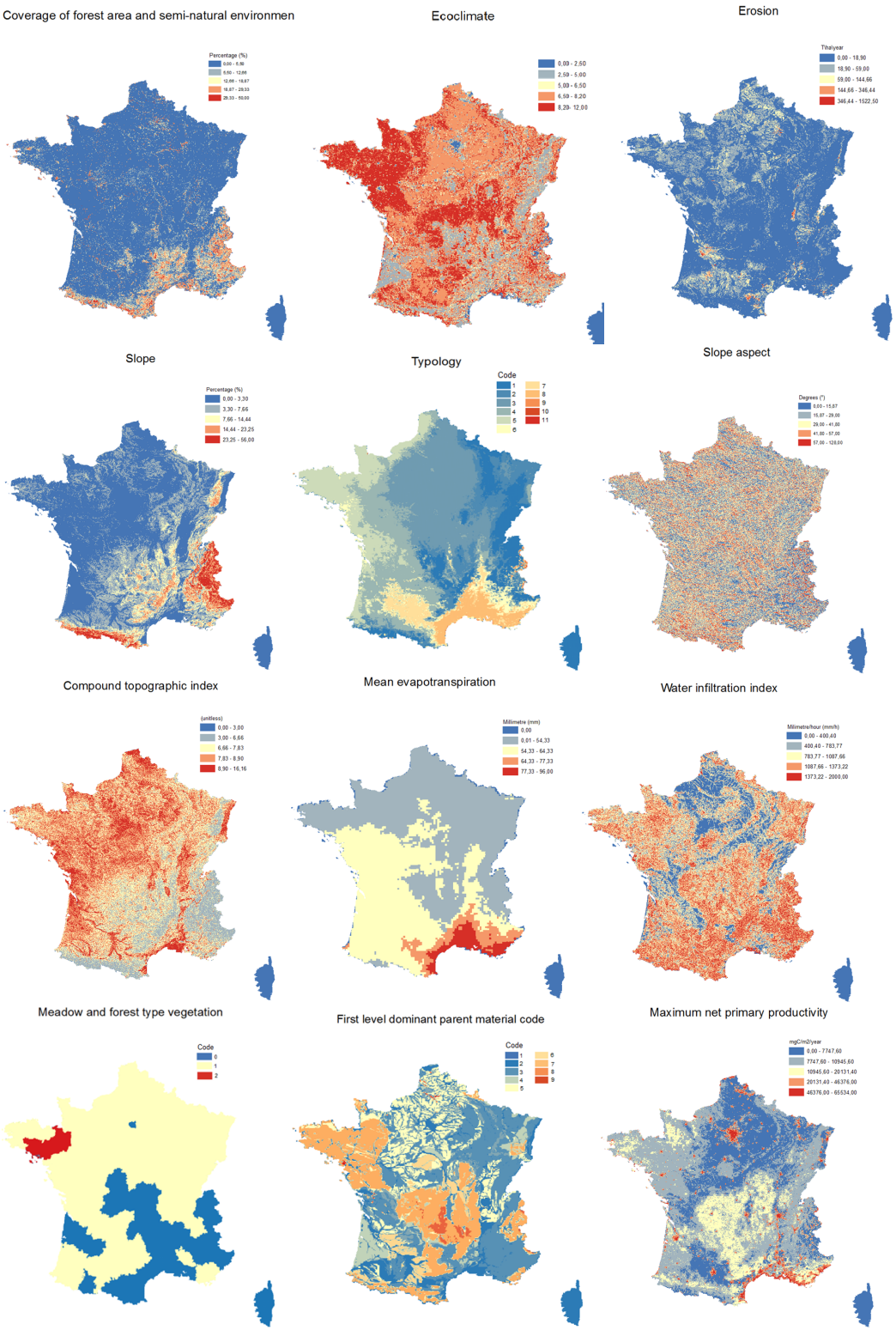
Figure 3.5: Histogram of concentrations measured in soil for Benzo[a]pyrene (mg/kg)

as potentially interesting covariates to spatialize the exposure to PAH, are presented in Figure 3.5. The available data on the polluted sites is a binary variable, denoting by the value 1 if a site is polluted and by 0 otherwise. For the further analysis, only the identified polluted sites are retained.

3.4.4 Methodology

The objective here is to estimate the concentration of B[a]P in the topsoil on the 9 km^2 reference grid, considering data from different available sources in order to increase data resolution.

Regression followed by kriging of the residual [153] is employed to interpolate the concentration measurements between the sample locations. This method is popular when predicting spatial concentrations of pollutants in soil and contamination mapping [6; 16; 125], as it allows to integrate secondary information from auxiliary variables in the prediction, exploiting linear relationships between the variables. The candidate covariates for the model include an indicator variable representing the probability of exceedance of the detection threshold, a function of the distance between the measurements and the polluted sites and a number of soil properties that may affect the spatial distribution of PAH in soil. Knowledge of the processes controlling the spatial variation of the property can be included in the


Coverage of forest area and semi-natural environmen
Ecoclimate
Erosion
Slope
Typology
Slope aspect
Compound topographic index
Mean evapotranspiration
Water infiltration index
Meadow and forest type vegetation
First level dominant parent material code
Maximum net primary productivity

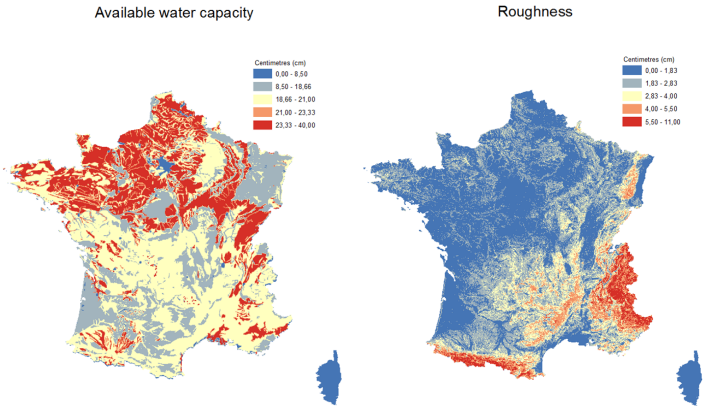


Figure 3.5: Soil physio-chemical properties to be potentially included in the model in the modeling of PAH concentrations as auxiliary variables in order to improve the representativeness.

model by selecting appropriate covariates. For example, we expect the variation of B[a]P in French soils to be influenced by the geological setting, inputs from anthropogenic activities such as agriculture, industry and mining and the transport and deposition of these elements from these sources. Therefore, covariates reflecting those processes, namely a classification of parent material, a classification of land use, the average annual precipitation and the average annual potential evapotranspiration, are introduced in the kriging.

The observations under the detection limit will be included in the modeling with the help of an auxiliary variable. The auxiliary variable is constructed using Indicator Kriging (IK) [68], a method that allows the transformation of continuous data in distinct classes.

Instead of assuming a normal distribution at each estimate location, IK builds the Cumulative Distribution Function (CDF) at each point based on the behavior and correlation structure of indicator transformed data points in the neighborhood. To achieve this, IK needs a series of threshold values between the smallest and largest data values in the set (for example [0, DL]. These threshold values, referred to here as IK cutoffs, are used to numerically build the CDF of the estimation point. For each IK cutoff, data in the neighborhood are transformed into 0s and 1s: 0 if the data are greater than the threshold, and 1s if they are less. IK then estimates the probability that the estimation point is less than the threshold value, given this neighborhood of transformed data and a model of the IK cutoff correlation structure. Performing this operation for each cutoff across the range of data approximates the cdf at the estimation point. Let $Z(x)$ be the random variable and$z$ its realization. The random variable Z is characterized by the probability distribution $F(x;z) = P\{Z(x) \leq z\}$. Employing IK with a single cutoff (here, the detection limit DL)

the probability of the measurements to be below a threshold value is modeled, by applying the nonlinear discrete variable transformation:

$$I\left[Z\left(x_i\right);DL\right]=\begin{cases}1, \text{ if } \mathrm{Z}\left(x_i\right)\leq DL\\ 0, \text{ otherwise}\end{cases}$$

Indicator variogram is then computed using the transformed data. The indicator variogram is estimated as:

$$\widehat{\gamma}\left(I\left(h;Z_{\mathrm{DL}}\right)\right)=\frac{1}{2N\left(h\right)}\sum_{a=1}^{N(h)}\left\{i(x_a;z_{\mathrm{DL}})-i(x_a+h;z_{\mathrm{DL}})\right\}^2$$

where $i\left(x_a;z_{\mathrm{DL}}\right)=\begin{cases}1, \text{ if } \mathrm{z}\left(x_a\right)\leq DL\\ 0, \text{ otherwise}\end{cases}$.

At this point, a least-squares estimate of the Conditional Cumulative Distribution Function (CCDF) is made at the cutoff. The CCDF ($F(x_a;z_{\mathrm{DL}}\,|n)$ is constructed by putting together the IK estimates for the cut-off. The CCDF signifies a probabilistic model for the uncertainty around the unknown value $z(x)$ [94]:

$$\left[F\left(x_a;z_{\mathrm{DL}}|n\right)\right]^*_{\mathrm{IK}}=\left[I\left(x_a;z_{\mathrm{DL}}\right)\right]^*_{\mathrm{OK}}=\sum_{a=1}^{N(x)}\lambda_a^{\mathrm{OK}}\left(x_a;z_{\mathrm{DL}}\right)I(x_a;z_{\mathrm{DL}})$$

The weights λ_a^{OK} are the results of the linear equations of the ordinary kriging system.

The estimator of the unknown z value is the mean of the CCDF, while the variance of the CCDF is considered as a predictor and an associated measure of uncertainty:

$$\widehat{z}_E(x)=\overline{z}_{\mathrm{DL}}[F(x;z_{\mathrm{DL}}|n)])$$

$$\sigma^2\left(x\right)=\left[\overline{z}_{\mathrm{DL}}-\widehat{z}_E\right]^2\left[F\left(x;z_{\mathrm{DL}}|n\right)\right]$$

A second auxiliary variable is built using the identified polluted sites data from BASOL database. In this case, the aim is to model the relation between the measurements and the polluted sites, by a distance decay function. To build the auxiliary variable, an analysis dataset is built by estimating the distance between the polluted sites and the measurement points. In order to determine the distance decay function ($\frac{1}{\sqrt{d}}$, $\frac{1}{d}$, $\frac{1}{d^2}$) than can explain

better the soil concentration variability, the Akaike criterion is minimized.

For each pollutant, the two auxiliary variables, along with the 14 soil properties retained from the INRA database will be used in a geostatistical model in order to spatially predict the concentrations of PAH in soil. Regression kriging is a framework usually used to predict the variable of interest Z through another variable that is exhaustively known in the area. A linear model is normally fitted between the variables, and the residuals of the regression are interpolated using ordinary kriging. The residual error is given by: $r(x) = Z(x) - Z^*(x)$, consequently, the variable $r(x)$ inherits the spatial variability of $Z(x)$.

Supposing Z is the variable of interest and$y$ the auxiliary variable known in the totality of the area. The linear regression model is of the form:

$$Z(x) = m(x) + T(x) = ay(x) + b + T(x)$$

with $E[T(x)] = 0$ and $E[Z(x)] = ay(x) + b$, where T are the real residuals.

The auxiliary variable$y$, is known for the model points as well as the grid points that$Z$ need to be predicted. Then:

$$Z(x) = m^*(x) + W(x) = a^*(x) + b^* + W(x)$$

where a^* and b^* can be estimated using, for example, the least squared method and W being the model residuals. Using this model, the variable$Z$ is calculated by the calculation of residuals: $W = Z - m$

An alternative to the linear regression is the use of a machine learning algorithm, namely the random forests. The use of linear regression implies assumptions about the distribution of the data which cannot always be fulfilled: data following a normal distribution, homogeneity of variance and independent of one another (no autocorrelation). On the contrary, regression trees allow the interactions between the different predictors variables and they are automatically incorporated into the regression model tree, while irrelevant predictors are excluded. That way complex, nonlinear interactions between the variables are easier to accommodate than in linear regression modeling [77]. Growing a regression tree is based on the subdivision or partitioning of the space into smaller groups where the interactions of the variables are easier to manage, and then again until simple models can be fitted.

Let y be the continuous response andx_i the predictors. At first, the space is split into

two regions on the basis of a rule of the form $x_i \leq s_i$, and modeling the response using the mean of y in the two regions. The optimal split is the one that reduces the residual sum of squares RSS:

$$RSS = \sum_{\text{left}} (y_i - y_L^*)^2 + \sum_{\text{right}} (y_i - y_R^*)^2$$

where: y_L^* is the mean$y$ value for left node and y_R^* for the right.

The split is selected over all variables$i$ and all possible split points$s$. The process is repeated recursively until a some stopping rule is applied. The end result is a piecewise constant model over the partition R_m of the form:

$$f(x) = \sum_m c_m I(x \in R_m)$$

where c_m is the mean of y_i in region m, (i.e. the mean of y_i for $x_i \in R_m$.

Breiman (1984) [25] extended the concept of regression trees by generating hundreds of regression trees, known as "random forests", which are based on a random selection of a subset of data from the training set (bootstrap sampling). The numerous regression trees fitted on the data are then averaged in order to predict the target variable with the smallest MSE [126].

A tree-predictor collection is given by

$$\{\widehat{h}(., \Theta_\ell), \quad 1 \leq \ell \leq q$$

where $(\Theta_\ell)_{\{1 \leq \ell \leq q\}}$ are iid random variables, independent the data L_n. $L_n = \{(X_1, Y_1), \ldots, (X_n, \ Y_n)$ are iid random variables, with the same distribution as (X, Y), where $(X^1, \ldots, \ X^p) \in R$ are the input variables and $Y \in R$ the response variable. Then, $\widehat{h} : R^p \to R$ is the random forests predictor, obtained by aggregating the collection of trees:

$$\widehat{h}(x) = \frac{1}{q} \sum_{\ell=1}^{q} \widehat{h}(x, \ \Theta_\ell)$$

The candidate covariates to be added in the model include the two auxiliary variables constructed, in addition to the 14 environmental covariates. The importance of the variables is assessed in the random forest framework, while the covariates selected are the ones that minimize the out-of-the-bag (OOB) error rate, which is defined as:

$$errOOB = \frac{1}{n} \sum_{i \in \{1,\ldots,n\}} (y_i - \widehat{y}_i)^2$$

Once the regression models (linear and random forests) are fitted, the variogram of the residuals is modeled. In this case, a spherical model was used with the three standard parameters; nugget,c_0, partial sill c_1 and range r:

$$\gamma(h; c_0, c_1, r) = \begin{cases} 0 & h = 1 \\ c_0 + (c_1 - c_0)\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right) & h \leq r \\ 1 & h > r \end{cases}$$

To assess the performance of the different models, leave-one-out and k-fold cross validation were employed.

3.5 Air spatial data processing

3.5.1 Introduction

The atmosphere is the most important mean of PAH dispersal, it receives the bulk of the PAH environmental load resulting in PAHs being ubiquitous in the environment [1]. PAHs are emitted to the atmosphere primarily from the incomplete combustion of organic matter. The combustion sources can be either natural (volcano eruptions, forest fires) or anthropogenic. While the anthropogenic sources include vehicle exhaust, agricultural fires, power plants, coke plants, steel plants, foundries and other industrial sources, PAHs tend to be found in greater concentrations in urban environments than in rural environments as most PAH sources are located in or near urban centers.

As PAH is persistent in the air compartment, assessing the exposure due to inhalation, is a crucial step when calculating the total exposure. However, concentrations measured in the air monitoring network don't always provide a sufficient basis to spatialize and characterize the exposure due to the limited number of monitoring stations. In this context, the objective of this work is to spatialize the air concentration measurements of PAH over the French territory on the grid of reference, compensating for the limited number of observations by including in the spatial prediction information from a correlated auxiliary variable.

3.5.2 Available Data

The PAH concentrations in the air compartment are available from the French national air quality database. The data are collected in France for each region in the context of regulatory surveillance. The available database consists of 76 and 77 annual mean concentration point measurements for the years 2010 and 2011, respectively. The monitoring stations are categorized according to the type of area in which they are located and the predominance (or absence of predominance) of near-by emission sources. Five categories are distinguished here: rural background (designated as rural), suburban background (suburban), urban background (urban), traffic and industrial. Often, at the national level, only background data are considered. However, for this study all types of stations are considered, including those that are locally influenced, as the objective is to identify possible areas of overexposure which requires taking into account the pollution at a local scale.

Atmospheric concentrations of PAHs on a grid 7 x 7 km in France are also available from the Eulerian atmospheric dispersion model CHIMERE. The CHIMERE multi-scale model was primarily designed to produce daily forecasts of ozone, aerosols and other pollutants and make long-term simulations (entire seasons or years) for emission control scenarios [136]. It runs over a range of spatial scale from the regional scale (several thousand kilometers) to the urban scale (100-200 km) with resolutions from 1-2 Km to 100 Km. The EMEP (European Monitoring and Evaluation Program) air emission inventory was used to describe emissions in the Eulerian dispersion modeling of PAH concentrations over the French territory.

In addition, cumulative annual emissions of three PAH substances (benzo[a]pyrene, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene) for 2007 from the National Spatialized Inventory are available on a 0.01° x 0.01° grid in France. They include the emissions from different industrial and non-industrial sources (combustion, use of solvents, urban, traffic, etc). To construct the database of emissions for 2012 the industrial sources were subtracted from the data of 2007 in order to add the industrial sources of 2012. The available emission data are for 2007 will be used as an auxiliary variable for the modeling of concentrations of 2010 and those of 2012 for the modeling of concentrations of 2011.

Moreover, population data based on INSEE 2011 official data and spatially distributed according to the methodology described in the work of Letinois [107] are available on a 1 km^2 grid. Data describing the altitude are available in the form of centroids of French municipalities through the National Inventory.

3.5.3 Exploratory analysis

An exploratory analysis is first employed in order to gain a deeper understanding of the available data before passing to the modeling. Summary statistics and visualization of the available data are employed in order to gain some potentially useful insights. The exploratory analysis is presented for the Benzo[a]pyrene substance.

The Benzo[a]pyrene annual concentrations for the two years (2010, 2011) were not always taken in the same monitoring stations (Figure 3.7). Small variations are noticeable between the two years, indicating slight improvement for the year 2011, however the areas with the highest concentrations remain the same for both years. Concentration above the European target value of 1 ng/m^3 (Directive 2004/107/EC) are less frequent for year 2011 (Figure 3.7 and Figure 3.8). Higher concentrations have been mainly observed in the vicinity of industrial emission sources (Figure 3.6).

The relevance of available auxiliary variables as possible determinants of the exposure is first assessed by examining the correlation between them and the monitored concentrations for the two years.

The concentration values predicted by the CHIMERE model [42] are significantly lower than those of the actual data. Even though there is some coherence in the structure of the predicted CHIMERE concentrations and the observed ones, there is less variability amongst the first ones, while the maximum values are scaled back. This is illustrated by Figure 3.10 where the high values observed for the year 2010 are significantly underestimated by the CHIMERE model, while measured and modelled values are more consistent in the low concentration range.

The annual emission data initially available on a 0.01° x 0.01° grid, are transformed to a 3 x 3 km grid (Figure 3.9). The emissions are measured in kg and have a big range, being higher in highly industrialized areas, close to highways and in urban environment. These are also he areas where the higher annual measurements are observed (Figure 3.7).

The altitude and population are also available in France (Figure 3.9). The altitude seems to be better correlated with the concentrations measured especially for the year 2010 and less for the year 2011 (Figure 3.10). Concerning the population, the relation between the variables does not appear to be linear, however the highest concentrations of Benzo[a]pyrene are observed in the areas where few people reside (Figure 3.10). The correlation between the measurements and the output of CHIMERE model for the year 2010 is examined as well (Figure 3.11). In the plot, it is clear that the high values observed

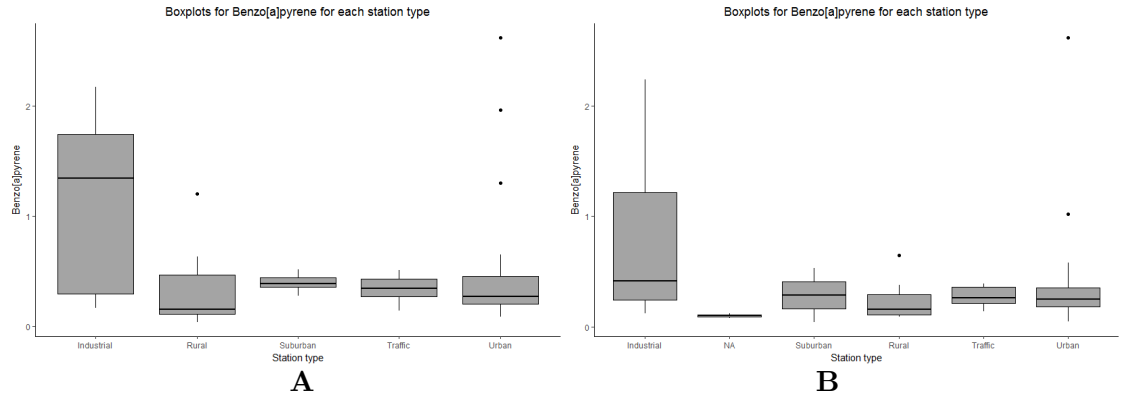


Figure 3.6: Measured concentrations of B[a]P (ng/m^3), by station type, for 2010 and 2011.

for the year 2010 are significantly underestimated by the CHIMERE model, while measured and modelled values are more consistent in the low concentration range.

The relationships between the variables examined in this initial exploratory analysis will be employed to increase the low representativeness of data due to the limited number of monitored observations, in the process of spatialization over the French territory.

	Benzo[a]pyrene 2010	Benzo[a]pyrene 2011	Benzo[a]pyrene (CHIMERE)
Min	0.035	0.040	0.024
Median	0.303	0.260	0.060
Mean	0.479	0.363	0.068
Variance	0.260	0.188	0.001
Max	2.621	2.620	0.544

Table 3.2: Summary statistics of Benzo[a]pyrene concentrations (ng/m^3) in France; 2010 and 2011; Output concentrations from the CHIMERE model (2010).

	Emissions 2007	Emissions 2012	Population
Min	0.00	9.85×10^4	0.00
Median	7.49×10^6	7.60×10^6	2.20×10^2
Mean	1.16×10^7	8.11×10^6	9.74×10^2
Variance	4.93×10^{14}	1.26×10^{13}	2.37×10^7
Max	3.95×10^9	3.67×10^8	3.11×10^5

Table 3.3: Summary statistics of Benzo[a]pyrene emissions (kg) in France; 2011 and 2012; Summary statistics for population.

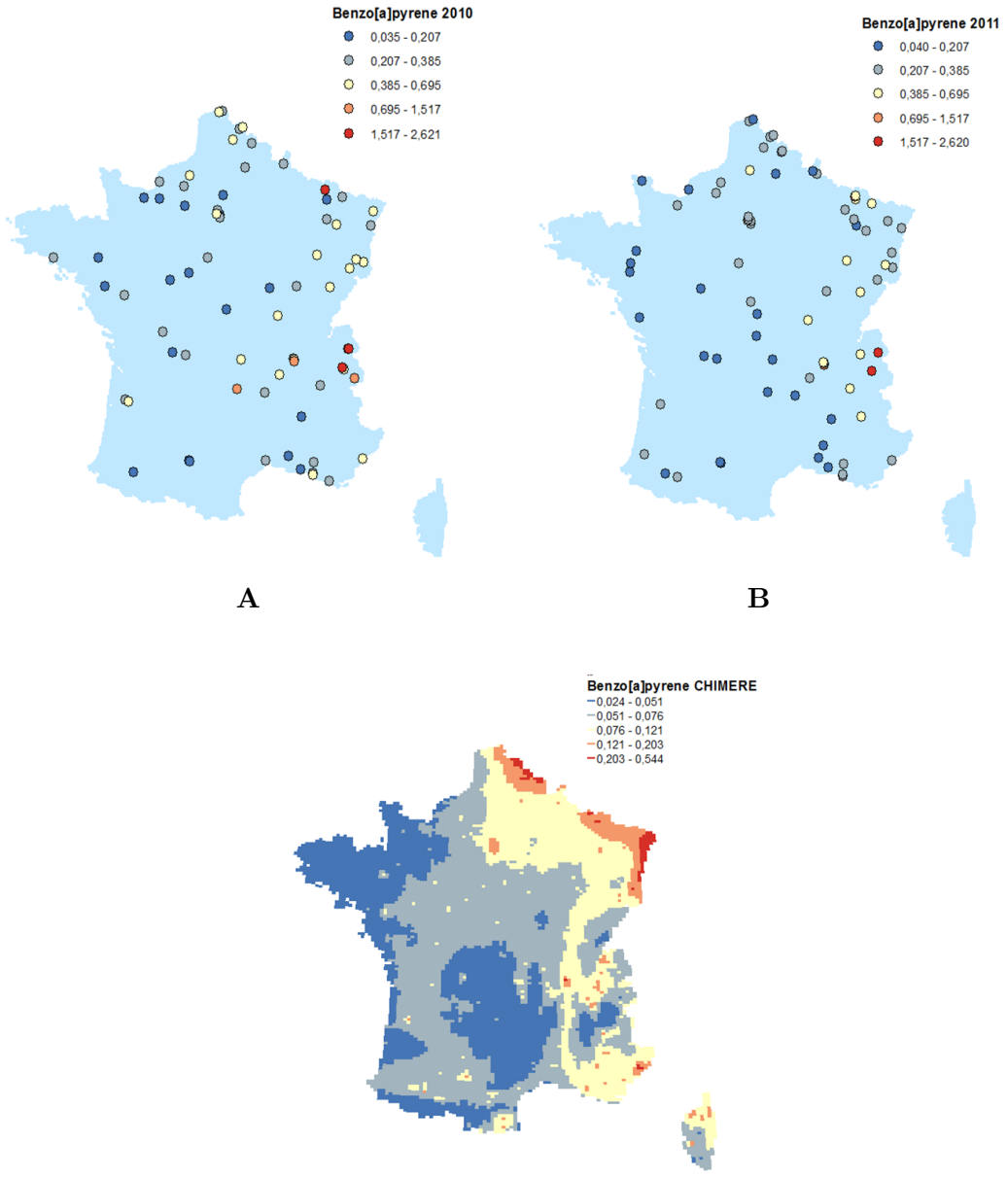


Figure 3.7: Annual mean concentration measurements of Benzo[a]pyrene (ng/m^3) in France; (A) 2010 and (B) 2011; (C) Simulated concentrations of Benzo[a]pyrene from the CHIMERE model (ng/m^3) (2010) [42].

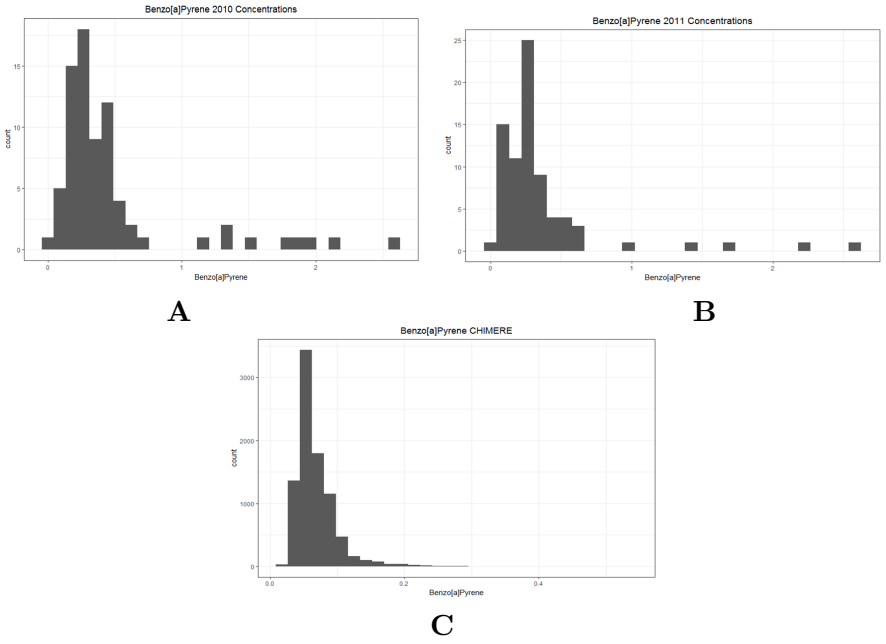


Figure 3.8: Histograms of Benzo[a]pyrene (ng/m^3) concentrations measured in monitoring stations in France; (A) 2010 and (B) 2011; (C) Modeled output concentrations from the CHIMERE model (2010).

3.5.4 Methodology

The main issue with the available PAH measurements for the air compartment is the limited number of observations present. Attempting to directly spatialize PAH concentrations over the whole French territory by ordinary kriging could possibly lead to a misrepresentation of the phenomenon. In an effort to compensate for this limited number of data, information from auxiliary variables can be included in the kriging prediction. This is possible using the method of regression kriging with auxiliary variable. With this method, auxiliary variables exhaustively known on the area of interest are introduced in the model in order to statistically explain the variable of interest. The auxiliary variables and the variable we want to explain need to be well-correlated.

Let Z be the variable, only known for a small set of points in the area of interest, then regression kriging allows the prediction of the variable in the ensemble of the area through another variable, which is exhaustively known. The variable Z is modeled as a random function $Z(x)$ and the predictor as a deterministic variable $q(x)$. The two variables are expected to be linearly related. Then, the estimation at the new locations is obtained by:

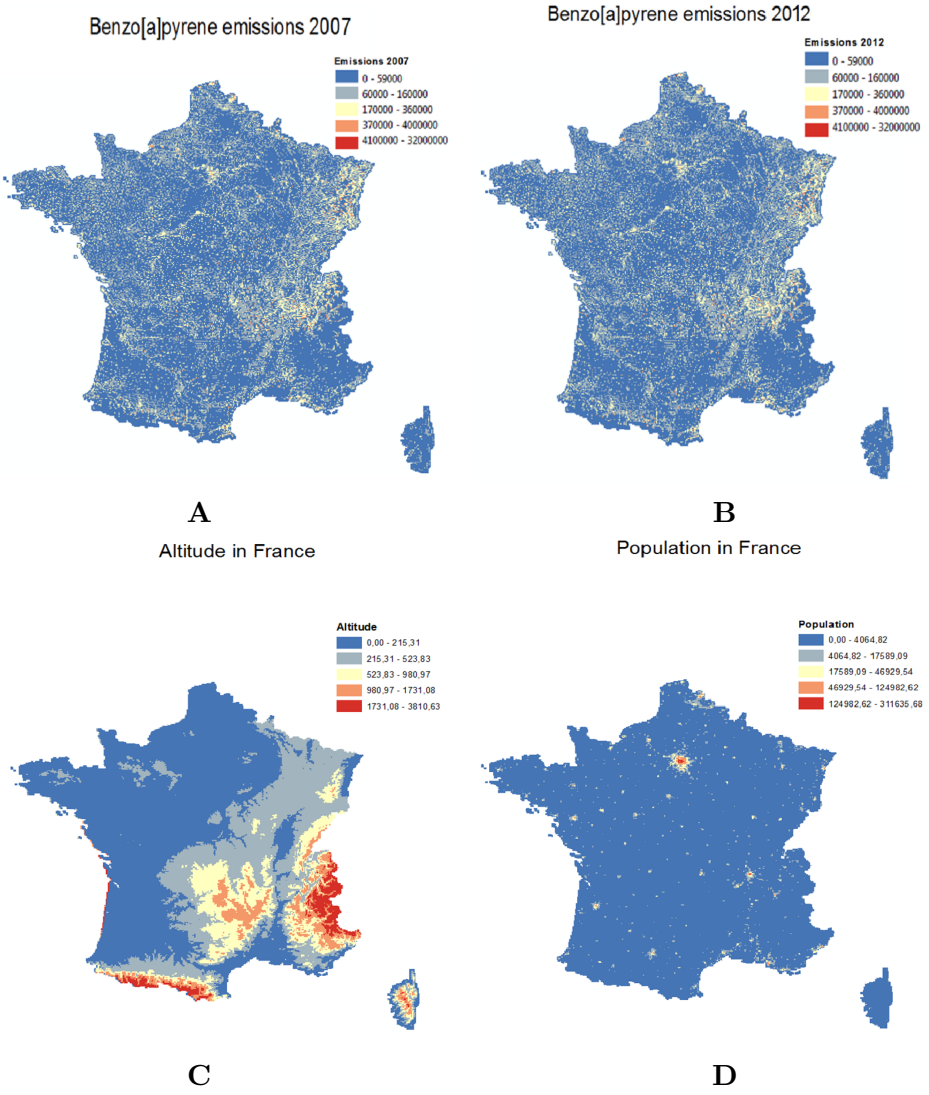


Figure 3.9: Annual emissions of Benzo[a]pyrene (kg) in France (2007 and 2012); Altitude (m) and Population.

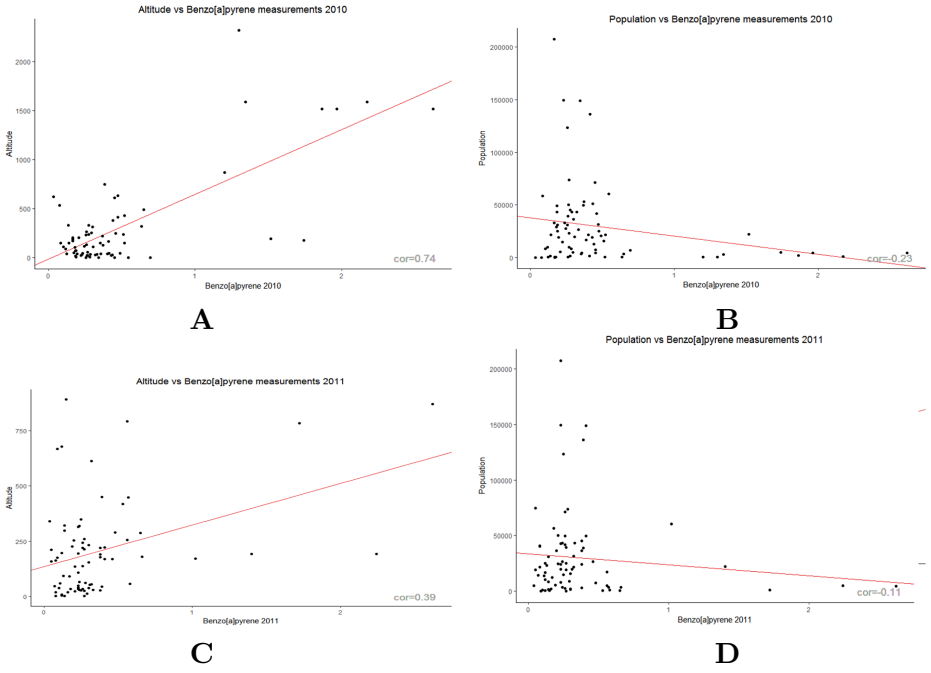


Figure 3.10: Scatterplot and correlation between the altitude and population and the measured concentrations of Benzo[a]pyrene for the years 2010 (A & B) and 2011 (C & D).

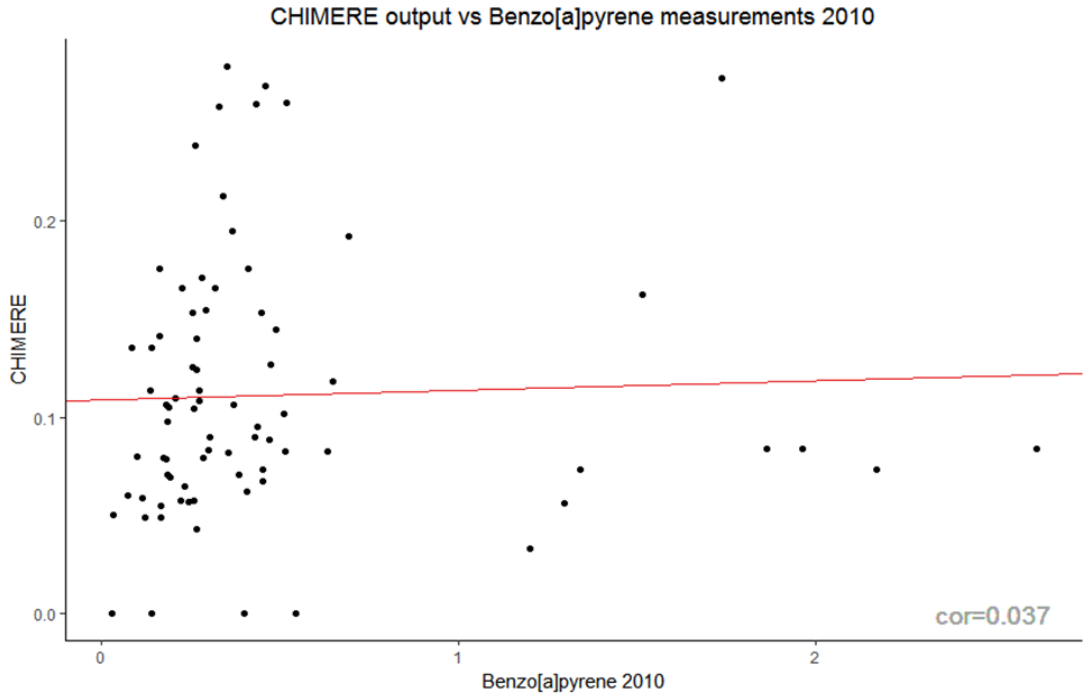


Figure 3.11: Scatterplot and correlation between the concentration of Benzo[a]pyrene, measured in the monitoring station in 2010 and the modeled output for the same year by CHIMERE model.

$$\widehat{Z}\left(x_0\right) = \sum_{i=1}^{n} w_{\mathrm{i}}^{\mathrm{RK}}(x_0) \cdot Z(x_i)$$

where $\sum_{i=1}^{n} w_{\mathrm{i}}^{\mathrm{RK}}\left(x_a\right) \cdot q_k\left(x_i\right) = q_k\left(x_a\right)$, for $k = 1, \ldots, p$ with z being the target variable, q_k the p number of predictors and w_i^{RK} the weights.

The weights can be solved using the extended matrices:

$$\lambda_0^{\mathrm{RK}} = \left\{w_1^{\mathrm{RK}}\left(x_0\right), \ldots, \ w_n^{\mathrm{RK}}\left(x_0\right), \phi_1\left(x_0\right), \ldots, \phi_p\left(x_0\right)\right\}^T = \mathbf{C}^{RK-1} \cdot \mathbf{c}_0^{\mathrm{RK}}$$

where λ_0^{RK} is the vector of solved weights, ϕ_p are the Lagrangian multipliers, $\mathbf{C}^{\mathrm{RK}}$ is the extended matrix of the residual covariance and $\mathbf{c}_0^{\mathrm{RK}}$ is the extended vector of covariances at new location:

$$\begin{bmatrix} C\left(x_1, x_1\right) & \ldots & C\left(x_1, x_n\right) & 1 & q_1\left(x_1\right) & \ldots & q_p\left(x_1\right) \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ C\left(x_n, x_1\right) & \ldots & C\left(x_n, x_n\right) & 1 & q_1\left(x_n\right) & \ldots & q_p\left(x_n\right) \\ 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ q_1\left(x_1\right) & \ldots & q_1\left(x_n\right) & 0 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ q_p\left(x_1\right) & \ldots & q_p\left(x_n\right) & 0 & 0 & \ldots & 0 \end{bmatrix} c_0^{\mathrm{RK}} =$$

$$\left\{C\left(x_0, x_1\right), \ldots, C\left(x_0, x_n\right), q_0\left(x_0\right), q_1\left(x_0\right), \ldots, q_p\left(x_0\right)\right\}^T ; q_0\left(x_0\right) = 1$$

The variogram is modeled using the semivariance function, but for computational efficiency the covariances are used, since when solving the kriging equation system, both the semivariance and covariances matrices give the same results. The relation between the covariance and semivariance is given by [185]:

$$C\left(h\right) = C_0 - C_1 - \gamma(h)$$

where $C(h)$ is the covariance and$\gamma(h)$ is the semivariance function. The covariance at zero distance is by definition equal to the mean residual error [43]. The distance $C(x_1, x_1)$ is equal to the variance:

$$C\left(0\right) = C_0 + C_1 = Var\{Z\left(x\right)\}$$

Finally, the RK variance, i.e. the variance of the estimation error is calculated as the weighted average of the covariances from the new point x_0 to all calibration points $(x_1, \ldots, x_n)$, plus the Lagrange multipliers [218]:

$$\sigma^2_{\mathrm{RK}}(x_0) = (C_0, C_1) - \left(\mathbf{c}_0^{\mathrm{RK}}\right)^T \cdot \lambda_0^{\mathrm{RK}} = C_0 + C_1 - \sum_{i=1}^{n} w_i(x_0) \cdot C(x_0, x_1) + \sum_{k=1}^{p} \phi_k(x_0) \cdot q_k(x_0)$$

In order for the spatial prediction to be accurate, the linear drift (auxiliary variable(s)) needs to be well defined. To construct a pertinent variable, the factors that influence the concentration measurements need to be determined. Different possible auxiliary variables were tested in order to define the best model.

In a first approach, the CHIMERE dispersion model output was employed as an auxiliary variable. The use of the output of a dispersion model as an auxiliary variable in an external drift model is a common practice [211; 121]. However, its pertinence as an auxiliary variable needs to be addressed.

Another auxiliary variable was constructed using the yearly PAHs emissions for year 2010, available through the National Spatialized Inventory. First, the emissions are disaggregated in a buffer zone around the PAH measurement locations. Different buffer radius are tested in order to determine the optimal bandwidth considering the decaying correlation between the emissions and the actual measurements as the distance grows. In order to determine more precisely the decaying relation between the emissions and the measurements two kinds of function are tested: exponential and inverse decay:

- Inverse distance decay: $f(d) = \frac{m^a}{d^\beta},\ \beta \geq 0$
- Exponential distance decay: $f(d) = m^a \cdot exp(-\beta d)$

Here m is the emissions, d the distance and α, β parameters to be estimated, always considering the correlation between the emissions and the measurements.

Other possible determinants of the PAH exposure are also tested in the process. These include the population and the altitude of the emissions. Linear and nonlinear combinations of these variables are tested in an effort to build a representative auxiliary variable.

The experimental semi-variogram is traced, and a model is fitted for each external drift model:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left(Z(x_i) - Z(x_i + h)\right)^2$$

where $h = (x_1, x_2) = (x_i, x_i + h)$ and N the number of observations.

The optimal model is selected as the one that minimized the Root Mean bSquare Error (RMSE) for each station type and the average.

To combine the results for the two years, a weighted average is employed. The weights are derived for each year, based on the accuracy of the individual predictions [82]. Specifically, the kriging weights are estimated the following way:

$$w_1 = \frac{\sigma_2^2 - \rho_{12}\ \sigma_1\ \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2} \qquad w_2 = \frac{\sigma_1^2 - \rho_{12}\ \sigma_1\ \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}$$

where σ_1^2 and σ_2^2 are the kriging prediction errors and ρ is the correlation coefficient between the prediction errors.

3.6 Water spatial data processing

3.6.1 Introduction

Contamination of water may lead to intake of PAH through drinking water and cooked foods [120]. The levels are usually below 1 ng/L in drinking water. However, this level can be higher when asphalt or coal tar coating of storage tanks and water distribution pipes are used. As drinking water is an important factor of the human exposure to potentially toxic substances, characterizing its quality is an essential step when estimating the total exposure to PAH substances.

Due to their hybrophobic nature, PAHs are found in water in small concentrations, therefore is not always possible to quantify the exact concentration considering the limitations on the sensitivity of the appropriate equipment. Then instead of an exact value, the limit of detections is reported. In this study, the objective is to estimate the concentrations of PAH substances in potable water in the grid of reference, taking into account the values that are too small to be detected (values under the detection limit). The concentration measurements are retained from the SISE'Eaux database [45], taken in the water distribution facilities in France.

3.6.1.1 Available Data

The concentration measurements of PAH substances in potable water in France which are retained from the French regulatory monitoring database, SISE'Eaux [45] are collected in the framework of monitoring and evaluating the quality of the distributed water in France, by the French ministry of health.

The concentrations are measured in the 24655 water distribution units (UDI) that serve all 28809 municipalities, constituting the water distribution network in France. The latter is constructed in such way that a single distribution unit can serve more than one municipality, while one municipality can be served by more than one unit, yielding that way 39062 combinations amongst them. The network complexity needs to be addressed before proceeding with the spatialization in order to get more accurate results.

The database includes 8 PAH substances: anthracene, benzo[a]pyrene, benzo[b]fluoranthene, benzo[k]fluoranthene, fluoranthene and indeno[1,2,3-cd]pyrene, as well as 2 indicators of cumulative concentrations of 4 and 6 substances respectively. The measurements were collected during the years 2000 and 2012, however there is no periodicity. The number of observations per substance is highly inconsistent as not all substances were measured at the same time and at the same units. The non-regular temporal sampling will be accounted for in the process of estimating the multi-annual PAH concentration means.

As it is often observed in the environmental databases, the rate for observations under the detection limit is quite high. In this database, it varies between 70-95% (Table 3.4). This high rate of values under the detection limit requires careful handling, in order to extract the maximum information from the available measurements, without introducing too much bias in the final results.

3.6.2 Exploratory analysis

The number of observations of the 8 substances and 2 cumulative indicator variables measured in potable water in France, as well as their rate of values under the detection limit vary greatly (Table 3.4). The limits of detection vary not only among the different substances but also within a single substance.

The substances in the benzo family were generally measured in the same water distribution facilities and at the same time, yielding that way a homogenous sample. In these substances, it is also observed the highest rate of under the detection limit. Within this family of substances important correlations are also observed (Table 3.5), that can be later

Substance	Observations	% values UDL
anthracene	33581	91.4 %
benzo[a]pyrene	162918	88.1 %
benzo [b] fluoranthene	161748	90.4 %
benzo [ghi] perylene	161627	92.4 %
benzo [k] fluoranthene	161640	91.1 %
fluoranthene	110380	73.8 %
indeno [1.2.3-cd] pyrene	159872	93.6 %
naphthalene	26171	91.7 %
hpat4	106166	79.8 %
hpat6	48980	59.2 %

Table 3.4: Number of observations and detection limit for each substance and indicator included in the water PAH database.

used in the imputation of values under the detection limit.

Substance	S1	S2	S3	S4	S5	S6	S7	S8	I1	I2
Anthracene	.	.79	.79	.98	.79	.95	.96	.15	.08	.97
benzo[a]pyrene	.79	.	.93	.71	.93	.89	.39	.03	.06	.35
benzo [b] fluoranthene	.79	.93	.	.73	.89	.89	.56	.03	.06	.31
benzo [ghi] perylene	.98	.71	.73	.	.75	.96	.56	.17	.08	.58
benzo [k] fluoranthene	.79	.93	.96	.75	.	.89	.56	.03	.07	.23
fluoranthene	.95	.89	.89	.96	.89	.	.55	.17	.02	.98
indeno [1,2,3-cd] pyrene	.96	.39	.56	.56	.56	.55	.	.17	.01	.02
naphthalene	.15	.03	.03	.17	.03	.17	.17	.	.05	.93
hpat4	.08	.06	.06	.08	.08	.07	.01	.05	.	.37
hpat6	.97	.35	.31	.58	.23	.98	.02	.93	.37	.

Table 3.5: Correlations between the PAH substances, measured in water.

In order to account for the complexity of the French water network, first it is important to identify its structure. The spatialization of the data will be carried out in a municipality level, therefore it is essential to associate the concentrations measured in each distribution facility with the district they serve. Most facilities as well as municipalities appear only once in the database (Figure 3.12), and the majority consists in 1-1 relationships.

The population of each municipality and the population that each facility served are available and will be used to aggregate the concentrations in the French territory. However, since data on the population served by each UDI for each municipality are not available, extra modeling is required. To estimate those populations, three cases are identified:

- UDI serves only one municipality: $\text{Population}_{\text{UDI/municipality}}=\text{Population}_{\text{UDI}}$

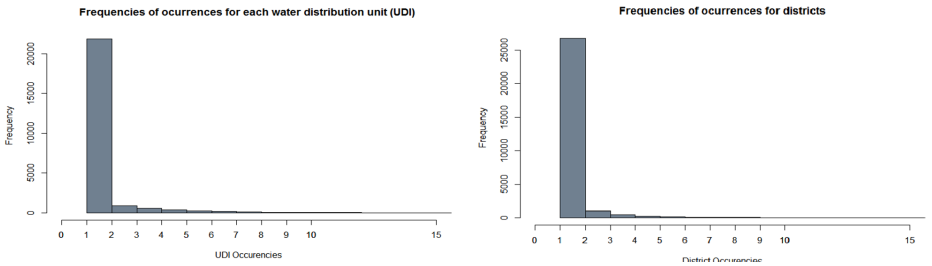


Figure 3.12: Histogram of the occurrences in the database for the water distribution units (A) and districts (B).

- UDI serves more than one municipality and the municipality is served by only one UDI: $\text{Population}_{\text{UDI/municipality}}$=$\text{Population}_{\text{municipality}}$

- UDI serves more than one municipality and the municipality is served by multiple UDIs: solve the system that occurs between the populations and the communes.

The first two cases concern the majority of the data, and they are easy to be attributed. The third cases accounts approximately for 8% of the data, and a system of equations is to be solved. For a municipality i, which is served by 3 UDIs that system would be:

$$\begin{cases} pop_{comm_i} = \sum_{j=1}^{3} pop_{UDI_{j/i}} \\ pop_{UDI_{1/i}} = pop_{UDI_1} \\ pop_{UDI_{2/i}} = pop_{UDI_2} - \sum_{k=1}^{n} pop_{UDI_{2/k}} \\ pop_{UDI_{3/i}} = pop_{UDI_3} - \sum_{l=1}^{m} pop_{UDI_{3/l}} \end{cases}$$

Certainly, there is some uncertainty associated to these estimations. Even when the relation is 1-1, there are cases where the population served by the UDI and the population of the municipality don't correspond. In this case, the population of the UDI is retained. To quantify the uncertainty introduced from these assumptions, the absolute difference between the population of municipalities initially available and the population of the municipalities estimated as the sum of the populations served by each facility for the given municipality is computed.

3.6.3 Methodology

In the process of spatializing the multi-annual drinking water concentrations there are three issues that need to be addressed:

- Include the observations under the detection limit, without introducing great bias in the model.
- Account for the temporally irregular sampling when estimating the mean multi-annual concentrations for each UDI.
- Estimate the mean multi-annual concentration for each municipality, taking into account the network complexity.

A multiple imputation method is employed to substitute the left censored measurements. The imputation method selected here is consists in a non-parametric bootstrap based expectation-maximization (EM) algorithm, the EMB.

The first step of the approach consists in randomly drawing vectors of means from an appropriate posterior distribution to account for the estimation uncertainty. The EMB algorithm replaces the complex process of random draws from the posterior by the resampling method of nonparametric bootstrapping, where samples are taken from the originally sampled data. The nonparametric bootstrap uses the existing sample data (size = n) as the pseudo-population and draws resamples (size = n) with replacement M times. Supposing data $Y_1, \ldots, Y_n$ are independently and identically distributed from an unknown distribution F, then an estimator of this distribution would be $\widehat{F}$(y), which is the empirical distribution F_n defined as:

$$F_n(Y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq y)$$

where $I(Y)$ is the indicator function of the set Y.

The bootstrap re-samples are generated based on the above equation for F_n. The empirical distributionF_n is a nonparametric estimator of F. Since incomplete data are being bootstrapped, there are good chances that the bootstrapped samples will also contain missing values redeeming the sample biased and inefficient. The EM algorithm is used to refine bootstrap estimates then.

The expectation maximization algorithm is a generalization of maximum likelihood estimation to the incomplete data case and it specifically attempts to find the parameters $\widehat{\theta}$,

that maximize the log probability logP $(x;\theta)$ of the observed data. The algorithm alternates between the two steps, the E-step and M-step. During the E-step, the algorithm selects the functiong_t which lowers the bounds logP $(x;\theta)$ overall, and so that $g_t(\widehat{\theta}^{(t)} = logP(x;\widehat{\theta}\,(t)))$.

In the M-step, the expectation maximization algorithm selects a new set of parameters $\widehat{\theta}^{(t+1)}$, so that it maximizes g_t. When the value of the lower-boundg_t matches the objective function at $\widehat{\theta}^{(t)}$, then logP $\left(x;\widehat{\theta}^{(t)}\right) = g_t\widehat{\theta}^{(t)} \leq g_t\widehat{\theta}^{(t+1)} = logP(x;\widehat{\theta}^{(t+1)})$ which means that the objective function increases monotonically at each iteration of expectation maximization algorithm.

The data are treated as time-series-cross-sectional-data. This means that the imputation model assumes that the missing values are linear functions of other variables observed values allowing that way to take advantage the strong relationships between the substances as manifested by the strong correlations. This is particularly important in the cases where all the measurements for a distribution facility are under the detection limit thus there is no temporal aspect to consider.

Let D be then$\times k$ data set, with D$^{\text{obs}}$ the observed part and D^{mis} the unobserved, with the assumption:$D \sim N_k\,(\mu,\ \Sigma)$, the multivariate normal distribution with mean vector μ and covariance matrix Σ. This model makes the assumption that the missing values are Missing At Random (MAR):$p\,(M|D) = p\left(M\middle|D^{\text{obs}}\right)$, where M is the missingness matrix:$m_{\text{ij}} = 1$ if d_{ij} is non observed, and $m_{\text{ij}} = 0$ otherwise. Then, under the MAR assumption:

$$p\left(D^{\text{obs}}, M\middle|\theta\right) = p\left(M\middle|D^{\text{obs}}\right)p(D^{\text{obs}}|\theta)$$

where $\theta = (\mu,\ \Sigma)$. The likelihood is:

$$L\left(\theta\middle|D^{\text{obs}}\right) \propto p(D^{\text{obs}}|\theta)$$

And:

$$p\left(D^{\text{obs}}\middle|\theta\right) = \int p\,(D|\theta)\,dD^{\text{mis}}$$

Then the posterior is:

$$p(\theta|D^{\text{obs}}) \propto p(\theta)\int p\,(D|\theta)\,dD^{\text{mis}}$$

A flat prior can be used, however due to the high rate of under the detection limit values, the use of a ridge prior no more than1% of the total number of observations is advised. That way, the model accounts for the fact that the simulated observations need to fall in the $[0, DL]$ interval.

A number of diagnostics are employed to evaluate the performance of the multiple imputation model. Moreover, the methodology is applied in a second similar data set, artificially truncated and the results are compared with other widely used methods, such as the substitution of under the detection limit by a constant.

Once the imputation results are retained and the now complete dataset is available, we proceed with the estimation of the concentrations for each water distribution unit. The concentration for each unit is calculated as the multi-annual weighted mean of the multiple available concentrations. The weighted method selected is the method of segments of influence, where the assigned weights are proportional to the time interval separating the measurements. It is a simple method for calculating the temporal weights but it is shown to yield comparable results to more complicated methods [14].

The method consists in tracing the median between two neighborhood events' dates and assigning a weight equal to the segment to which is included, divided by the total number of days between the starting date (1/1/2000), and the finish date (31/12/2012), so 4834 days in total (Figure 3.13). The main advantage of this method is in its simplicity to implement however, In the case of a "closed" sampling, with measurements located at the extremities of each section, the weights assigned to the first and last measure are calculated differently from the others. Nonetheless, it is possible that this does not entail a large difference in the estimation, especially when the measurements are numerous. Moreover, this method doesn't take into account the temporal correlation, but only for the irregular sampling.

Once the weights are assigned, the multi-annual PAH concentration measurement for each unit can be calculated as follows:

Let C_{in}^{t} be the drinking water concentration of the i PAH substance for the$n$ water distribution unit at the t time instance, then the multiannual mean for the given unit is given as:

$$C_{\text{ij}}^{n} = \frac{\sum_{t=1}^{T} C_{\text{in}}^{t} \cdot w_t}{\sum_{t=1}^{T} w_t}$$

where T is the total number of temporal observations registered for the distribution

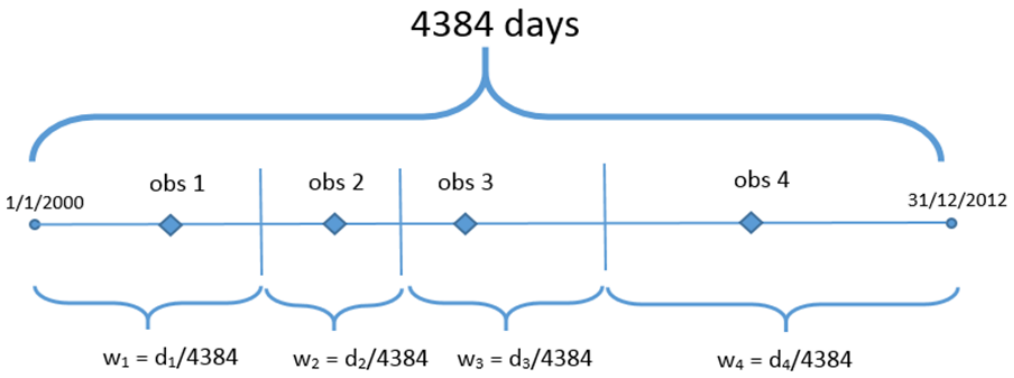


Figure 3.13: Schematic representation of the temporal weighted method "segments of influence" (Bernard-Michel, 2006).

unit and w_t the associated weight of observation t.

Finally, to estimate the concentrations for each municipality with accurate precision with regard to the complexity of the water distribution network, the concentrations calculated for the water distribution units are weighted by the populations affected by each unit, by each municipality that have been previously calculated and summed.

Let N be the total number of water distribution units that serve commune m, and p_n the associated population that the unit serves. Then, the concentration of the PAH substance i for the district m is calculated as the sum of the concentrations of the N units that serve the district, weighted by the population p_n:

$$C_{\text{im}} = \frac{\sum_{n=1}^{N} C_{\text{ij}}^{n} \cdot p_n}{\sum_{n=1}^{N} p_n}$$

Once the concentrations for each municipality are calculated, the results are spatialized by georeferencing them on the 3 km^2 grid of reference. Specifically, the concentrations calculated are assigned to the characteristic surface of each commune, by intersecting the geographical layers that correspond to the communes and to the grid of reference. The concentrations are then distributed homogeneously on the grid using the surface ratio:

$$C_{ig} = \sum_{m} \frac{C_{ig} \cdot S_{mg}}{S_g}$$

where C_ig is the concentration of a pollutant i in the grid g ($\mu g/l$), C_{im} the concentration of pollutant i in the district m ($\mu g/l$), S_{mg} the surface intersection between the reference

grid and the district m (km^2) and S_g the surface of the grid cell g (km^2).

Chapter 4

Results

In this chapter, the results of the study are presented. In the first section the results of the spatialization methodology employed for each environmental compartment are presented. The spatialized dataset then serves as an input for the multimedia model. The modeling results are presented in the second section.

Those results allow the analysis of the determinants of exposure through the estimation of the contributions of the different exposure routes, environmental compartments and transfers. They can also be used to estimate for which pollutants the risk indicators are elevated and to identify overexposed population. Finally, they allow the mapping of the environmental inequalities due to exposure through the spatialization of risk indicators.

In this study, modeling focused on mainland France for benzo[a]pyrene, benzo[ghi]perylene, and indeno[123-cd]pyrene. The calculation of exposure dose and risk indicators is carried out for the two age classes defined (2-17 years old and 17-70 years old) with the available data described in Chapter 3. A 3-km mesh is considered for the study (around 63,000 grid cells for the total area). The concentrations defined in each environmental compartment (soil, air, drinking water) as well as parameters for the various exposure scenarios are described for each grid cell.

The results correspond to simulations over a period of 30 years, assuming there is no modification on the polluting practices and no environmental restoration policy has been set up. Concentrations in the water and air compartments are considered to be constant over this period, while for the soil compartment the concentrations are considered a balance of losses (degradation phenomena) and inputs (deposits) over the duration of exposure.

Most of the results are presented only for benzo[a]pyrene. The other two substances

of interest, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene are highly correlated with benzo[a]pyrene, and the results are equivalents.

4.1 Spatialization results

4.1.1 Soil

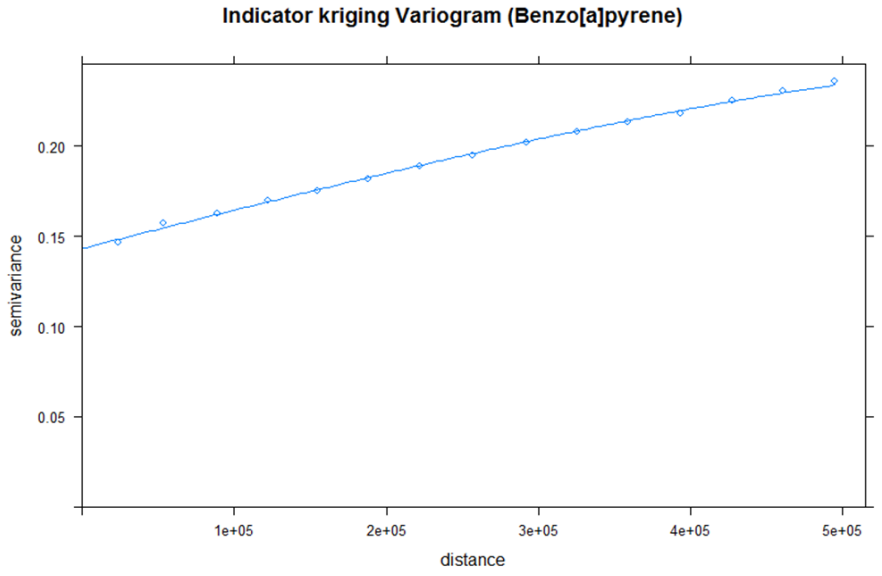
Indicator kriging is a popular method to treat highly skewed variables. In this case, the spatial prediction for the values under the detection limit will be used as an auxiliary variable in the final model. A spherical variogram is used to model the spatial relationship between the observations, with the optimal sill, nugget and range parameters (Figure 4.1). The result obtained is the probability map of the measurements to be under the detection limit in the French territory (Figure 4.1).

A second auxiliary variable is constructed from the available polluted sites data. The optimal buffer radius and the distance decay function is selected according to the Akaike criterion (Figure 4.2 A). The auxiliary variable estimated for the totality of the French territory is presented in Figure 4.2 B.

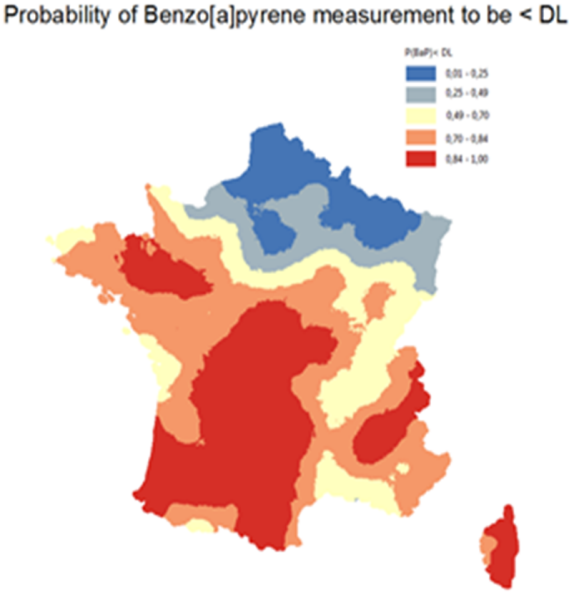
The two modeling approaches (linear and random forest regression) were tested in order to determine the best approach as well as the best model. For the linear model approach, three models are tested: two models with each auxiliary variable constructed and one with both. For the random forest regression, the two auxiliary variables described above, along with the 14 soil properties are initially all used in the regression model.

The calibration of the random forest model is a crucial step of the process. The parameters that need to be defined are the number of trees in the forest (ntree) and the number of variables randomly sampled as candidates at each split (mtry). Concerning the number of variables at each split for regression problems, a usual tactic is to set it as one third of the total number of variables, in this case 16, so the number for variables at each split should not overpass 5, as setting the mtry parameter higher could result in overfitting.

Concerning the number of trees, it should be enough to stabilize the error (error convergence) but not that many that they will over correlate the ensemble, leading to overfitting or unnecessary variance. the mean squared error (MSE) is used as a measure of the prediction accuracy of the random forest model. Two MSE error estimates are used in the validation procedure: the OOB error and the cross-validation error. To tune these parameters, a search is launched, with respect to the minimization of MSE. The parameters

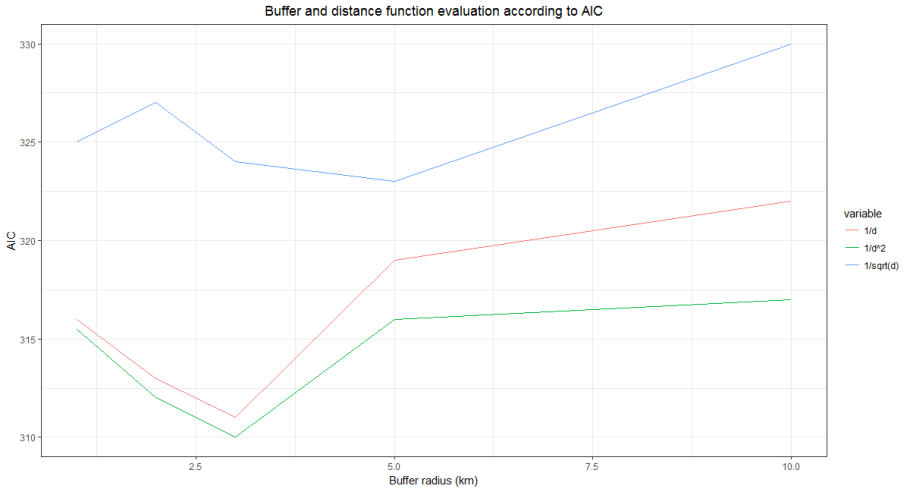


A

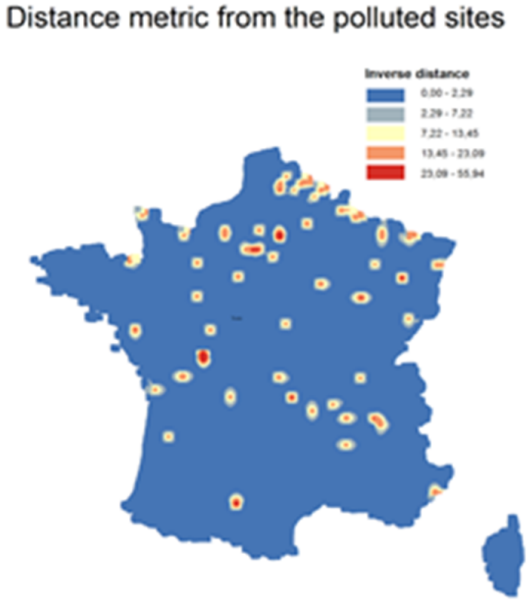


B

Figure 4.1: (A) Variogram for the indicator kriging for Benzo[a]pyrene; (B) Indicator kriging map for Benzo[a]pyrene in France, where the blue color indicates a smaller probability of a measurement to be under the detection limit while the red color indicates a greater probability.



A



B

Figure 4.2: (A) Evaluation of the buffer and distance function according to AIC criterion; (B) Map of the auxiliary variable constructed for the spatialization of Benzo[a]pyrene in France, from the available data on the polluted sites.

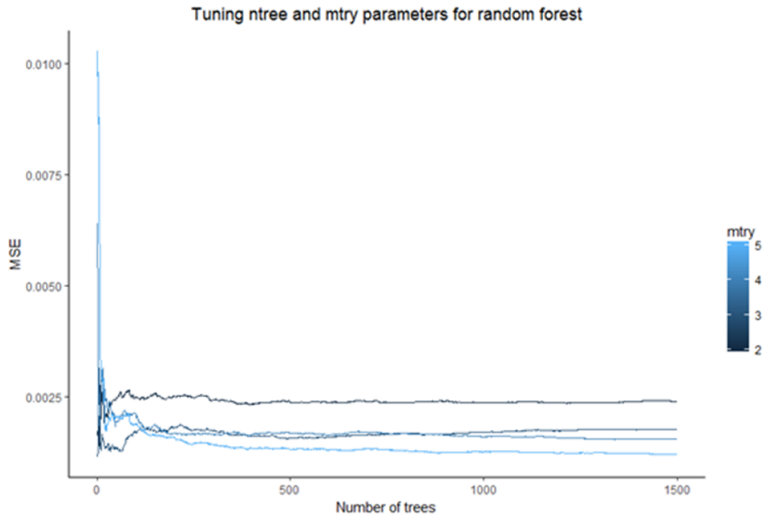


Figure 4.3: Comparison of performance of random forest models by comparing two parameters (number of trees in the forest and number of variables at each split (mtry)), by evaluating the Mean Squared Error (MSE).

selected are 500 number of trees, as the MSE stabilizes after and 5 variables (mtry), as for these values the MSE is the lowest (Figure 4.3).

In the framework of random forests, the 16 covariates are evaluated, in order to obtain a parsimonious random forest regression model that explains the variability of Benzo[a]pyrene sufficiently. The Out-Of-Bag error is employed as a diagnostic (Figure 4.4). The model that satisfies the above is the one including the two auxiliary variables constructed as well as the following soil properties: coverage of forest area and semi-natural environment, surface roughness, slope, slope aspect, mean evapotranspiration and ecoclimate.

The fitted residuals of the different regression models are then used in the regression kriging to spatially predict the PAH concentrations. The coefficient of determination (quotient of the variances of the fitted values and observed values of the dependent variable, R^2) of each regression model as well as the cross validation results are compared in order to determine the best model for the spatial prediction. The results are presented in Table 4.1.

Model	R^2	CV correlation
Random forest: complete (all possible variables)	0.80	0.75
Random forest: optimal (8 covariates)	0.84	0.81
Liner model: complete (all possible variables)	0.16	0.67
Linear model: optimal (8 covariates)	0.17	0.66

Table 4.1: Comparison of the different models, by comparing the R^2 and the cross-validation result

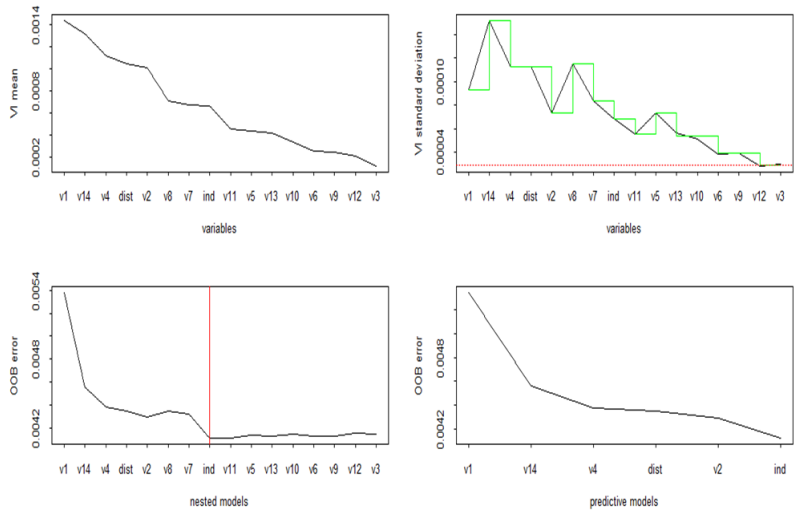


Figure 4.4: Diagnostic plot for random forest model; Top graphs illustrate the thresholding step: Sorted variable importance mean associated with the 16 explanatory variables (top left) and standard deviation (top right); Bottom graphs are associated with the interpretation and the prediction steps respectively.

The percentage of variance explained for the random forest varies between $80 - 84\%$ for the random forest and $16 - 17\%$ for the linear model. The difference in the 5-fold cross-validation after kriging are more similar though (Figure 4.5), as the influence of the spatial variation of the original observations is the dominant factor in the final prediction. The best model selected according the above criteria, is the random forest model with 8 covariates. The final spatialization results are presented in Figure 4.6.

4.1.2 Air

To map the air PAH concentrations, regression kriging is employed for each of the two years for which data is available. In both cases, auxiliary variables are constructed using the annual emission inventory, the altitude and the population, where they are relevant.

Once the emissions are disaggregated in the 20-km buffer, the optimal distance decay function describing the decreasing influence of emissions with distance is determined considering the correlation with the measurements:

- $f_1(d) = \frac{m^6}{d^8}$, for 2010 and
- $f_1(d) = \frac{m}{d^{1/3}}$ for 2011

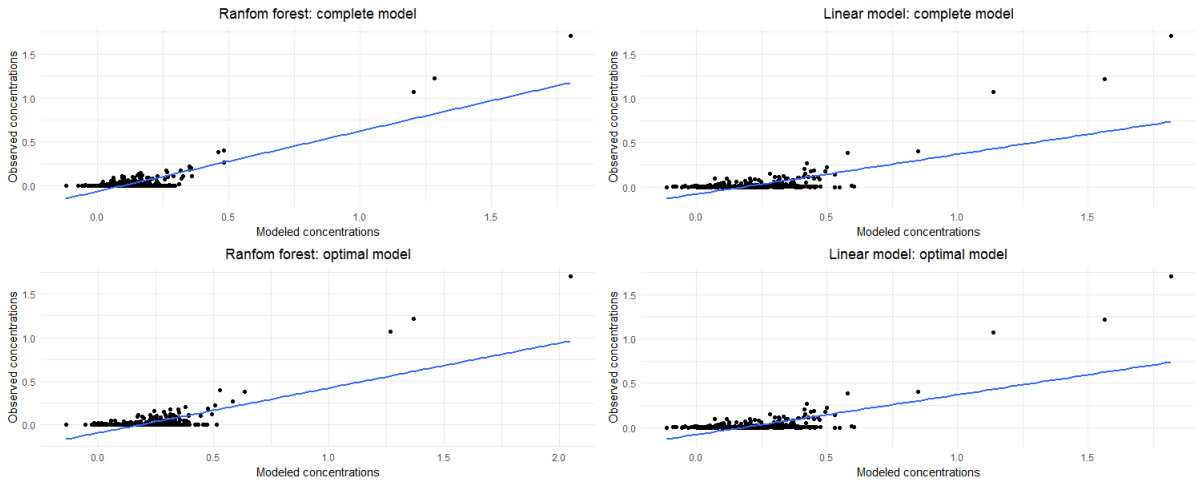


Figure 4.5: Scatter plots of 5-fold cross-validation errors for B[a]P concentration (mg/kg).

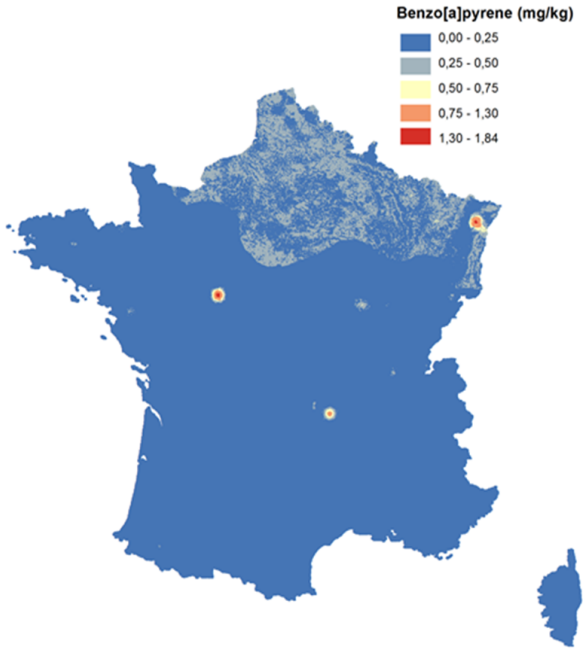


Figure 4.6: Benzo[a]pyrene spatialization results for the French territory, using regression kriging with the random forest regression model.

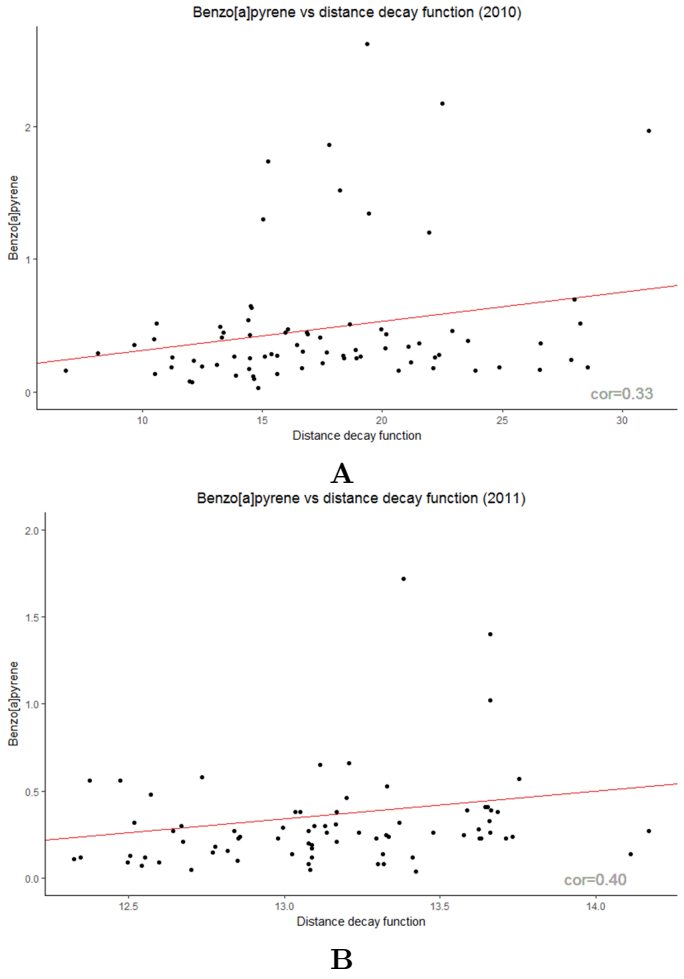


Figure 4.7: Benzo[a]pyrene concentrations vs the optimal distance decay function of atmospheric emissions: (A) for year 2010 and (B) for year 2011.

where m is the emission and d the distance between the emissions and the measurements. The correlation between the above function and the respective measurements are shown in Figure 4.7.

Different combinations of the possible concentration determinants are tested to determine the best model for each year and substance. The output of CHIMERE model is also considered as a potential auxiliary variable. The best auxiliary variables and models for each year are determined according to cross validation results, the average mean quadratic error and the average mean quadratic error divided by the kriging variance. The models tested and the statistics for the two years for Benzo[a]pyrene are presented in Table 4.2 and Table 4.3.

The model retained for the year 2010 is the one that includes the function of distance

from the emission as well as the altitude and the population:

$$C_{\text{BaP}}^{2010} \sim a \cdot \log\left(\frac{m^6}{d^8}\right) + p$$

For the year 2011, using only the function of distance from the emission sources proved to be the best amongst the other models:

$$C_{\text{BaP}}^{2011} \sim \frac{m}{d^{\frac{1}{3}}}$$

The individual results for the two years are presented in Figure 4.8.

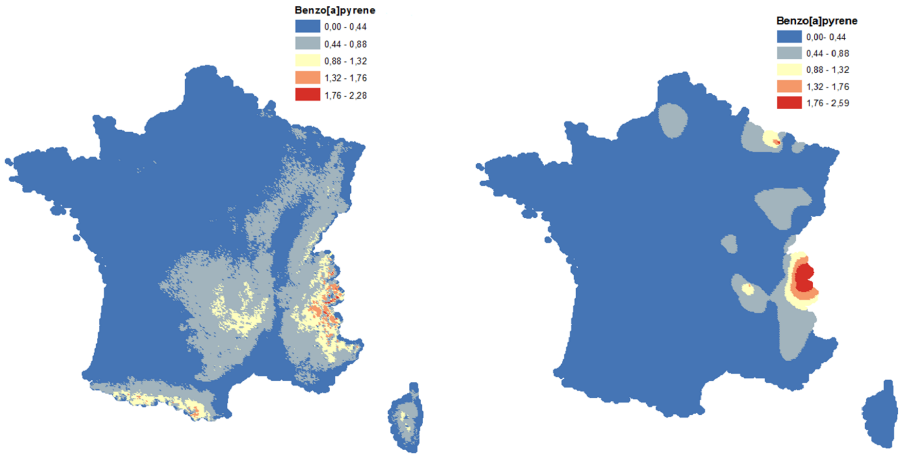
Model	CV correlation	MQE	MQE/Var(kriging)
$\sim a \cdot log(\frac{m^6}{d^8})$	0.71	0.17	2.29
$\sim a \cdot log(\frac{m^6}{d^8}) \cdot p^{\frac{1}{2}}$	0.67	0.18	2.77
$\sim a \cdot \log\left(\frac{m^6}{d^8}\right) + p$	0.74	0.15	1.99
$\sim a \cdot \log\left(\frac{m^6}{d^8}\right) + C$	0.69	0.18	2.41
$\sim a \cdot \log\left(\frac{m^6}{d^8}\right) \cdot p^{\frac{1}{2}} + C$	0.65	0.20	2.50

Table 4.2: Models tested for the year 2010; is the altitude, is the emission, the distance, is the population and the output of the CHIMERE model.

Model	CV correlation	MQE	MQE/Var(kriging)
$\sim \log\left(\frac{m}{d^{1/3}}\right)$	0.21	0.17	2.25
$\sim \log\left(\frac{m}{d^{1/3}}\right) + a$	0.33	0.17	2.36
$\sim \log\left(\frac{m}{d^{1/3}}\right) + p$	0.33	0.17	2.35
$\sim a \cdot \log\left(\frac{m}{d^{1/3}}\right)$	0.22	0.18	2.42
$\sim \frac{m}{d^{\frac{1}{3}}}$	0.40	0.16	2.19
$\sim \log\left(\frac{m}{d^{1/3}}\right) + a + p$	0.34	0.17	2.38

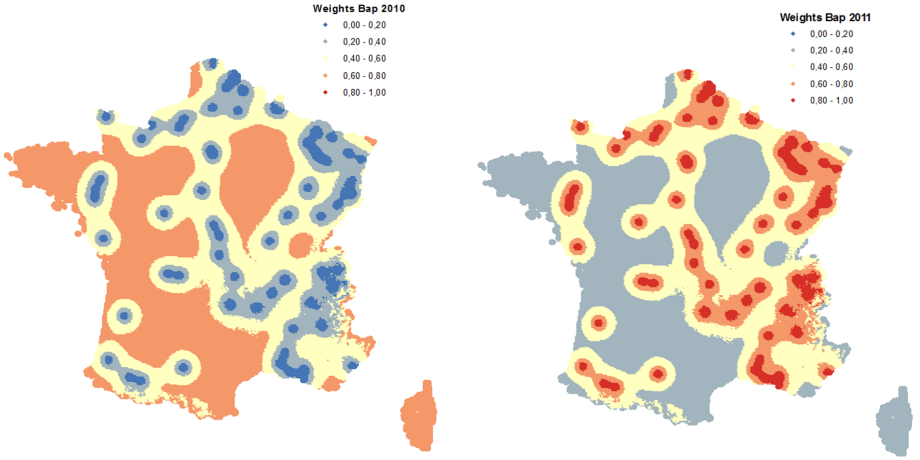
Table 4.3: Models tested for the year 2010; a is the altitude, m is the emission, d the distance and p is the population.

To combine these results in a single map, a weighted average is employed. The weights were calculated with the method presented in section 3.5.4, and are presented in Figure 4.9. The final result is obtained by the weighted average of the two years' individual results and it is presented in Figure 4.10.



A **B**

Figure 4.8: Spatialization results for Benzo[a]pyrene (ng/m^3); (A) Results for 2010; (B) Results for 2011.



A **B**

Figure 4.9: Estimated weights for the two years: (A) 2010 and (B) 2011.

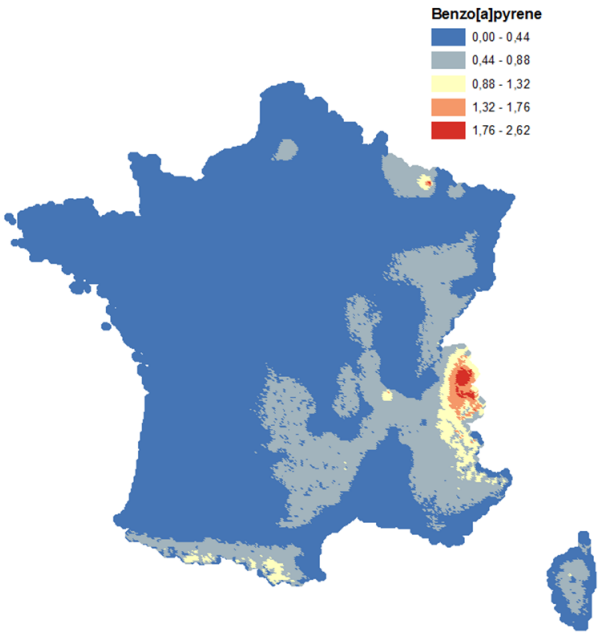


Figure 4.10: Spatialization of Benzo[a]pyrene in air (ng/m^3) by combining the results of the two years in a weighted average.

4.1.3 Water

The multiple imputation and model diagnostics were carried out in R using the Amelia II package [85]. Amelia II performs multiple imputation using an EMB algorithm (Expectation-Maximization algorithm and Bayesian classification model). The assumption of multivariate normality is required for Amelia II and the available data are tested against it, in order to make sure it is fulfilled. To better meet the assumption that variables are missing at random the maximum possible number of substances is to be used for the imputation. In this context, since not all measurements were taken at the same time nor in the same facilities, the substances that are measured in the same water treatment facilities and at the same time (-/+ 24 hours) were grouped together. That way the initial dataset is partitioned in smaller, homogenous datasets concerning time and space (UDIs). The positioning of the treatment facilities (longitude and latitude) was not used in the model. The imputation limits were also inserted in the model, in the form of conditionals, forcing the imputations to stay always below them.

The imputation results varied greatly among the water treatment facilities. In the facilities where all the available measurements were under the detection limit, there was not temporal aspect to consider and the imputation was mainly inferred by their relationship with the other substances included in the model. In this case, the variance between the imputed sets is lower than when more measurements above the detection limit are available (Figure 4.11).

The plausibility of the imputation model is assessed by comparing the distribution of imputed and real values. For each cell that is missing in a variable, the diagnostic will find the mean of that cell across each of the m data sets (m depending on the case) and use that value for the density plot. However, there may be no reason to expect a priori that the distributions will be identical, so these plots are not a definitive test. It is possible however check to see whether the imputed distributions are a reasonable shape and that they peak within the known bounds of the actual distribution of the data [85].

Finally, the disperse function is used to test whether the analysis converged on the global maximum of the likelihood surface and starting values. This diagnostic runs EM chains from multiple starting values that are overdispersed from the estimated maximum. As displayed in the following graphs, the EM algorithm converged on the global maximum of the likelihood, indicating that Amelia was in the right parameter space when choosing starting values for imputation runs (Figure 4.12). Once the complete datasets are retained, the multi-annual concentrations per district are computed: first, the multi-annual mean

concentration of a given pollutant is estimated for each water distribution unit, weighting the individual measurements by the temporal weights already attributed with the method of segments of influence; then in each district those mean concentrations are weighted by the population served by the corresponding distribution units.

The concentrations that are initial spatialized in the district level, and are finally georeferenced in the 3-km grid of reference (Figure 4.13).

4.2 Results of multimedia modelling

Once the spatial database is assembled, the multimedia exposure model is launched. The results of the modeling presented in this chapter include:

- the analysis of the exposure media contributions and the daily exposure doses,
- the national and local contributions for the whole area of study, as well as for particular areas,
- the transfers' and environmental compartments' contributions, and
- the cartography of environmental inequalities.

4.2.1 Exposure media contributions

Analyzing the Lifetime Average Daily Dose (LADD) obtained by each exposure media allows the assessment of the variability of the contributions for each exposure media. The LADD concerns the intake due to ingestion and does not include inhalation. The Figure 4.14 shows the LADD for each ingestion exposure pathways for the two age classes: 2-17 and 17-70 years old in the totality of the French area for benzo[a]pyrene.

At the national scale, the variance between the modeled values of the LADD, for the same pollutant and exposure pathway is on one hand the result of the differences in the contaminant concentration spatial distribution in the given exposure media and on the other hand, the result of the difference diet scenarios (production rate of locally grown food, type of food, quantity ingested).

For ingestion, the predominant pathway for Benzo[a]pyrene for both classes is commercial aliments consumption (excluded vegetables) (Figure 4.14). The concentration of PAH in commercial food is considered the same in the entire France and stable for the examined

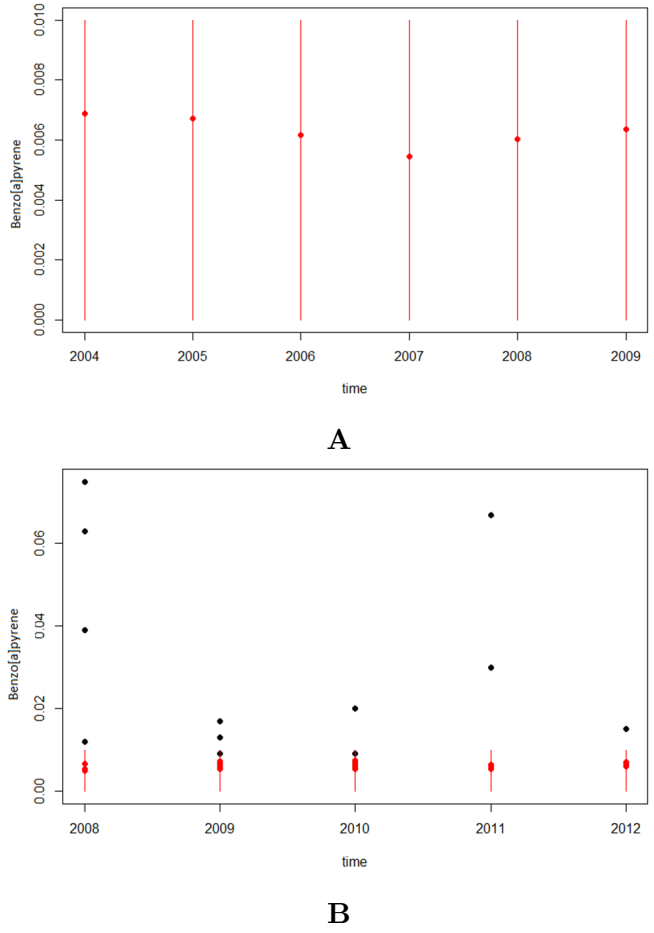


Figure 4.11: Example of imputation for two water distribution facilities (UDIS), with red points the imputed values ($\mu g/l$) and red lines the prediction domain, with black the observed ones: (A) For this facility there are no observed values, and therefore no temporal effect to take into account, the imputations are solely made based on the correlation between Benzo[a]pyrene and the other substances of the same family and it is the same for all the imputed datasets; (B) In this case the plurality of observed values allows the combined approach for the imputation, resulting in bigger variance between the imputed datasets.

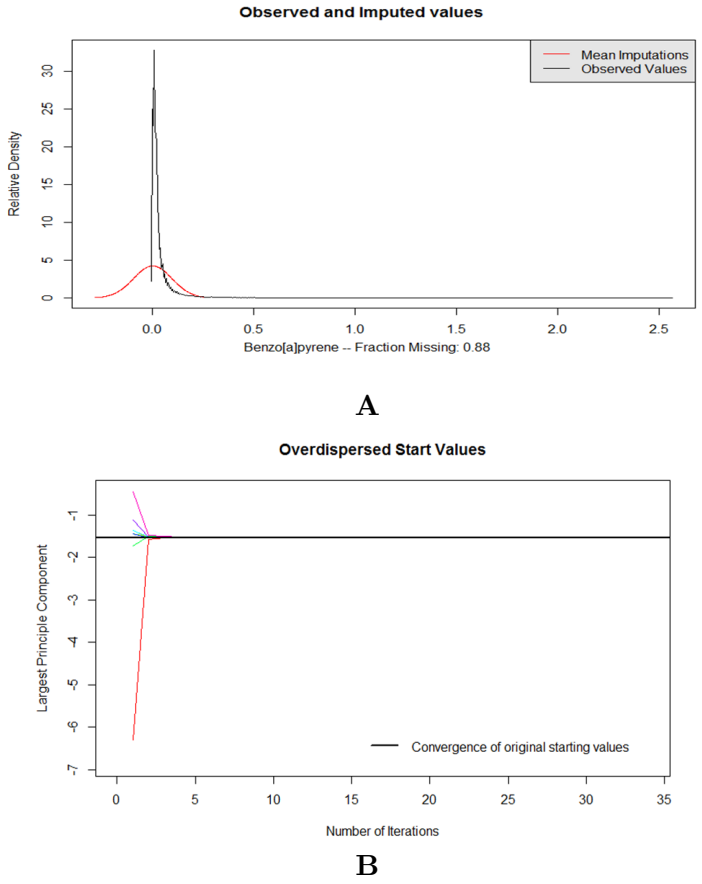


Figure 4.12: Diagnostic plots for the EMB imputations: (A) Comparative plot of densities between the observed and the mean of the imputed datasets retained; (B) Overdispersion diagnostic plot that indicates that the chains from the Expectation - Maximization algorithm have converged to the same parameter estimates which means that the global maximum is achieved rather than a local one.

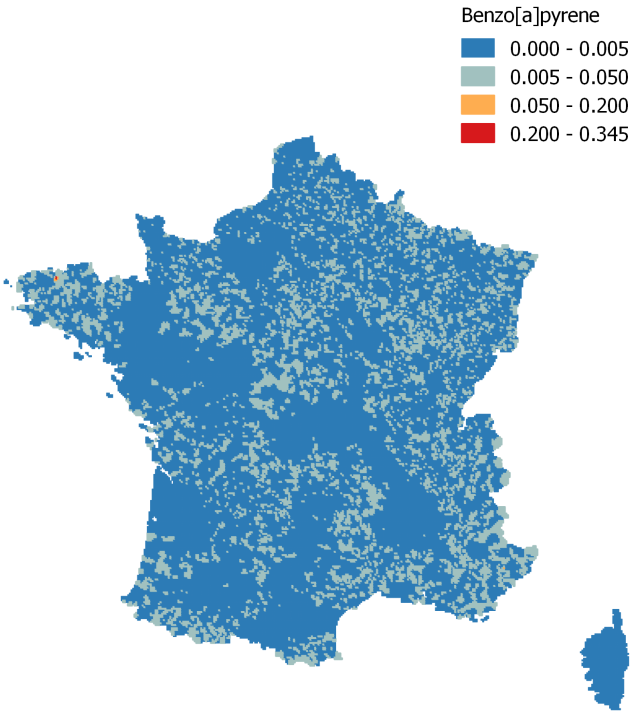


Figure 4.13: Spatialization of benzo[a]pyrene in water ($\mu g/l$) for France in the 3×3 km grid.

period. The second greatest contribution comes from the drinking water, followed by the consumption of vegetables. Finally, the ingestion of soil has the smallest impact in the daily exposure dose.

4.2.2 National and local vegetation contributions to ingestion

In this section, the plant contributions on the global vegetation ingestion exposure is presented, based on their origin (local or commercial). Root-vegetables and fruits account for the biggest contribution of commercial vegetation for both age classes with 48.6% and 42.3% for class 1 (children) as well as 45.8 and 39.2% for class 2 (adults) (Figure 4.16). Leafy vegetables and vegetable-fruits are the main contributors towards the exposure due to local vegetation consumption with 33.8% and 31.7% for class 1 while 32.8 and 32.7 for class 2 (Figure 4.16). For illustrating purposes, the national and local contributions for two different study zones are also presented (Figure 4.17).

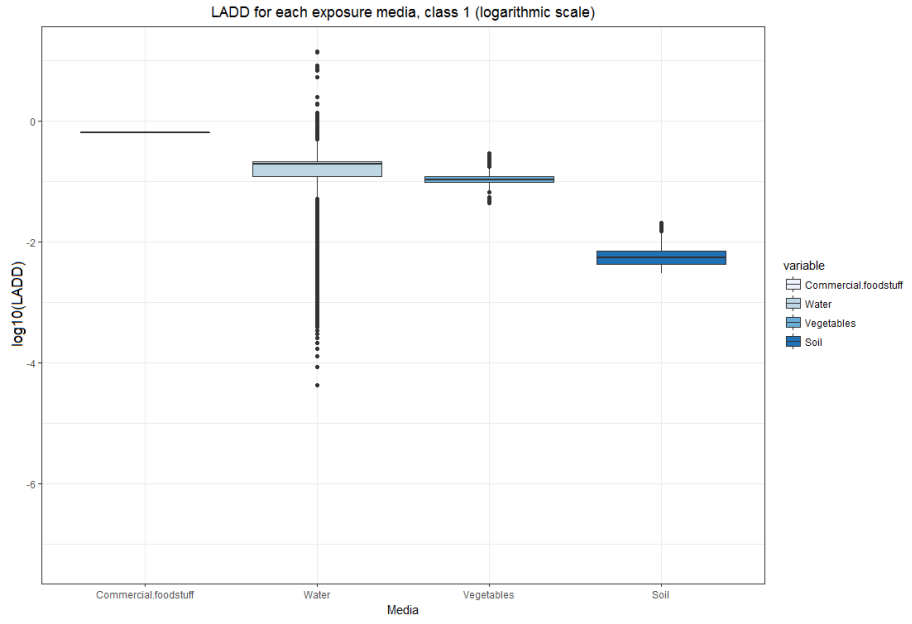
4.2.3 Comparison of estimated and measured values

The measurements considered for simulation of plants and animal products are based on the Total Feeding Study (Etude d'Alimentation Totale, EAT 2), which provides data on average concentrations of contaminants in food of commercial origin, consumed by the population (INRA-DGAL-AFSSA, 2011) at a national level in France. The modeled values are obtained by deterministic simulation of contaminant transfers in local plant matrices taking into account root transfers, deposits and resuspension of soil particles.

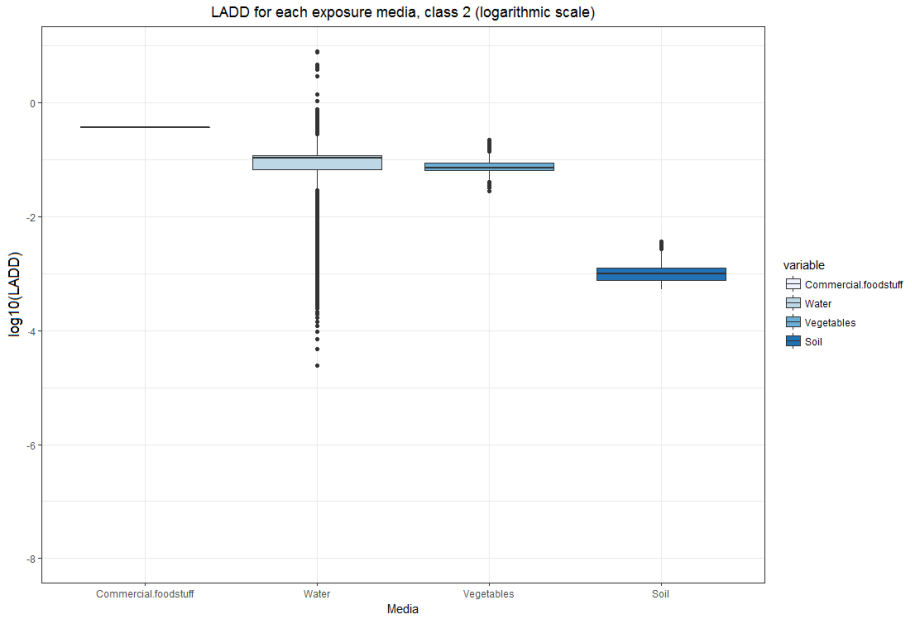
Overall, for the three categories (fruit, leafy vegetables and vegetable fruit) the modeled values seem to be overestimated for all three substances when compared to the measured ones, while for the root vegetables, the concentrations modeled are underestimated.

The biggest difference between the modeled concentrations obtained and the measured concentrations is observed for leafy vegetable (128 times greater for B[a]P, 92 for Bgh and 82 for Ind) and fruit (61 times greater for B[a]P, 45 for Bgh and 39 times for Ind), followed by vegetable fruit (25 times greater for B[a]P, 18 for Bgh and 1.6 for Ind). On the other hand, root vegetable concentrations are underestimated by a a factor of 9 for B[a]P, 39 for Bgh and 26 for Ind.

It is important to define which of the above differences can be explained by comparing soil concentrations in the study area (relative to data modeled in locally sourced foods) compared to those the commercial origin.



A



B

Figure 4.14: LADD by exposure media in base 10 logarithmic scale, for the totality of the study area: (A) Age class 1: 2-17 years; (B) Age class 2: 17-70 years.

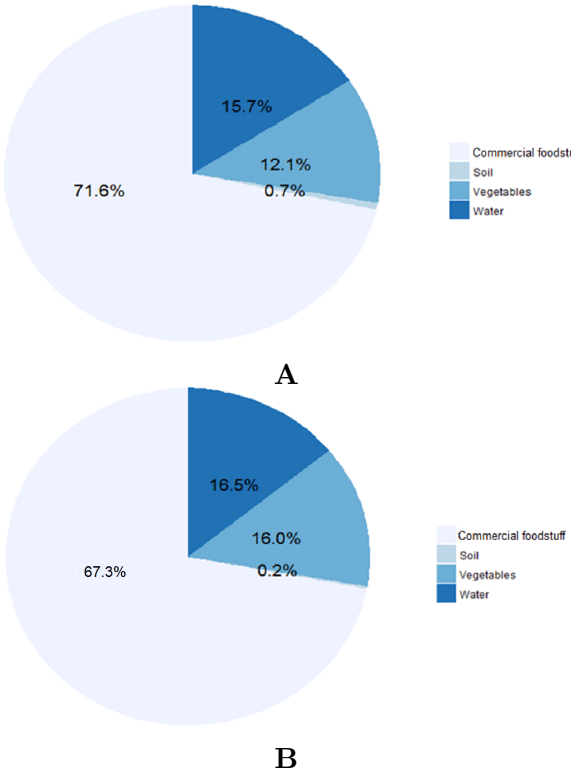


Figure 4.15: Contributions of exposure routes to the LADD for the totality of the study area: (A) Age class 1: 2-17 years; (B) Age class 2: 17-70 years.

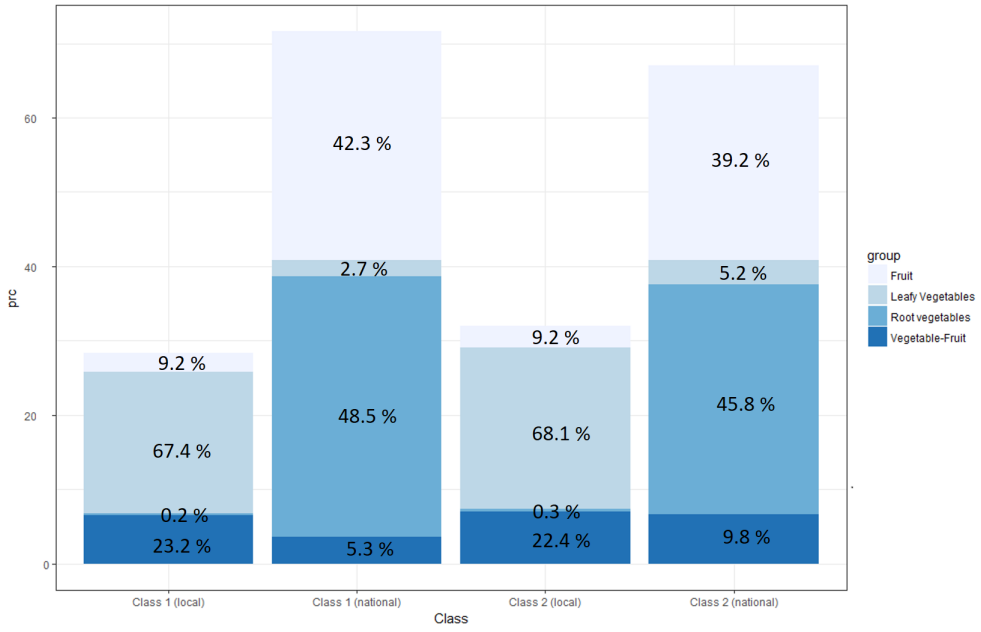


Figure 4.16: Contributions of vegetable subgroups according to their origin (local and national), for both age groups (class 1: 2-17 and class 2: 17-70 year old) in percentage of the global vegetation ingestion exposure.

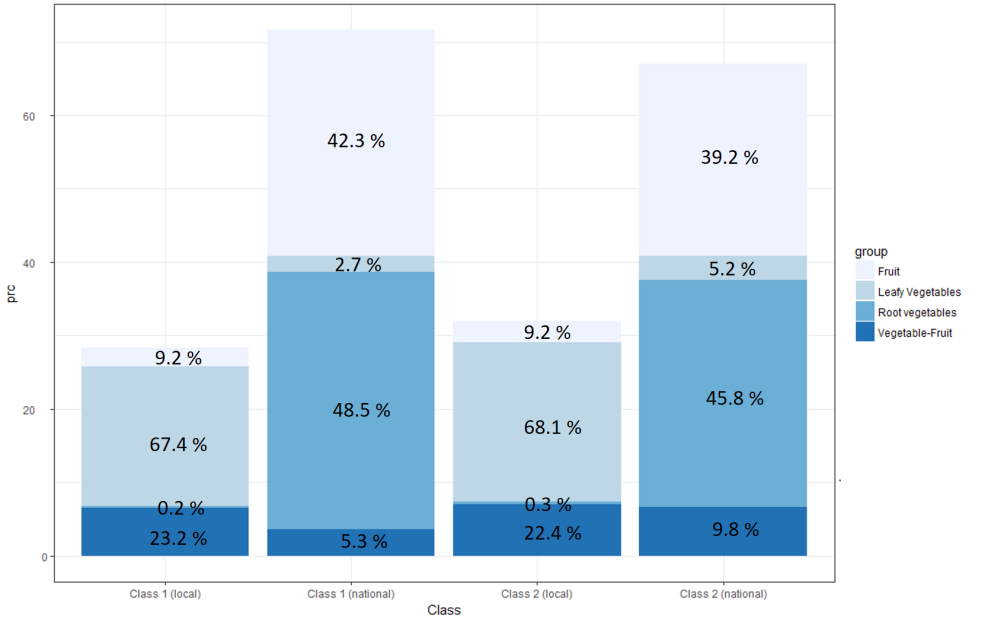


Figure 4.17: Comparison of local and national contributions to the exposure for two different study zones for the whole lifetime considered (2-70 years old) in percentage of the global vegetation ingestion exposure.

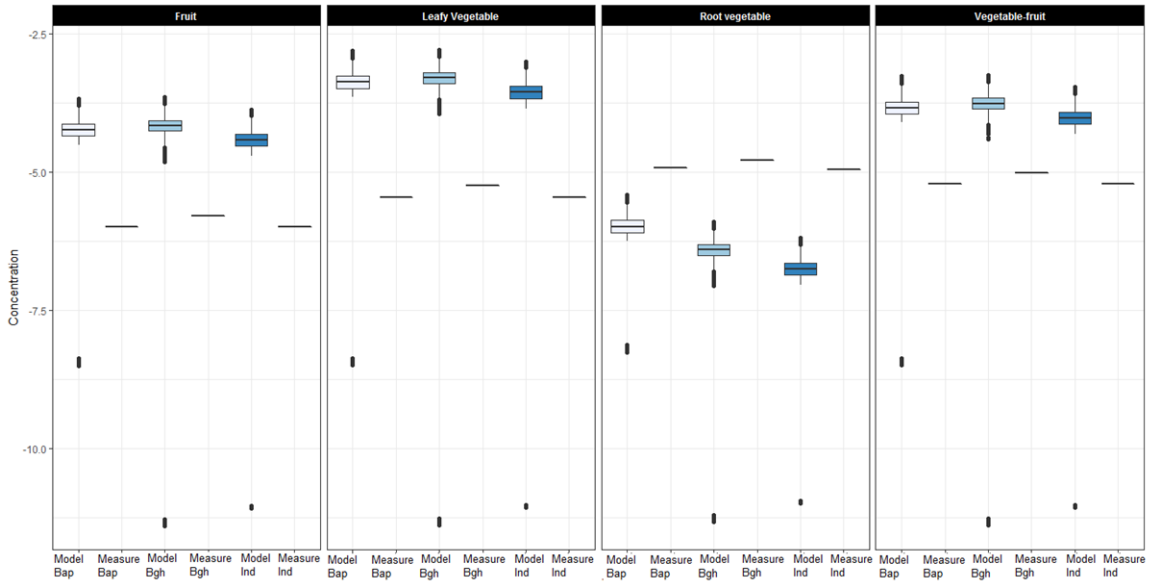


Figure 4.18: Boxplots describing the distributions of the modeled concentrations (mg/kg) of benzo[a]pyrene in the four categories of fruit and vegetables versus the measured concentrations in the EAT 2 study (ANSES, 2011) (in base 10 logarithmic scale).

B[a]P	Quantile 25	Median	Quantile 75	Mean	Measured value
Fruit	$4.5 \cdot 10^{-5}$	$5.7 \cdot 10^{-5}$	$7.4 \cdot 10^{-5}$	$6.1 \cdot 10^{-5}$	$1.0 \cdot 10^{-6}$
Leafy vegetable	$3.2 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$5.4 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$3.5 \cdot 10^{-6}$
Root vegetable	$8.1 \cdot 10^{-7}$	$1.0 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$	$1.2 \cdot 10^{-5}$
Vegetable fruit	$1.1 \cdot 10^{-4}$	$1.4 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	$6.0 \cdot 10^{-6}$

Bgh	Quantile 25	Median	Quantile 75	Mean	Measured value
Fruit	$5.4 \cdot 10^{-5}$	$6.8 \cdot 10^{-5}$	$8.5 \cdot 10^{-5}$	$7.2 \cdot 10^{-5}$	$1.6 \cdot 10^{-6}$
Leafy vegetable	$3.9 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$6.2 \cdot 10^{-4}$	$5.2 \cdot 10^{-4}$	$5.6 \cdot 10^{-6}$
Root vegetable	$3.1 \cdot 10^{-7}$	$3.8 \cdot 10^{-7}$	$4.8 \cdot 10^{-7}$	$4.1 \cdot 10^{-7}$	$1.6 \cdot 10^{-5}$
Vegetable fruit	$1.4 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$2.2 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$	$9.6 \cdot 10^{-6}$

Ind	Quantile 25	Median	Quantile 75	Mean	Measured value
Fruit	$2.9 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$4.8 \cdot 10^{-5}$	$3.9 \cdot 10^{-5}$	$1.0 \cdot 10^{-6}$
Leafy vegetable	$2.1 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$3.5 \cdot 10^{-4}$	$2.9 \cdot 10^{-4}$	$3.5 \cdot 10^{-6}$
Root vegetable	$1.3 \cdot 10^{-7}$	$1..10^{-7}$	$2.3 \cdot 10^{-7}$	$1.8 \cdot 10^{-7}$	$1.1 \cdot 10^{-5}$
Vegetable fruit	$7.3 \cdot 10^{-5}$	$9.5 \cdot 10^{-5}$	$1.2 \cdot 10^{-5}$	$1.0 \cdot 10^{-5}$	$6.0 \cdot 10^{-6}$

Table 4.4: Comparison of modeled values versus the measured ones (mg/kg wet) for the three PAH substances, for every vegetable subgroup studied.

4.2.4 Transfers' contributions

It is possible to assess the main contributions of the various transfers towards the total concentrations of contaminants measured in the vegetables. The main contributions in the three out of the four vegetable categories (excluding the root vegetables) are almost exclusively due to deposition transfer: 99.8% for leafy vegetables, 99.5% for fruit vegetables and 98.8% for fruit for benzo[a]pyrene, equivalent for the other two substances. Root transfers account only for a small percentage towards the total concentration: 0.15% for leafy vegetables, 0.45% for fruit vegetables and 1.1% for fruit, for benzo[a]pyrene. For root vegetables, as expected the totality of the concentration is due to the transfer from the roots.

4.2.5 Risk exposure pathway and environmental compartment contributions

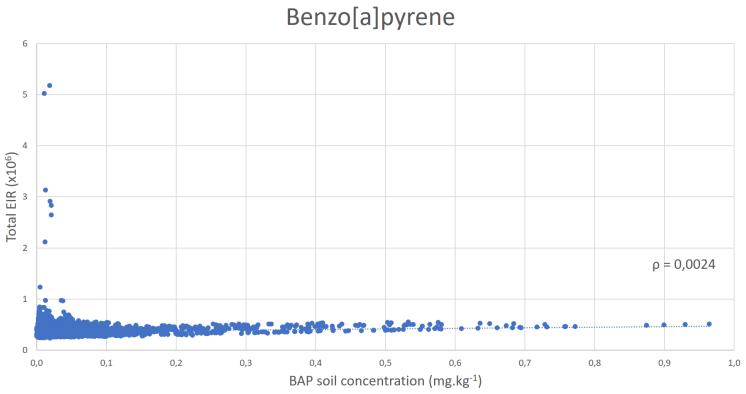
By aggregating the inhalation and ingestion exposure pathways in an integrated indicator using toxicological reference value, the Excess of Individual Risk (EIR) permits to analyze the contribution of exposure pathway and environmental compartment in term of carcinogenic risk.

In Figure 4.19 the EIR for benzo[a]pyrene, obtained for France as a function of the concentrations in the environmental compartments (for the environmental compartments per substance) is shown.

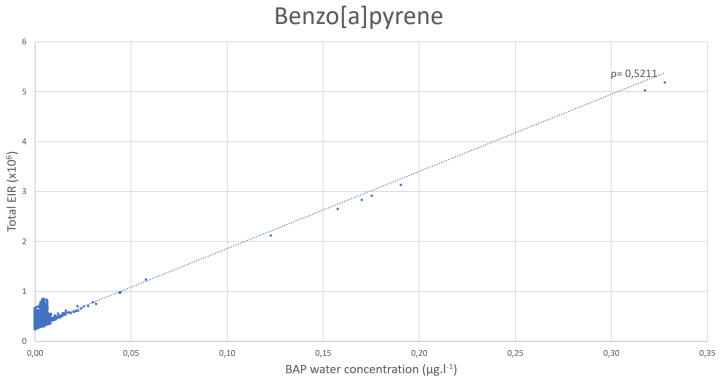
The strongest correlation ($\rho = 0.53$) is observed between benzo[a]pyrene concentrations in the water compartment and the EIR (Figure 4.19), followed by the air compartment with similar correlation ($\rho = 0.5$). Weaker concentrations are observed between the EIR and soil compartment ($\rho = 0.002$). The contribution of the environmental compartments to the other two substances follow the same pattern with benzo[a]pyrene, as the three substances are highly correlated.

4.2.6 Cartography of environmental inequalities

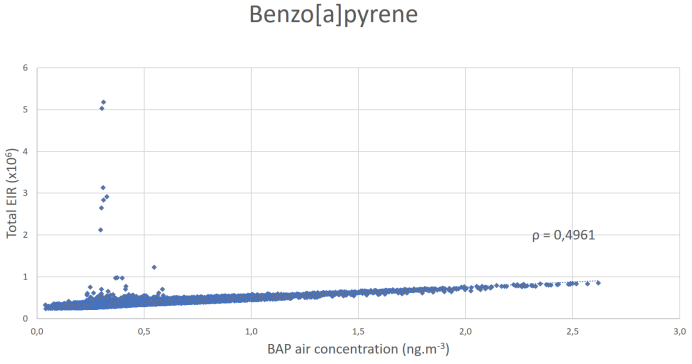
GIS could be used to map environmental inequalities of exposure by spatializing the EIR on a fine resolution (9 km^2 grid). The spatial structure of risk maps reflects the influence of a complex set of spatial and environmental factors with great variation and operating in different spatial scales. The analysis of these structures allows to quantify the



A



B



C

Figure 4.19: Correlations between EIRs and benzo[a]pyrene concentrations in environmental compartments obtained in France. (A) EIR vs. soil concentrations; (B) EIR vs water concentrations; (C) EIR vs atmospheric concentrations.

combination and the contribution of different scales of spatial variability.

The results obtained allow to identify the areas that high exposures are more likely to occur. The similarity between the map of the exposure by ingestion (Figure 4.20) and the spatialized concentration of B[a]P in water (Figure 4.13) shows that the exposure due to ingestion corresponds mainly to the drinking water intake. However, by examining the map of the total exposure, the inhalation route is the dominant contribution and the spatial structure of the total spatialized risk indicator follows the same pattern as the spatialized atmospheric concentration of B[a]P in France.

The two principal components for BAP appear in Figure 4.19. The first corresponds to the drinking water concentration, with $\rho = 0.53$ (in the very high EIR range values). The second component is due to the air concentration.

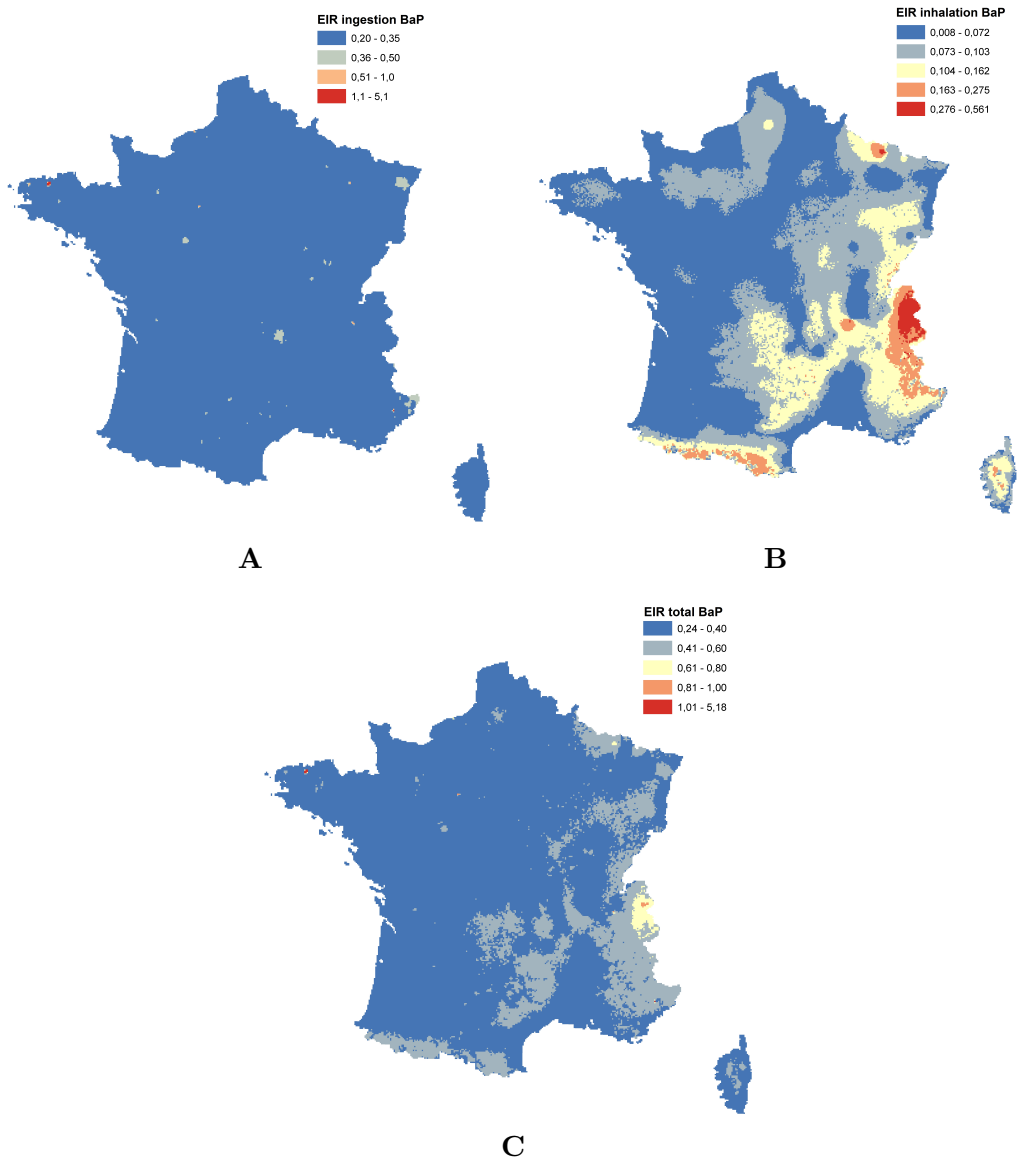


Figure 4.20: Mapping of benzo[a]pyrene EIRs in the study area for the lifetime of exposure: (A) EIR due to ingestion; (B) EIR due to inhalation and (C) EIR total.

Chapter 5

Discussion

In this chapter, the final results of the study are discussed. The principal developments and processes applied on the available data in the context of environmental inequalities characterization are discussed. The limitations and results concerning spatial and temporal support as well as data homogeneity and resolution are presented. The coupling approaches put in place to combine the different data are also detailed: spatial data, environmental compartments, exposure pathways. Finally, exposure assessment results are analyzed and compared to other studies and uncertainties are discussed in order to provide efficient exposure map to prioritize public action.

5.1 Evaluation of exposure in the context of environmental inequalities

PAHs released to the atmosphere are subject to short- and long-range transport and are removed from the atmosphere by wet and dry deposition onto soil, water, and vegetation. In surface water, PAHs can be volatilized, photolyzed, biodegraded and bind to suspended particles or sediments, or accumulate in aquatic organisms. Since PAHs are present in various environmental compartments, a multipathway exposure assessment was put in place. To evaluate environmental PAH exposure in the environmental inequalities context , it was necessary to collect data to characterize environmental compartment contamination. The data selection focused on measured data, which allows the integration of past and present, local and diffuse sources. This prerequisite required to couple multimedia modeling approaches with spatial analysis techniques, interfaced in a GIS.

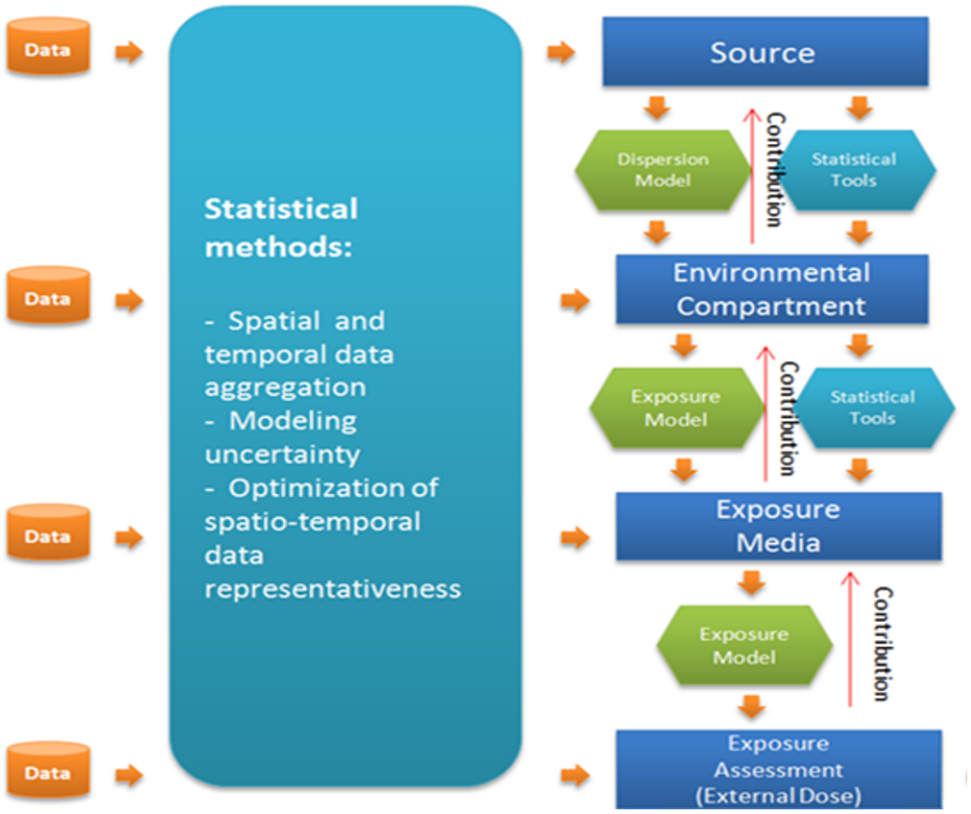


Figure 5.1: Diagram of the thesis procedure.

A source-to-exposure continuum approach was used for connecting the source of the exposure with the target exposure. The exposure assessment integrates a data processing workflow able to connect the global source-effect chain. This approaches is presented in Figure 5.1. This approach addresses the need of a methodological tool particularly adapted towards the construction of an integrated platform from the source to the health risk. The modeling is based on the use of measurement data in environmental media, from environmental databases, rather than the classical approach that employs a fate and transport model using source data as principal input. Despite the greater need for georeferenced data that this method requires in comparison to the other approaches, it also allows to limit the coupling of environmental compartments and the horizontal substance transfer assessment needs.

The definition of pertinent exposure scenarios is crucial in order to identify the population at risk. The different scenarios include various age classes as well as the different

dietary habits. Thus, to assess the exposures due to ingestion, it was necessary to integrate the exposures related to the consumption of commercial and homegrown products. Self-consumption factors for each food category were spatialized over the entire study area. This approach allows the determination of the exposures related to the ingestion according to the global or local character of the origin of products, as well as during the cooking process. It should be noted that the data used for the self-consumption factors are collected from older sources (1991), that could be potentially leads to overestimate the contributions of the local exposure towards the total. 0 to 2 years old age class were not included in the analysis since consumption data are not overestimated using French Ciblex database.

PAH exposure has been estimated by other multimedia models such as MERLIN-Expo [191] and INTEGRA [175] answering different problematics. INTEGRA was initially developed to address consumer exposure by establishing detailed consumer use scenarios, and hereby accounting for dermal, oral and inhalation pathways. MERLIN-Expo on the other hand, focuses on the evaluation of environmental exposures including indirect exposures due to aquatic and terrestrial food web. The modeling approach employed in our work, considers only the inhalation and ingestion pathways due to commercial foodstuff as well as home-grown vegetation. An extension of our approach could include the integration of other exposure pathways (like dermal) or a more detailed description of exposure due to biota.

In general, addressing environmental inequalities through exposure assessment poses certain limitations (such as the scope of assessment, the level of detail, etc.) that influence the choice of data and modeling. This demands the implementation of methodological developments to process additional data. That way, interactions between the different determinants and geographical levels can be evaluated, by acquiring longitudinal environmental, health, and socio-demographic data.

Because no sufficient model parameter data were available to characterize sensitive phenomena (soil-plant transfer factors for example), a deterministic model was used. however stochastic models should be more efficient to deal with the variability of exposure that can occur in a given population. To be more realistic and representative, individuals and events are segmented in similar groups (in a given area discretize on meshes) for which the statistically parametrized or actual measurements are used to describe them in the input. The output is an average of exposure per defined group.

The predicted values can to be verified by either comparing them with actual high-end observations or biomonitoring data. However no biomonitoring data were available easily

to compare predicted values with biological impregnation.

Conceptualization involved making fundamental choices about what aspects of reality were relevant to the purpose of the specific modeling process and which aspects of reality were left out. For example, except for deposition from the CHIMERE model output, the approach fails to take in account long-range transport and chemical transformations.

5.2 Data processing workflow

A large number of data is required in order to characterize the external dose starting from environmental data, including: emissions, concentrations in the environmental compartments (water, air, soil), exposure (food, drinking water), population data (behavioral, physiological, demographic). To construct the spatial dataset to serve as an input for the model, heterogeneous data from different sources were employed, which were often initially collected for a different purpose than that of the study. To ensure their relevance to the purpose, additional processing was required to increase spatial and temporal representativeness of the population and contamination of the studied area.

Environmental inequalities characterization is based on specific assumptions and hypotheses, which influences and limits the data selection. However, data that do not initially correspond to these objectives can be used after appropriate processing or modeling, allowing to partially avoid the biases that would result from their direct use. Data that are not available at the resolution of interest are aggregated or desegregated in the desired resolution. Moreover, additional variables correlated with the pollution processes are tested and constructed from available database to increase resolutions. Finally, additional processing is needed to homogenize all available data on the same spatio-temporal unit of the study.

5.2.1 Soil compartment

Concentration measurements from sample locations do not provide a robust base to directly spatialize the PAH concentrations in soil due to the grid resolution and the values under the detection limit. Therefore additional information is used to address those limitations and improve the accuracy of concentration estimation. In this regard, 16 auxiliary variables were introduced: Two of them being artificially constructed and 14 physico-chemical soil properties.

The first artificial auxiliary variable is constructed to address the censored values problem. It corresponds to transform the measured concentration in term of probability of concentrations to be under a threshold using indicator kriging method. This way, the sampled locations where the measurements were not exhaustively known, can be included in the spatialization process, instead of being disregarded. In the map in Figure 4.6, the northern part of France (Hauts-de-France region) concentrations tend to be higher than the concentrations observed in the central and south west of France.

Recent studies [233; 112; 183] showed that the anthropogenic (coal combustion, traffic related) emissions are positively correlated with the distribution of PAHs in soil as well. In the extreme north of France, polluted areas are attributable to the metallurgic industry.

The second auxiliary variable is constructed in order to include the available information from the polluted sources localization. The processed variable represents the polluted sites proximity. Spatial analysis techniques were used to construct metric in each study meshes according to their proximity to potential exposure sites: being far away from a polluted site might have null impact on measured concentrations, therefore it is necessary to define a buffer (i.e. a distance threshold) within which the impact is strong enough to be considered. The buffer is a good compromise to summarize the effect of the polluted sites, within which several distance decay models are considered to model the pollution decay. Nevertheless, proximity to a potentially contaminated site does not involve systematic population exposure. Finally, the 14 remaining auxiliary variables correspond to physico-chemical properties in soil that have either shown or could potentially affect the spatial distribution of PAH.

Regression followed by residual kriging provides the framework to include information from the auxiliary variables to the prediction, increasing the precision of the estimation. The 16 variables are combined into a geostatistical model to spatially predict the PAH distribution.

Two methods were employed to fit the regression: The traditionally used linear regression and the random forest regression. To assess the quality of the model and the prediction we use the coefficient of determination (R^2) and a cross validation. It is to note that the performance of the model depends on the sampling. The evaluation in Table 4.1 shows that random forest outperforms linear regression every time. A reason for the poor performance of the linear regression could be the nonlinear nature of relationships between the variables. The random-forest regression selected to t the residuals before the spatial interpolation allows to include the various variables in the model (soil properties, distance from polluted

sites, indicator kriging result) in a more flexible framework, than linear regression which is usually used. The fact that the random forest outperforms notoriously the linear model indicates that the relation of the explanatory variables with the PAH concentrations is highly nonlinear.

Moreover, random forests provide a setting to evaluate the importance of the variables inserted in the model. In this case the variables retained by order of importance are: land coverage of forest area and semi natural environment, surface roughness, slope aspect, mean evapotranspiration, ecoclimate, the indicator and the distance from polluted sources variables. As a remark, the importance does not necessary imply a positive correlation. Including only these 8 variables instead of the 16 improved the performance of the model prediction indicators (according to R^2 and cross validation score): including non-pertinent variables can add noise to model and decrease prediction quality.

The fact that forest coverage and slope as predictors of B[a]P concentrations are retained as the more important variables is consistent with the biochemistry of PAH. This dictates that gaseous and particulate phase PAH are affected by dry deposition and precipitation when they travel to colder environment and higher altitudes. During this process, PAH could be trapped and deposited to soil surface. Soils enriched with organic matter (as it could happen in the forest) act as an environmental sink of persistent organic pollutant. Previous studies have indeed concluded the altitudinal and forest based impact in the PAH distribution [192].

After spatialization of topsoil B[a]P concentration, higher concentrations are observed in the north of France as well as some hotspots around certain polluted sites. Similar patterns have been observed for other PAH substances in [208] and correspond to highest levels in the north and northeastern part of France while smaller concentrations in the south west and south east.

The kriging variance is retained as a measure of uncertainty in this model as well. Additional uncertainty analysis in proximity of polluted sites and not well characterize by measure should provide interesting information about population exposure and to improve statistical model quality.

5.2.2 Air compartment

Data for toxic air pollutant exposures like PAHs are usually scarcer in geographic and temporal coverage than data for other regulated pollutants such as nitrogen dioxide (NO_2),

ozone (O_3) and particulate matter (PM). The available dataset consisted of a limited number of observations, measured in air quality monitoring stations (76 for 2010; 77 for 2011). In order to adequately spatialize them in the French region, a geostatistical method allowing to take into account auxiliary variables was employed. Two different models were constructed for each year and then combined in a single map.

The distance-decay principle was used to construct the auxiliary variable from the emissions inventory. This principle is based on the assumption that proximity to the source is a main determinant of emission intensity. This along with a buffer zone allows us to estimate the emission contribution from each source, and it can be then used as an auxiliary variable in the model. It does not exist a recognized model when selecting the distance-decay function and the buffer zone radius, therefore multiple models were tested. For 2010, the optimal model (i.e. the one better correlated with the initial measurements) showed a steeper decline on the effect of the emissions (equation), while for 2011 a more gradual decline was privileged.

PAHs have been shown to concentrate in colder environments at higher latitudes and altitudes [192], therefore altitude was considered as a potentially pertinent variable to increase the modeling quality. Indeed for 2010 was included in the model, increasing the accuracy. The population was inserted in the model to limit the effect of the altitude for the mountainous areas, which are sparsely inhabited. For 2011 no strong correlation was observed. This difference between the two years could be explained by the differences in climatological conditions. For example, a particularly cold winter could increase the need for domestic combustions in the areas of high altitude, and consequently observe increased emissions of PAH. Since the available data correspond to annual mean observations, it is not possible to account for this seasonal effect. For a more refined temporal approach, a seasonal modeling taking into account the climatological data could be established.

Combining the models of two consecutive years allows to limit the climatological effect of a particular year. The method employed to combine the results in each cell, consisted in a weighted average of the two years, where the kriging output with the smallest variance is privileged.

The associated uncertainties and models' limits are difficult to estimate where there are no or limited available data. In order to avoid the improper characterization of exposure, we define the model limit: areas with altitude above 1800 m and population below 200/3 km^2 are not-relevant in the estimation since no monitoring station permits to assess the linearity of the model in this condition range. That way, the results can be presented by

omitting the areas where the uncertainty due to the altitude is too high to adequately assess concentrations (Figure 5.4.4). Nonetheless, by combining the results of two consecutive years, we minimize the effect of extreme events that may appear in a year but not in the following.

The quality of those air quality concentration maps relies primarily on the quality of the input data used, namely the spatial density of the observations and the relevance of auxiliary information introduced in the interpolation method.

A more refined model could better address the underlying differences in the exposure. To do that, the emissions used as auxiliary variable could be separated into industrial and non-industrial, as their dispersion could be potentially different. In this context, different distance-decay functions could be fitted for the two cases. The climatic conditions subjected to seasonality or to other specificities (such as the micro-climate along coastlines) have been shown to affect exposure to PAHs [234; 143]. To better address this, longitudinal models capable to capture the seasonality can be constructed, including climatic data as auxiliary variables (such as temperature, wind speed, etc). In the final result, the concentration of B[a]P are higher in the mountainous areas (particularly in the south west of France and in the south east of France). However, given the limited number of observations initially available, it is difficult to identify the exact limitations of the model, where the altitude is high and the density population is low. Additional sampling in those areas as well as in the highly industrialized areas where contrasting situations are observed, could provide a robust base to better quantify the associated uncertainty.

The results obtained in this study are in accordance with those obtained in the context of regional cartography of Benzo[a]pyrene in Rhône-Alpes (Cartographie régionale du Benzo(a)Pyrène, Atmo 2012). Specifically, the average of four measurement campaigns carried out in the Rhône-Alpes region between 2010 and 2011 indicates similar high values on those areas.

Finally, given the high correlation between the observed measurements of B[a]P, Bgh and Ind, the estimated concentrations of B[a]P is used as the external drift to guide spatial interpolation for the two other substances.

5.2.3 Water compartment

The available concentration measurements of PAH substances for the drinking water were taken in the water distribution network in France. Due to its hydrophobic nature, B[a]P

is found in water in small concentrations, therefore the exact measurement cannot always be reported, requiring additional data processing before spatialization. Reconstructing these concentrations is not trivial, especially when the rate of observations under the detection limit is high. An imputation method was selected to process these values under the detection limit in an effort to avoid the distortion when estimating the contribution of the drinking water as an exposure pathway towards the total exposure. The imputation was performed taking the estimated relationship between B[a]P and other PAH substances as well as the temporal correlation into account. This allows to capture a more realistic spatial distribution compared to a simple substitution method that would result in a smooth, uniform result. As a result, a rened spatial distribution of B[a]P may also reect the pollution origins, and the variance between the imputed datasets provides a quantication of the associated uncertainty.

Honaker et al. [85] suggest that a total of five imputed sets is usually sufficient, however given the high missingness (above 80% of data under the detection limit in our case) , specifically for B[a]P, eight imputed sets were obtained instead. The number of iterations needed to reach convergence for the substances with missingness up to 80% is rather similar (120-150) for each set, indicating the stability of the algorithm. On the contrary, the first tries with the datasets showing high missingness (above 80%) displayed high variability in the number of required iterations (220-600). For those datasets, the imputation was re-run including a ridge prior to 1% as recommended by Honaker et al. [85]. This significantly improved the evenness in the number of iterations (between 160-200). Therefore, these imputed sets were used in further analysis.

Because of the nature of the missing data mechanism, it is impossible to tell whether the prediction of the imputation model is close to the unobserved value that is trying to be recovered. A way to assess how accurate the imputed data consists in sequentially treating the observed values as if they were missing and generating several hundreds of imputed values (allowing the construction of confidence intervals). The observed and imputed data can then be compared graphically. In particular it is checked that the confidence interval includes the actual value, which is the case when it crosses the x=y line (perfect imputation), indicating that the imputation model is doing well. However, in the case of missingness due to values under the detection limit, the accuracy of the imputation cannot be evaluated with this approach since the observed values are above the detection limit while the missing ones are below this limit.

To evaluate the performance of the multiple imputation, the method was applied

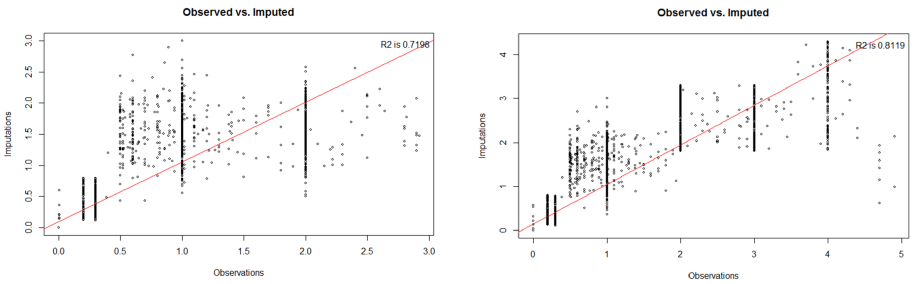


Figure 5.2: Scatter plot of Observed vs. Imputed values for (A) DL=3 $\mu g/l$ and (B) DL=5 $\mu g/l$.

using a second dataset, that was artificially imputed. The database selected, contains concentrations of heavy metals in the drinking water in France, that were measured in the same conditions as the PAH (same temporal sampling in the water distribution units). The goal is to determine how close the imputed values are to the actual observed values, by considering part of the observed measurements as under the detection limit. The variable of interest in this case is Arsenic (As), and the auxiliary variable is Selenium which were found to have the strongest correlation with (0.467).

The complete dataset, consists of 4718 observations. Two different limits are selected and tested. At first, the observed values of Arsenic and Selenium that are less than 3 $\mu g/l$ are set as under the detection limit, resulting in 82.1% of the complete data set as under the detection limit. In the second try, the limit was raised to 5 $\mu g/l$, resulting in 88.3% of measurements under the detection limit. The AMELIA II was launched, receiving 8 imputed datasets for each case. To compare the imputed values to the observed ones, the means of the imputed sets were taken. In Figure 5.2 the scatter plot of the observed values versus the mean of the imputations sets is presented.

The variation that is observed can be partially explained due to auxiliary variable selected and their weak correlation. However, comparing with the methods of replacing the values under the detection limit either by $\frac{\mathrm{DL}}{2}$ or $\frac{\mathrm{DL}}{\sqrt{2}}$, the imputation method has the smallest absolute mean error (Table 5.2).

As it is expected, by raising the number of under the detection limit values in the sample, the average mean error rises (Table 5.1), but still, the imputation method compared to the constant substitution methods has a clear advantage.

For the districts were no measurements were available, the concentrations were considered to be 0. However, this could have an impact in the distortion of exposure route, by

Method	AME (3 $\mu g/l$)	RMSE (3 $\mu g/l$)	AME (5 $\mu g/l$)	RMSE (5 $\mu g/l$)
Imputation	0.17	0.37	0.21	0.45
$\frac{DL}{2}$	1.18	1.29	1.95	2.12
$\frac{DL}{\sqrt{2}}$	1.37	1.49	1.73	0.068

Table 5.1: Absolute mean error and root-squared-mean error for the three methods, when limit of detection is 3 $\mu g/l$ and 5 $\mu g/l$.

underestimating the contribution of the drinking water. To avoid this potential distortion, these measurements can be substituted by a global mean.

The uncertainty associated with the imputation process is considered in the form of variance between the imputed datasets. This development finalizes to provide uncertainties for each environmental spatial data process.

5.3 Modeling results

5.3.1 Coupling administration and exposure routes

The daily exposure dose and the EIR allow the combination of all the exposure media. For ingestion, the contributions of each exposure media are summed, allowing this way the estimation of the total dose and the associated relative contribution. The total EIR is also calculated as the sum of the risks of ingestion and inhalation, also permitting the combination as well as the comparison of the contributions of the different administration routes of exposure. Nevertheless, the use of toxicological reference values to transform dose into risk adds a significant level of uncertainty. These values are usually defined in a conservative logic that could potentially generate distortions when comparing the administration routes.

For example, the Figure 5.3 present the EIR estimated for each exposure pathway using toxicological reference value proposed at different period (by INERIS in 2009 and by US EPA in 2017). For similar exposure media contamination, this figure shows that inhalation was more important in the total risk in comparison with the ingestion using 2009 toxicological reference value recommendation. This trend is inverted when using US EPA 2017 recommendation [54].

The indicator combines also biological responses which are associated with two different sensitive organs. This kind of indicator allows the identification of overexposed population

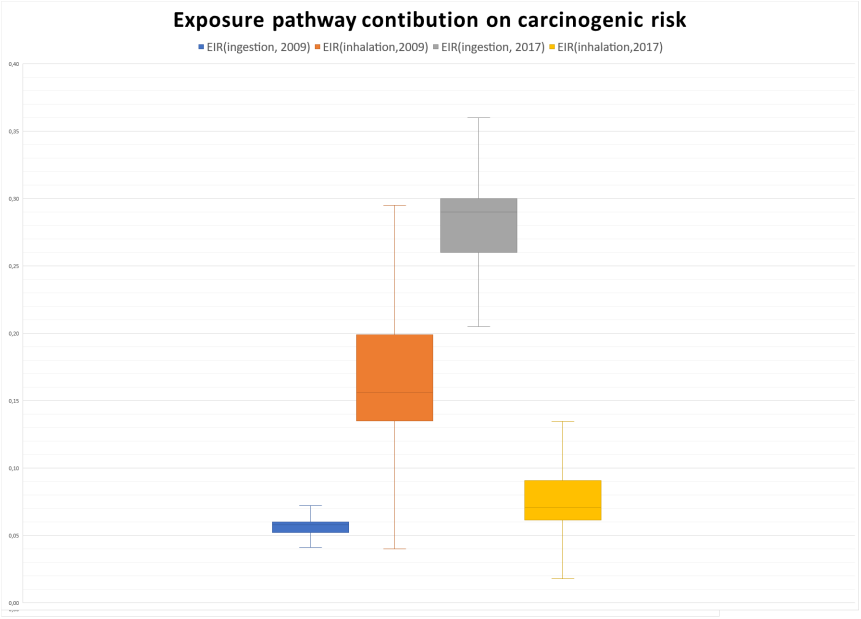


Figure 5.3: Estimation of EIR for each exposure pathway using toxicological reference value (2009 and 2017).

subgroups that are more submitted to potential health effects. It is common for studies that address exposure to multiple PAH mixtures.

Here, the single pollutant (B[a]P) risk analysis was favored. B[a]p is generally found in higher concentrations than other PAH substances and is commonly used as an indicator species for PAH contamination due to its higher toxicological reference value and high correlation with other PAH substances (Figure 5.4). It contributes to approximately 50% of the total carcinogenic potential of the PAH group. Although unmetabolized PAHs can have toxic effects, a major concern is the toxicity of their reactive metabolites [29; 4; 1]. A strong limitation of the employed model here is that it does not allow to capture the potential toxicity due to the formation of these reactive metabolites of PAH substances.

One other way to combine inhalation and ingestion pathways could use physiologically based pharmacokinetic (PBPK) models [111; 164] to assess the internal dose in different fluids, organs or tissues. With these models, mathematical modeling is used to predict the absorption, distribution, metabolism and excretion of chemical substances. Those prediction could be directly compared with exposure biomarkers. Those provide a direct measure of general exposure including all the possible exposure pathways, and they are extensively used in epidemiological studies to evaluate the dose-effect chain. However, evaluating the biomarkers is not always possible as it is not only a costly procedure but it

	BaP	BaA	BbFA	BkFA	BghiP	CHR	DBahA	IP	BjFA	CPP	DBaeP	DBahP	DBaiP	DBalP	MCH	BcFL	PAH2	PAH4	PAH8
BaP		0.92	0.89	0.89	0.94	0.80	0.88	0.96	0.90	0.41	0.74	0.04	0.39	0.65	0.08	0.24	0.86	0.90	0.92
BaA	0.92		0.92	0.91	0.87	0.91	0.81	0.91	0.90	0.38	0.65	0.02	0.35	0.51	0.05	0.22	0.94	0.97	0.97
BbFA	0.89	0.92		0.95	0.91	0.91	0.89	0.93	0.90	0.23	0.71	0.03	0.37	0.52	0.02	0.14	0.94	0.96	0.97
BkFA	0.89	0.91	0.95		0.87	0.86	0.84	0.91	0.88	0.23	0.68	0.05	0.37	0.58	0.02	0.12	0.89	0.92	0.94
BghiP	0.94	0.87	0.91	0.87		0.82	0.89	0.97	0.87	0.33	0.77	0.04	0.40	0.59	0.04	0.16	0.87	0.89	0.92
CHR	0.80	0.91	0.91	0.86	0.82		0.76	0.84	0.87	0.30	0.64	0.01	0.32	0.36	0.14	0.36	0.99	0.98	0.96
DBahA	0.88	0.81	0.89	0.84	0.89	0.76		0.92	0.81	0.18	0.81	0.15	0.49	0.60	-0.01	0.04	0.81	0.84	0.87
IP	0.96	0.91	0.93	0.91	0.97	0.84	0.92		0.90	0.28	0.80	0.08	0.45	0.63	0.02	0.12	0.89	0.92	0.94
BjFA	0.90	0.90	0.90	0.88	0.87	0.87	0.81	0.90		0.44	0.69	0.12	0.43	0.50	0.24	0.63	0.90	0.92	0.93
CPP	0.41	0.38	0.23	0.23	0.33	0.30	0.18	0.28	0.44		0.20	0.13	0.20	0.04	0.39	0.71	0.33	0.33	0.32
DBaeP	0.74	0.65	0.71	0.68	0.77	0.64	0.81	0.80	0.69	0.20		0.49	0.78	0.51	0.00	0.03	0.68	0.69	0.72
DBahP	0.04	0.02	0.03	0.05	0.04	0.01	0.15	0.08	0.12	0.13	0.49		0.86	0.04	0.01	0.01	0.01	0.02	0.03
DBaiP	0.39	0.35	0.37	0.37	0.40	0.32	0.49	0.45	0.43	0.20	0.78	0.86		0.30	0.02	0.02	0.34	0.35	0.37
DBalP	0.65	0.51	0.52	0.58	0.59	0.36	0.60	0.63	0.50	0.04	0.51	0.04	0.30		-0.01	0.04	0.43	0.47	0.51
MCH	0.08	0.05	0.02	0.02	0.04	0.14	-0.01	0.02	0.24	0.39	0.00	0.01	0.02	-0.01		0.67	0.13	0.10	0.08
BcFL	0.24	0.22	0.14	0.12	0.16	0.36	0.04	0.12	0.63	0.71	0.03	0.01	0.02	0.04	0.67		0.34	0.28	0.25
PAH2	0.86	0.94	0.94	0.89	0.87	0.99	0.81	0.89	0.90	0.33	0.68	0.01	0.34	0.43	0.13	0.34		0.99	0.99
PAH4	0.90	0.97	0.96	0.92	0.89	0.98	0.84	0.92	0.92	0.33	0.69	0.02	0.35	0.47	0.10	0.28	0.99		1.00
PAH8	0.92	0.97	0.97	0.94	0.92	0.96	0.87	0.94	0.93	0.32	0.72	0.03	0.37	0.51	0.08	0.25	0.99	1.00	

Figure 5.4: Correlation coefficient matrix calculated for 1375 upper bound food samples analyzed for all 16 priority PAH and combinations of PAHs (source: EFSA report).

also requires the collaboration of individuals. It also does not allow the evaluation of the exposure sources, which constitute the route of the problem. Then, PBPK model could be easily used at territorial scale by using spatial exposure multimedia model output as data input. For example, Sarrigianis et al. [176] found a mean B[a]P blood levels up to 0.00012 ng/L for adults and 0.0007 for children, while the respective urinary concentration of the major specific metabolite 3-OH-B[a]P is 0.01 ng/L.

5.3.2 Coupling environmental compartments and exposure media

The estimation of the concentrations in the exposure media from the environmental compartments requires the coupling of certain compartments, according to the choice of equations in order to simplify the calculation of the transfers between the matrices. In this modeling, only vertical transfers are taken into account, and the pollutant transport from a grid mesh to another is not considered. Atmospheric transport is integrated in the concentration and deposition data used. The soil and plant compartments are coupled with the atmospheric deposition. The other environmental and exposure compartments are considered independent and static over the period considered. Certain simplifications in the modeling approach could introduce bias in the estimations. The output data of each compartment is used as input for the adjacent compartment, accounting that way for the inter-compartment transfers. In the work of Sarigiannis and Karakitsios [176] a more complex modeling approach of PAH exposure is used, where the micro-environmental

concentrations are calculated taking into account the interactions among different media. This required to solve mass-balance differential equations and increase drastically time computing.

Using MODUL'ERS estimation of exposure and associated risk took 10 days. Mass-balance integration at a fine resolution and at the national scale requires to change calculation platform (as python or C++) and recode all equations. This option was discarded at the beginning of the thesis.

In the case of carcinogenic risks, the risks considered are those corresponding to the mean values, obtained during all life. Nevertheless, only the soil compartment was submitted to temporal change depending on competition terms as degradation (first-order kinetics are assumed for degradation processes) versus deposition. During time, it appears that topsoil concentrations reduce rapidly. In contrast, since no degradation term is applied on deposition, this transfer is overestimated for plant contamination (photolysis process is not integrated in the multimedia modeling). Many PAHs such as benzo[a]pyrene are, however, rapidly destroyed by UV light. In the absence of local sources a pronounced seasonal trend has been demonstrated by the Co-operative Program for Monitoring and Evaluation of the Long-range Transmissions of Air Pollutants in Europe (EMEP, 2001).

5.3.3 Combining measured and modeled data

Measured data were privileged in the process of exposure modeling permitting to take in account past sources. Specifically for the soil compartment, measured data are used to estimate initial concentrations though appropriate processing. Because soil is able to store, its integration in the exposure modelling permits to reflect potential past population exposure.

Measurements were also used for the construction of atmospheric concentration maps of B[a]P. To compensate for the limited number of monitoring sites and improve the mapping, measurements were coupled with emission-based data. This requires method development and data pre-processing to find the most relevant variables to be used in combination with PAH concentrations. Kriging variance may then be used to weight the resulting maps as a function of the quality of te interpolation.

Food categories were also constructed for the sake of consistency. The determination of these categories permits the homogenization and the data coupling by constructing homogeneous calculation. Based on data included on scientific articles or available national

or international databases, those categorizations were made for estimating:

- local vegetation modeling parameters origin;
- national residue concentration in commercial food products;
- food intake by age class and homegrown product rates;
- soil-plant transfer factors.

The multiple aggregations often lead to approximations and a possible information loss, however they are necessary to allow the coupling of different types of data. Constructing additional subcategories or including qualitative variables in the modelling could potentially improve the modeling accuracy.

5.3.4 Interoperability of the methodological framework

The management, manipulation, analysis and modeling of spatially referenced data were conducted using a statistical language and a GIS. That way it is possible to integrate all available data, despite their heterogeneity in a common spatial support. The zone system selected, allows the reflection of local variations, and the integration of environmental monitoring databases in France.

The GIS modeling platform, enables the coupling and interoperability of all spatial data via the reference grid on which the input data and the variables of interest are discretized after processing:

- environmental contaminant concentrations through indicator construction from water-air-soil environmental data;
- population data describing population densities and nutritional habits of reference subgroups;
- concentrations in exposure media, exposure doses and risk indicators from the multimedia modeling.

The multimedia approach permits to include the concentrations from the different environmental compartments and their interconnections. The exposure assessment provided a methodological framework and the definition of an integrating metric that include all the exposure media of interest.

5.3.5 Final results

The final results of this work showed that PAH exposure map is a result of the inhalation and ingestion contribution combination (Figure 4.20). Concerning the ingestion, commercial products have the highest contribution towards the total exposure dose, with drinking water and vegetables following with similar contributions. Soil appears to have the smallest contribution for both age groups. Higher intake estimates for children is a result of the higher bodyweight normalized ingestion rate, amount of soil and food daily ingested; this is illustrated also in the analysis of various pathways contributions as well as when integrating among exposure pathways. To understand the dietary intake of food contaminants like PAHs, one needs to describe the relationship between contamination quantities contained in food products as well as nutritional habits [231; 50].

Average values for the French two age classes and from other surveys are presented in table 5.2 (C1: children and C2: adults) in terms of average daily dose (ng/kg bw/day). Conversion of unit is based assuming a ventilation rate of 20 m3/day, a water ingestion of 2 L/day. and a body weight of 60 kg for an adult.

	C1	C2	INTEGRA-C1	EFSA report C2
Air	0.22	0.12	1.0E-06	0.32
Soil	0.07	0.01	8.3E-03	
Water	0.15	0.08	6.5E-04	0.003-0.03
Food	0.65	0.37		3.95
Vegetables	0.12	0.08		
Cigarettes (10)				0.9
Dermal			9.0E-08	
Total	1.21	0.65	9.0E-03	4.45

Table 5.2: Average daily dose (ng/kg bw/day) (C1 : children and C2 : adults).

Diet (including vegetables) is the dominant exposure pathways (uptake rate of 0.77 ng/kg bw/day for children), followed by inhalation (uptake rate at 0.12- 0.22 ng/kg bw/day), water ingestion (uptake rate for children at 0,15 ng/kg bw/day) and soil ingestion (uptake rate of 0.07 ng/kg bw/day for children). In term of dose, oral is the dominant route of exposure (uptake rate at 0.99-ng/kg bw/day) followed by inhalation (uptake rate at 1E-06 ng/kg bw/day).

In a similar study that was conducted on human exposure assessment to B[a]P in

Europe using the INTEGRA multimedia exposure model [176], soil and dust ingestion are shown to be the dominant exposure pathways for children followed by diet and inhalation (where only continental emission sources were considered). However, when local emissions are included in the exposure scenario as in this study, inhalation becomes the dominant route of exposure. The INTEGRA survey differ in terms of modeling hypothesis and data input but data reported in the abstract [175] looks to be erroneous.

For B[a]P, contributions from particulate deposition transfers is the most important in local food, with the exception of root-vegetables. In general, the stronger the solubility of a contaminant, the more the root transfers are favored. The solubility of PAH pollutants in water is particularly low which explains why the impact of root transfers is so weak in comparison with other pollutants as trace metals (specifically for cadmium [33]) and also why concentration in drinking water is also low. The EFSA Panel on Contaminants in the Food Chain (CONTAM Panel) reviewed the available data on occurrence and toxicity of PAHs. In total, results from 9714 PAH analyses in 33 food categories/subcategories were evaluated with a specific focus on B[A]P. The median dietary exposure across European countries was calculated both for mean and high dietary consumers and varied between 235 ng/day (3.9 ng/kg bw/day), 389 ng/day (6.5 ng/kg bw/day) for benzo[a]pyrene. The value estimated for France is 3.95 ng/kg bw per day). The difference between EAT2 and European reporting for France is probably due to different monitoring campaign reporting and to food categorization differences.

In a study by Lodovici et al. [114], assuming 6.5 ng as a mean delivery and that 80% of inhaled particle-bound BAP from the mainstream smoke is deposited in the respiratory tract (WHO/IARC, 1986), the additional benzo[a]pyrene intake for a person smoking 10 cigarettes/day is estimated to 57 ng, which is below but in the same order of magnitude (0.9 ng/kg bw/day) of C1-C2 intake by ingestion of food (0.65-1.21 ng/kg bw/day). Our results for water intake is quite high in comparison with EFSA. It may be due to imputation method used in our work. Inhalation become the dominant pathway for BaP exposure for smokers, while for non-smokers the main route of exposure is through food. Those results are coherent with other research [96; 222; 224].

The concentrations in the predicted local food products and those measured in the second Total Diet Study (EAT2) are significantly different for all three substances (Figure 4.18). This can be partially explained by the contributions from transfers related to particulate deposits and resuspension of soil particles are higher in plant compartments. However, the potential overestimation of the local plant matrices does not have a significant influence

on the total LADD calculation, as it is essentially dependent on the from commercial diet intake contribution.

5.4 Uncertainties

Uncertainty is inescapable in the assessment of environmental exposure and the consequent risks to human health, and it arises at every stage in these assessments. It causes an increased risk of incorrect decisions being made in the assessment In modeling exercises, uncertainty of the model output is the combination of the three main sources generally described as: model structure uncertainty, model parameter uncertainty and model input uncertainty. Those three have to be considered specifically when results is used to prioritize public exposure reduction action. These include the systematic identification of the variability and uncertainty sources, so that the uncertainty analysis can be conducted through the whole process of exposure assessment.

Although, the main focus is on quantifying variability and uncertainty in model inputs and assessing their effect on model outputs, uncertainties in scenarios and models are also important.

5.4.1 Model structure uncertainty

The multimedia exposure used is composed of equations and parameters. The equations describe quantitatively the natural processes being modelled, usually as equilibrium or rate expressions. The parameters describe relevant properties of the specific system under investigation. For our modeling these parameters can be classified as chemical properties, environmental characteristics, exposure scenarios, toxicological values. Together the equations and parameters produce the model results to be used to map exposure indicators. In this way, our work select or adapt the subjective most relevant equations and the parameters in regard with the thesis objective. However, a model is necessarily a simplification of a complex system with an associated uncertainty due to the mathematical approximations and the uncertainty and variance in the parameters.

The relationship between model complexity and uncertainty is complex itself: Increasingly complex models may reduce model framework/theory uncertainty as more scientific understandings are incorporated into the model, however, as models become more complex by including additional their performance can degrade because they require more input variables, leading to greater data uncertainty At first, transfers are estimated only in one

direction, from one compartment to another, without considering the return exchanges (from the plant to the soil, for example). Mass-balance approach was excluded for time calculation reason. That way, the pollutant's mass is not conserved between the environmental compartments and the secondary transfers are not accounted for completely. This was shown to lead to an overestimation of transfers. Even though the decoupling of the environmental compartments facilitates the implementation and simplifies the system of equations and initial assumptions, it also limits the dynamic dimension of the model and the estimation of the temporal evolution of the environmental concentrations.

The role of uncertainty and sensitivity analyses differs with model applications. When model is being used to rank areas, population, exposure media and environmental contribution, it is necessary to determine whether differences in model outcome are significant and whether such differences are real or artifacts due to lack or in data. For example, since deposition transfer constitute the most important plant contamination contribution, it is important to identify why some additional phenomena (as photolysis degradation) could be integrated in the modeling to gain accuracy. An accurate description of the exposure using multimedia modeling, would require additional physical, chemical, or biological processes. However for computational efficiency, a simplified description of the phenomena was preferred and certain aspects were disregarded (such as the bioconcentration due to the gas-plant transfers). Scenario uncertainty is typically associated with qualitative issues regarding what is included or excluded and is dependent on the state of knowledge and judgment regarding the real-world system and judgment regarding which aspects of the real-world system are salient with respect to human exposures. Decisions regarding what to include or exclude from a scenario could be recast as hypotheses regarding which pathways, micro-environments and so on contribute significantly to the overall exposure of interest. Here, decisions whether to include or exclude a particular aspect of a scenario has been subject to a first bibliography that permits to identify principal pathway. For example, we exclude dermal contact as a significant pathway. The results of Sarrigiannis [176] permit to confirm our choice afterwards. However, more scientific approach could be used using sensitivity analysis.

With the term stochastic or probabilistic is described the model that includes the presence of randomness in one or more of its input parameters or variables. That way, the output is probabilities of occurrences of exposures in a given population.

Most often, exposure models are too complex to be solved analytically, and approaches such as Monte Carlo simulations are used to predict output distributions. Random model

input variables are represented as probability density distributions from which values are selected randomly and substituted into the equations of the model to calculate the output. In order to simulate and predict the probability density function for the modelled event in the target system, this process is repeated. This approach was initially considered and research of important parameter variables were made (specifically for transfer factor). Nevertheless, the lack of this kind of value for PAH did not permit us to build relevant statistical distribution. Consequently, deterministic approach was preferred to avoid useless additional calculation time (thus prevented in the same time to propagate parameter uncertainties).

5.4.2 Model parameter uncertainty

This work aims to compare areas and exposure media contribution. In contrast to the conservative classical risk assessment approach, our exposure framework required to build realistic parameter in order to avoid potential distortions between exposure pathway contribution, scenarios or spatialize contamination estimations. Those potential distortions may affect the spatial pattern and lead to identify erroneous populations.

Transfer chemical-to-chemical differences in the environment are controlled by physical-chemical properties. The chemical property parametrization is determined using experiment data. In deterministic approach, this is given as a mean value. For some properties such as degradation half-lives, there is also variation unrelated to experimental precision. Factors make it impossible to assign the substance as a single mean. For the solubility constant, these variations are also generally characterized by lognormal distributions that correspond to a distribution of possible values in the general environmental conditions.

A source of parametric uncertainty in the multimedia modeling could be introduce by parameters selected to describe the different transfers, such as the soil-plant. The values used in the modeling process are an average from the rare findings in the associated bibliography that can vary by a factor of 1 to 100. This lack of exact values to describe the phenomena results in uncertainty and could maybe explain possible underestimation of root vegetables (by comparing modeling results with measured EAT 2 value).

The data used to describe the concentrations found in commercial products were taken from the EAT 2 study. In this study, the undetected values were substituted by a constant proposed by ANSES. These undetected values may lead to bias the evaluation of the relative contributions of the exposure pathways and the calculation of the LADD. Moreover, the homegrown product rates considered for this study were elaborated using data from

1990 and therefore might not correspond to the current trends. This could result in an over-estimation of the exposure due to ingestion of local products and generate potentially distortions between the considered scenarios. Local products are consumed mainly in in rural areas, while commercial products are being consumed more in urban areas thus the impact of contamination related to ingestion due to local environmental compartments is also related to the level of urbanism of the area considered.

A more complex model does not necessarily imply less uncertainty. Those different examples permit to illustrate all questions that have been posed during our work. At the end, we tried to find a best compromise between work time and accuracy to select collect, and process relevant data related to our work objectives and model framework.

5.4.3 Model input uncertainty

In this work, one of the main objective was to optimize data input spatial database. Different limitations were imposed by the quality of available databases for each environmental compartment. Efforts are made to select and complete, accurate and up-to-date available datasets. Geostatistics provide straightforward estimates of the environmental contamination by kriging concentration. Moreover, those method allows one to address the lack of direct measurements of pollutant concentrations by incorporating denser, soft information.

In regard to the soil compartment, a first approach was employed to address the issue of values under the detection limit. Indicator kriging permit to consider a part of the monitoring uncertainty, while other supplementary variables were processed in order to improve the precision of the estimation and get a more refine description of concentrations in the soil.

The cross-validation method computes the quality of the spatial interpolation for each measurement station point from all available information except from the station point under examination, i.e. it withholds one data point and then makes a prediction at the spatial location of that point. This procedure is repeated for all measurement points in the available set, enabling the evaluation of the quality of the predicted values at locations without measurements, as long as they are within the area covered by the measurements. The results of the cross-validation are described by the statistical indicators and scatter plots. The indicators used are root mean squared error (RMSE). The uncertainty estimation of the French topsoil B[a]P concentration map is based on cross-validation and on the interpolation standard error map, based on kriging theory.

As the distance to polluted soil and soil properties have been shown to affect the distribution of PAH [161], including them in the modeling allows to account for these differences in the soil composition and their effect in the final map of PAHs. This way the local uncertainty due to the local variation of these properties can be diminished. However the non-linear relationship between them and the PAH measurements make the modeling challenging.

Concerning the air compartment, the spatialization of the limited and sparse number of observations could lead to misclassification of the exposure. To address this issue, the use of auxiliary variables such as the emissions, the altitude and the population improved the spatial prediction. The relative uncertainty is mostly driven by the density of the measurement network, because residual kriging results are largely driven by observations. Thus, more measurements, particularly in the black areas in X , would reduce the uncertainty in the B[a]P concentration map. On the other hand, the chosen auxiliary variables also have an impact in the spatial interpolated results, especially where the measurement data are scarce. The quality of these air quality concentration maps relies primarily on the quality of the input data used, namely the quality of the observations, modelled results and auxiliary information introduced in the spatial interpolation. Additional works have been made to capture uncertainties due to the integration of those additional data. We failed to develop a relevant method to address this issue and characterize model limit due to lacking data in specific situation. For example the introduction of altitude layer in the model permits to correct the domestic emission underestimation due to wood burning. Nevertheless, because no monitoring stations are localized in empty areas (no inhabitant) the model is unable to take in account the limit of this trend specifically in mountainous area where predictions are overestimated. A simplified approach was chosen to make appear this model limitation in the final map. Areas are reported censored when the uncertainty is considered too high to be reliable, in order to avoid the improper characterization of exposure. Specifically, we define area conditions of not-relevant estimation: altitude should be above of 1800 m and population below population below 200/3 km^2.

A way to present these results is by omitting the areas were the uncertainty is considered outside of our model limit, in order to prevent erroneous conclusions (Figure 5.4.4).

In the modeling process, two years were considered in order to have a representative average of a chronic exposure. It should be noted that there are several factors of temporal variability (meteorology, emission evolution,...) that could lead to great deviations over the exposure period considered (30 years).

Despite the plurality of available data in the water SISE'Eaux database, the majority of reported values was under the detection limit. Instead, the multiple imputation method selected acknowledges the uncertainty by generating m plausible values for each unobserved data point. This way the uncertainty can be reported in the form of variance between the imputed datasets. due to the complexity of the water network in France, assessing the population served by each distribution unit implies an additional assumption in the modeling process. In a similar way than for air concentration map, areas where no data permit to characterize potential drinking water contribution. A validation method has been elaborated using more detected pollutant dataset (arsenic and selenium), that was artificially imputed in order to test the imputation method. The results showed that the method employed provided more reliable results than the method of simple substitution.

5.4.4 Reporting results to peers, stakeholders and decision makers

As regulators rely increasingly on maps as decision-making tools, it is vital that the variance of the model outcome and its sources be thoroughly understood. There is inevitably uncertainty about potential exposure at unsampled areas, and this should be taken into account when decisions are made to reduce exposure.

The way the results from the spatial exposure assessment should be reported varies according to the objective and the audience. Uncertainty should be included in the decision making, as ignoring it could result in to improper interpretation.

A way to report the uncertainties on a map can be in the form of intervals. In this regard, the use of ranges of values for a particular metric of decision-making relevance (e.g., range of uncertainty associated with a particular estimate of exposure) may be adequate, as long as the bounds of the range are properly placed with a probabilistic context (e.g., 95% confidence intervals, interquartile range). One other approach that we tested was to present map with adapted grid size related to data precision or population density. Finally we have chosen is to add additional information on the total EIR map (Figure 5.4.4):

- Grey borders are associated with grids where water data were missing (no data);
- Black cross-hatch is used to hide areas outside of the validity limit of the air statistic model.

Those information's should be clearly reflected in the presented results when the associated uncertainty is too high for the results to be trustworthy. This way, decision-

makers are ought to understand whether their decisions are likely to be improved by waiting for additional information on critical factors influencing these decisions, or regarding the need to consider an adaptive strategy based on iterative reassessment.

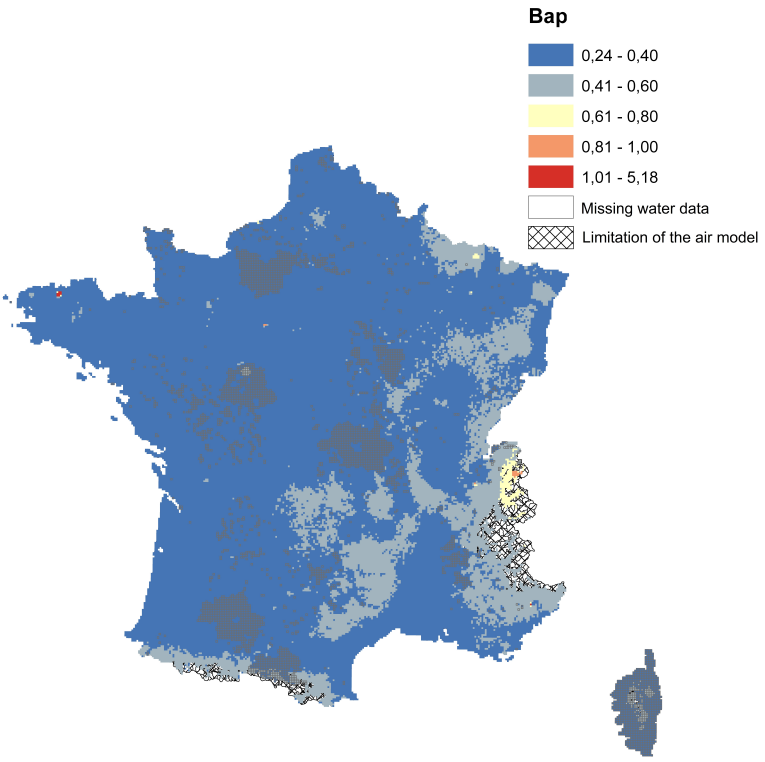


Figure 5.5: Final map of B[a
P EIR including the model's limitations.

Chapter 6

Conclusion and Perspectives

6.1 Conclusion

Spatial characterization of environmental inequalities requires a fine description of human exposure. This involves the integration of data able to describe the environmental compartments quality, and describe contaminant transfer phenomena that operate in different spatio-temporal scales. Moreover, it requires the description of the principal exposure routes, past and present sources and realistic exposure scenarios.

This thesis work aimed to develop statistical method to improve the limited representativeness of available data. The processed data are used as input of the exposure model to map PAH exposure indicator at a fine resolution in France. The present study required significant preparatory work in terms of data collection, data processing and implementation of modeling tools. Indeed, the use of databases not initially assembled for the objective of characterizing exposure may lead to biases in the estimation. Common problems include limited number of observations, spatial and temporal heterogeneity of geographic supports, lack of meta-data and measurements accuracy. Different statistical and geoprocessing techniques have been put in place to improve the variables of interest.

Atmospheric concentration data were collected in France in the context of regulatory surveillance for two years (2010 and 2011). Estimation of concentrations over France by classical interpolation method could lead to a misrepresentation of the spatial distribution due to the limited number of observations. To address this issue, auxiliary variables in the context of external drift kriging are employed. The best auxiliary variable to define linear drifts was found to be the one that includes the atmospheric emissions as well as the

population and the altitude.

Measurements of PAH topsoil concentrations are available through the French Soil Monitoring Network. Qualitative data on the polluted sites localization are integrated by processing distance-to-polluted soil proxy. These along with 14 variables about physic-ochemical soil properties were combined in a hybrid regression-kriging and fitted using Random Forest models, was shown to outperform the traditionally used linear regression.

Due to its hydrophobic nature, B[a]P is found in water in small concentrations, therefore the exact measurement cannot always be reported. The observations under the detection limit rate is quite high, which requires careful handling. A complex multiple imputation method was developed in order to extract the maximum information from the available measurements without introducing too much bias in the final results. This one permits to take advantage of the temporal aspect and correlations between substance of interest and other PAH substances. Spatial estimation of water concentrations is carried out by taking in account the multi-annual data and the network water distribution complexity using a bootstrap based expectation-maximization algorithm. The above methods permitted the construction of a representative spatial database in a 9 km^2 grid of reference used to perform the integrated exposure assessment.

The use of integrated tools enabled the analysis of exposure area, and the contribution of environmental compartments and exposure media. Two age group were defined to assess the exposure media contribution: one for children (2-17 years old) and one for adults (age 17-70 years old).

Concentrations of PAH estimated in the different compartments were integrated in a multimedia exposure model to yield the aggregated exposure estimation and associated environmental determinant. Correlation analysis between the spatialized exposure indicator and concentration found in the environmental compartments allows to highlight the most sensitive environmental data in the model. The single pollutant (B[a]P) exposure analysis was favored: We found very high correlation between B[a]p and the other 16 prioritized PAH in each environmental compartments and due to its higher toxicological reference value it has been used as an indicator species for PAH contamination.

Diet is the dominant exposure pathways, followed by inhalation, water ingestion and soil ingestion. Children are the most highly exposed.

The exposure map spatial pattern depends on the local combination of the different pathways and their local interrelationships. Our results permitted us to identify pollutant sources, determinants of exposure, and potential hotspot areas.

Some hotspots correspond to high air concentration values localized in the mountain areas where air dispersion is limited. Concerning the ingestion, commercial products have the highest contribution towards the total exposure dose, with drinking water and vegetables following with similar contributions. Soil appears to have the smallest contribution for the different age group scenarios. The highest hotpot due to water ingestion is localized in the proximity of an old gas factory.

In summary:

- GIS provides geoprocessing tools to assembled databases from different sources;
- developed statistic methods improve spatial and temporal data representativeness;
- the exposure multimedia links environmental and exposure compartments;
- the exposure assessment framework provides a methodological framework to combine and integrated all exposure pathways.

Uncertainty is inescapable in the assessment of environmental exposure and the consequent risks to human health, and it arises at every stage in these assessments. It causes an increased risk of incorrect decisions being made in the assessment. In our modeling exercises, uncertainty of the model output is the combination of three main sources described as: model structure uncertainty, model parameter uncertainty and model input uncertainty.

The analysis of all the uncertainties of each link in the modeling chain would allow the characterization of the sources of errors, the improvement of the interpretability of the results and the attribution of confidence levels relative to the results obtained. A first examination would consist of an analysis of the propagation of errors during the data processing, parametric uncertainties and modeling. Most of them weren't assessed for different reasons: calculation time, too complex uncertainty propagation to characterize, missing data... The approach here consisted in characterizing, on one hand, the representativeness of the input data in order to estimate the potential overestimations or underestimates and, on the other hand, the areas for which no information is available to detect potential overexposure.

Mapping the PAH concentrations is the preliminary step towards decision making such as the declination of regional and national plans for preventative or curative actions. The use of the above results in a management framework, however, should be accompanied by the variable maps to allow the interpretation of the representativeness of the prediction.

In this way, we adopted to present the exposure indicator map with additional information's to clearly reflected in the results when the associated uncertainty is too high for

the results to be trustworthy. This way, decision-makers are ought to understand whether their decisions are likely to be improved by waiting for additional information on critical factors influencing these decisions, or regarding the need to consider an adaptive strategy based on iterative reassessment.

The lack of available raw data poses challenges in the exposure assessment process, in the desired aggregation level. Access to collected data through the establishment of partnerships would permit the improvement of the characterization of risk and limit the uncertainties associated with data availability. Crossing the daily predicted exposure doses with internal doses and their metabolites, as they are measured in blood, urine, saliva or tissue would provide a more robust basis for the risk assessment, despite the potential high cost of such campaign. Focusing our studies specifically on the areas at risk of overexposure, would allow us to acquire more precise information to carry out a validation for each pollutant.

6.2 Perspectives

This work has been carried out here under the "Health Risk Assessment" research axis of the INERIS and is included in the framework of the PLAINE project. The objective of this project is the construction of a GIS-based modelling platform for quantifying human exposure in France, by integrating environmental, socio-economic and health (spatial) information. In this context, an effort was made in this work to respond to the challenges of the management of complex environmental situations, such as the issue of environmental inequalities. The aim was to develop tools to provide knowledge on pollutant exposures to environmental contaminants in order to guarantee adequate and proportionate management measures at different scales. In this objective, the algorithm developed in the context of this thesis to process environmental data will be integrated in the platform to be used for other pollutants. For example, since water processing method permits to provide associated variance, the characterization of lead (Pb) exposure in a project with US EPA and INERIS will integrate propagation of uncertainties (water compartment assessment was the last part of the model without processing method able to generate uncertainties for this pollutant approach).

As a perspective, the coupling of the spatial approach with the multimedia modelling in a stochastic/probabilistic mode would permit the characterization of the total uncertainty. A first examination would consist in the analysis of error propagation along the different

stages of data processing and modelling: initially in the spatialization process in the different environmental compartments and then in the multimedia modeling. Nevertheless, due to time calculation required this update implies to change code platform. In this way, the thesis of Corentin Regrain (INERIS/LAMFA laboratory-University of Picardy) aims to optimize INERIS exposure model calculation tool.

The approach developed in this thesis allows the identification of possible exposure inequalities of the populations, and their determinants, by the construction of exposure indicators for three PAH substances. In the same way, processing method will be used in the context of the CartoExpo project (lead by INERIS and UMR PériTox-University of Picardie) focusing in pesticide substances.

In the process of environmental inequalities characterization to help manage the related risk, it was required to construct continuous variables, over a specific area, for which data are not exhaustively monitored, thus introducing that way uncertainties. It is then necessary to provide elements that allow the evaluation of the representativeness of the input data and their impact to the overall risk calculation. The use of the proposed results in a management framework should therefore be accompanied by additional maps. The characterization of the information density and the type of data from which the indicators were constructed allows the interpretation of the representativeness of the predictions. Such additional maps will include the kriging variance for air and soil or the emission zones characterization for inhalation and deposits. Those maps will be used for the orientation of additional data collection campaigns in areas where limited data is available or those where overexposure of the population is suspected. Finally, the additional maps will allow:

- to estimate the level of confidence of the results;
- to acknowledge the intrinsic biases of the input data;
- to evaluate the potential over or underestimations in the predictions;
- to determine the areas for which additional data collection is needed.

Each region drew up a Regional Environment and Health Action Plan (Plan Régional Santé Environnement-PRSE) to implement the main objectives of the French National Action Plan according to its own specific needs. Different regions in France have included environmental health inequalities reduction in their planning, and need assessment to guide priorities for voluntary action. The Ile-de-France region has chosen this priority in its plan and PAH map built here is supposed to integrate the data catalog that will be used to

build environmental health diagnostic. Approach will be developped for evaluating spatial relationships between health outcomes, socioeconomic indicator, and PAH map as a first step to build deprivation indicators.

In this work an effort was made to respond to the issues raised at the national level and in support of the public authorities, as they were established in PNSE 3. Our work participates to the realization of 4 actions specifically dedicated to the environmental inequality thematic:

- Action no 38: develop and disseminate, via a common platform, reference methodologies at the national level for the characterization of locally disparate environmental inequalities, taking into account the vulnerable situations of the populations
- Action no 34: identify and analyze the methods of constructing spatialized and integrated exposure indicators
- Action no 39: using tools for analyzing environmental inequalities to cross exposure models and population data (biomonitoring, epidemiological data, social and health vulnerabilities)
- Action no 40: implement in the context of the PRSE multi-exposure studies in several territories, based on methodological references.

The exposure data produced in this work will also be integrated in the European data platform IPChem (Information Platform for Chemical Monitoring) in the context of the HBM4EU program (H2020 Initiative: coordinate and provide better evidence of the actual exposure of citizens to chemicals and the possible health effects to support policy making) to be compared with biomonitoring data.

Bibliography

[1] H. Abdel-Shafy and M. Mohamed-Mansour. A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. (25: 107-123), 2015.

[2] S. Ahmed and G. De Marsily. Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, (23(9): 1717-1737), 1987.

[3] A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and M. Giovis C Jerrett. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, (15(2): 185-204), 2005.

[4] B. Armstrong, E. Hutchinson, J. Unwin, and T. Fletcher. Lung cancer risk after exposure to polycyclic aromatic hydrocarbons: A review and meta-analysis. *Environmental Health Perspectives*, (112(9): 970 - 978), 2004.

[5] St. Aronoff. Geographic information systems: A management perspective. *Geocarto International*, (4(4): 58-58), 1989.

[6] A. M. Astel, L. Chepanova, and V. Simeonov. Soil contamination interpretation by the use of monitoring data analysis. *Water, Air, & Soil Pollution*, (216 (1): 375-390), 2011.

[7] T. C. Bailey, P. J. Diggle, and A. C. Rowlingson BS Gatrell. Spatial point pattern analysis andits application in geographical epidemiology transactions. *Institute of the British Geographers*, (21(1): 256–274), 1996.

[8] S. Banerjee, A.E. Gelfand, and B.P. Carlin. *Hierarchical modeling and analysis for spatial data.* Monographs on Statistics and Probability, 2003.

[9] D. K. Barupal, D. Wishart, P. Vineis, and S. M. Scalbert A Rappaport. The blood exposome and its role in discovering causes of disease. *Environmental Health Perspectives*, (122(8): 769–774), 2014.

[10] O. Baskan and O. Erpul G, Dengiz. Comparing the efficiency of ordinary kriging and cokriging to estimate the atterberg limits spatially using some soil physical properties. *Clay Minerals*, (44(2): 181-193), 2009.

[11] B. Beckerman, M. Jerrett, and S. B. Brauer M Henderson. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environ Sci Technol*, (41(7): 2422–2428), 2007.

[12] R. Beelen, G. Hoek, E. Pebesma, D. Vienneau, K. de Hoogh, and D.J. Briggs. Mapping of background air pollution at a fine spatial scale across the european union. *Science of Total Environment*, (407:1852–1867), 2009.

[13] V. Been. Locally undesirable land uses in minority neighborhoods: Disproportionate siting or market dynamics? *The Yale Law Journal*, (103(6): 1383-1422), 1994.

[14] C. Bernand-Michel. *These: Indicateurs geostatistiques de la pollution dans les cours d'eau*. PhD thesis, Ecole Nationale Superieure des Mines de Paris, 2006.

[15] E. Bielecka. A dasymetric population density map of poland. *Proceedings of the 22nd International*, 2005.

[16] F.P. Bierkens. Using stratification and residual kriging to map soil pollution in urban areas. (996-1007), 1997.

[17] G. Boeije and F. Koormann. Great-er ii, chemical fate models, georeferenced regional exposure assessment tool for european rivers. Technical report, 2003.

[18] F. C. Bolstad PV Collins. A comparison of spatial interpolation techniques in temperature estimation. *Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, NM. Santa Barbara, CA: National Center for Geographic Information and Analysis, Santa Barbara*, 1996.

[19] G. Bolte and H. Fromme. Gme study groupsocioeconomic determinants of children's environmental tobacco smoke exposure and family's home smoking policy. *European Journal of Public Health*, (19(1): 52-8), 2009.

[20] R. Bonnard. Evaluation de l'impact sur la santé des rejets atmosphériques des tranchescharbon d'une grande installation de combustion - partie 2 : Exposition par voiesindirectes. *INERIS*, 2003.

[21] R. Bonnard and T.E. McKone. Integration of the predictions of two models with dose mea-surements in a case study of children exposed to the emissions of a lead smelter. *Human and Ecological Risk Assessment*, (15(6): 1203-1226), 2010.

[22] J. S. Brainard, A. P. Jones, I. J. Bateman, A. A. Lovett, and P.J. Fallon. Modelling environmental equity: access to air quality. *Environment and Planning A*, (34(4): 695-716), 2002.

[23] M. Braubach and J. Fairburn. Social inequities in environmental risks associated with housing and residential location—a review of evidence. *European Journal of Public Health*, (20(1): 36–42), 2010.

[24] M. Brauer, G. Hoek, P. Van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, and B. Brunekreef. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information. *Epidemiology*, (14:228–239), 2003.

[25] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and regression trees*. Taylor & Francis, 1984.

[26] D. Briggs. The role of gis: Coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A*, (68(13-14): 1243-1261), 2005.

[27] D. J. Brus, J. J. De Gruijter, B. A. Marsman, R. Visschers, K. Bregt, A. Breeuwsma, and J. Bouma. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. *Environmetrics*, (7: 1-16), 1996.

[28] R. Bullard. *Unequal protection: environmental justice and communities of color*. Sierra Club Books, 1994.

[29] S.W. Burchiel and M.I. Luster. Signaling by environmental polycyclic aromatic hydrocarbons in human lymphocytes. *Clinical Immunology*, (98(1): 2 - 10), 2001.

[30] P. A. Burrough. Principles of geographical information systems for land ressources assesment. *Geocarto International*, (1(3): 54), 1986.

[31] M. Callahan and K. Sexton. If cumulative risk assessment is the answer, what is the question? *Environmental health perspectives*, (115: 799-806), 2007.

[32] J. Cattle, A.B. McBratney, and B. Minasny. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. *Environmental Quality*, (31(5): 1576-1588), 2002.

[33] J. Caudeville. *Développement d'une plateforme intégrée pour la cartographie de l'exposition des populations aux substances chimiques.* PhD thesis, University of Technology of Compiegne, 2011.

[34] S. C. Chen and C. M. Liao. Health risk assessment on human exposed to environmental polycyclic aromatic hydrocarbon pollution sources. *Science of the Total Environment*, (366: 112-123), 2006.

[35] N. Cheruiyot, W.J. Lee, J. Kennedy, L.C. Wang, N.H. Lin, J. Cao, R. Zhang, and G.P. Chang-Chien. An overview: Polycyclic aromatic hydrocarbon emissions from the stationary and mobile sources and in the ambient air. (15: 2730-2762), 2015.

[36] J.P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty.* Wiley, 2012.

[37] A. C. Cohen. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, (1(3): 217-237), 1959.

[38] S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. Van Reeuwijk, K. Smallbone, and D. J. Van Der Veen A Briggs. Mapping urban air pollution using gis: a regression-based approach. *International Journal of Geographical Information Science*, (11(7): 699–718), 1997.

[39] P. Congdon. *Bayesian Statistical Modelling.* Wiley, 2001.

[40] S. M. Corwinb DL Lescha. Prediction of spatial soil property information from ancillary sensor data using ordinary linear regression: Model derivations, residual assumptions and model validation tests. *Geoderma*, (148(2): 130-140), 2008.

[41] I. T. Cousins, M. Hauck, J. V. Harbersb, and J. M. Huijbregtsb MAJ Armitage. Empirical evaluation of spatial and non-spatial european-scale multimedia fate models: results and implictions for chemical risk assessment. *Environmental Monitoring*, (6: 572-581), 2007.

[42] F. Couvidat. *Modélisation des particules organiques dans l'atmosphère*. PhD thesis, University of Paris-Est, 2012.

[43] N. A. Cressie. *Statistics for Spatial Data*. John Wiley and Sons, Inc., 1993.

[44] S. Cutter. Race, class and environmental justice. *Progress in Human Geography*, (19(1): 111-122), 1995.

[45] H. Davezac, G Grandguillot, and A et al. Robin. L'eau potable en france 2005 -2006. *Ministère de la Santé, de la Jeunesse, des Sports et de la Vie associative*, 2008.

[46] A. Davidson JV MacEachren. Sampling and isometric mapping of continuous geographic surfaces. *The American Cartographer*, (14(4): 299-320), 1987.

[47] H. Demougeot-Renard and C. De Fouquet. Geostatistical approach for assessing soil volumes requiring remediation: Validation using lead-polluted soils underlying a former smelting works. *Environmental Science & Technology*, (38(19): 5120-5126), 2004.

[48] D. C. Dolinoy and M.L. Miranda. Gis modeling of air toxics releases from tri-reporting and non-tri-reporting facilities: Impacts for environmental justice. *Environmental Health Perspectives*, (112(17): 1117-1124), 2004.

[49] L. Downey. Environmental injustice: Is race or income a better predictor? *Social Science Quarterly*, (79(4): 766-778), 1998.

[50] X. Duan, G. Shen, H. Yang, J. Tian, F. Wei, J. Gong, and J.J. Zhang. Dietary intake polycyclic aromatic hydrocarbons (pahs) and associated cancer risk in a cohort of chinese urban adults: Inter- and intra-individual variability. *Chemosphere*, (144: 2469-2475), 2016.

[51] P. Egeghy. Determinants of temporal variability in nhexas-maryland environmental concentrations, exposures, and biomarkers. *Journal of Exposure Analysis and Environmental Epidemiology*, (15: 388–397), 2005.

[52] P. Elliott, D. Briggs, S. Morris, de Hoogh C, C. Hurt, T. Kold Jensen, I. Maitland, S. Richardson, J. Wakefield, and L. Jarup. Risk of adverse birth outcomes in populations living near landfill sites. *British Medical Journal*, (323: 363–368), 2001.

[53] Donnay J. P. Ali M. Emch M. Spatial filtering using a raster geographic information system: methods for scaling health and environmental data. *Health Place*, (8(2): 85–92), 2002.

[54] US EPA. Toxicological review of benzo[a]pyrene. Technical report, 2017.

[55] A.M. Evans, G.E. Rice, J.M. Wright, and L.K. Teuschler. Exploratory cumulative risk assessment (cra) approaches using secondary data. *Human and Ecological Risk Assessment: An International Journal*, (20(3): 704-723), 2014.

[56] D. Fassin, H. Grandjean, M. Kaminski, and A. Lang T Leclerc. *Les inégalités sociales de santé*. Editions La Découverte, 2000.

[57] Adam Finkel. Confronting uncertainty in risk management: A guide for decision makers. 1990.

[58] M.M. Finkelstein and K.V. Dave. Exposure estimation in the presence of nondetectable values: Another look. *American Industrial Hygiene Association*, (62(2): 195-198), 2001.

[59] R.A. Fisher. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford Science Publications, 1990.

[60] C. Fritsch, M. Cœurdassier, P. Giraudoux, F. Raoul, F. Douay, D. Rieffel, A. De Vaufleury, and R. Scheifler. Spatially explicit analysis of metal transfer to biota: Influence of soil contamination and landscape. *PLoS ONE*, (6(5): e20682), 2011.

[61] B. Fu, Y. Lü, F. Sun, S. Wang, and X. Liu M Yao. Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment. *PLoS ONE*, (8(1)), 2013.

[62] T. C. Gatrell AC Bailey. *Interactive Spatial Data Analysis*. Longman, Harlow, 1995.

[63] A. E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, (85(410): 398–409), 1990.

[64] S. Geman D Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, (6(6): 721-742), 1984.

[65] R.O. Gilbert. 27 an overview of statistical numbers related to environmental cleanup. In *Environmental Statistics*, number 12: 867 - 880. Elsevier, 1994.

[66] R.J. Glavin and P.S. Hooda. A practical examination of the use of geostatistics in the remediation of a site with a complex metal contamination history. *Soil and Sediment Contamination: An International Journal*, (14(2): 155-169), 2005.

[67] T. Glickman. Measuring environmental equity with geographic information systems. *The RFF Reader in Environmental and Resource Management*, (192–203), 1999.

[68] P. Goovaerts. Kriging vs stochastic simulation for risk analysis in soil contamination. *Geostatistics for Environmental Applications*, (9: 247-258), 1997.

[69] P. Goovaerts. Geostatistical tools for deriving block-averaged values of soil properties. *Annals of GIS*, (5(2): 88-96), 1999.

[70] P. Goovaerts. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. *International Journal of Health Geographics*, (5: 52), 2006.

[71] P. Goovaerts. Auto-ik: A 2d indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers and Geosciences*, (Volume 35, number 6, June 2009, number 1255-1270), 2009.

[72] P. Goovaerts, G. Avruskin, J. Meliker, M. Slotnick, and G. Jacquez. Modelling uncertainty about pollutant concentration and human exposure using geostatistics and a space-time information system: Application to arsenic in groundwater of southeast michigan. *Proceedings of the 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Portland*, 2004.

[73] P Goovaerts and Marc Van Meirvenne. Accounting for measurement and interpolation errors in soil contaminant mapping and decisionmaking. In *2001 Annual Conference of the International Association for Mathematical Geology, Cancún, Mexico*, 2001.

[74] C. Guerreiro, J. Horalek, F. de Leeuw, and F. Couvidat. Benzo(a)pyrene in europe: Ambient air concentrations, population exposure and health effects. (214: 657-667), 2016.

[75] J. T. Hamilton. Testing for environmental racism: prejudice, profits, political power? *Journal of Policy Analysis and Management*, (14(1): 107-132), 1995.

[76] P. Harremoës, J. Rotmans, Van der Sluijs JP, M. B. A. Van Asselt, P. Janssen, and W. E. Krayer Von Krauss MP Walker. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Journal of Integrated Assessment*, (4(1): 5-17), 2003.

[77] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[78] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, (57(1): 97–109), 1970.

[79] J. E. Hay, C. D. Stow, and K. N. Harris D Dirks. High-resolution studies of rainfall on norfolk island part ii: interpolation of rainfall data. *Journal of Hydrology*, (208(3–4): 187-193), 1998.

[80] O. HemstrOm. Health inequalities by wage income in sweden: the role of work environment. *Social Science and Medicine*, (61(3): 637-647), 2005.

[81] T. Hengl. A practical guide to geostatistical mapping of environmental luxembourg: Office for official publication of the european communities. Technical report, 2007.

[82] G.B.M. Heuvelink and M.F.P. Bierkens. Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma*, 55(1-15), 1992.

[83] G. Hoek, R. Beelen, K. De Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, (44(33): 7561–7578), 2008.

[84] B. Holgate ST Brunekreef. Air pollution and health. *The Lancet*, (360(9341): 1233-1242), 2002.

[85] J. Honaker, G. King, and M. Blackwell. Amelia ii: A program for missing data. *Journal of Statistical Software, Articles*, (45(7): 1-47), 2011.

[86] R. W. Hornung and L. Reed. Estimation of average concentration in the presence of non detectable values. *Applied Occupational and Environmental Hygiene*, (5: 46-51), 1990.

[87] A. G. Huijbregts CJ Journel. *Mining geostatistics*. Academic press, 1978.

[88] J.E. Hunter. Needed: A ban on the significance test. *Psychological Science*, (8(1): 3-7), 1997.

[89] Wolpert R. L. Best N. Ickstadt K. Spatial poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, (95(452): 1076-1088), 2000.

[90] M. Jerrett, R. T. Burnett, P. Kanaroglou, J. Eyles, N. Finkelstein, C. Giovis, and J.R. Brook. A gis–environmental justice analysis of particulate air pollution in hamilton, canada. *Environment and Planning A: Economy and Space*, (33(6)), 2001.

[91] M. Jerrett, P. DeLuca, N. Finkelstein, D. K. Verma, K. Chapman, and M. M. Sears MR Finkelstein. Relation between income, air pollution and mortality: a cohort study. *Canadian Medical Association Journal*, (169(5): 397-402), 2003.

[92] T. Jiani, Z. Yan, M. Weicun, Y. Qi, W. Jian, and C. Limin. Impact of spatial resolution on air quality simulation: A case study in a highly industrialized area in shanghai, china. *Atmospheric Pollution Research*, (6(2): 322 - 333), 2015.

[93] A. F. Jorm, A. E. Korten, H. Creasey, E. McCusker, G. A. Broe, W. Longley, and A. S. Anthony JC Henderson. Environmental risk factors for alzheimer's disease: their relationship to age of onset and to familial or sporadic types. *Psychological Medicine*, (22(2): 429-436), 1992.

[94] A.G. Journel and C. V. Deutsch. *GSLIB geostatistical software library and users guide*. Applied Geostatisics Series, 1998.

[95] P. Juarez, P. Matthews-Juarez, D. Hood, W. Im, R. Levine, B. Kilbourne, M. Langston, M. Al-Hamdan, W. Crosson, M. Estes Jr, S. Estes, V. Agboto, P. Robinson, S. Wilson, and M. Lichtveld. The public health exposome: A population-based, exposure science approach to health disparities research. (11: 12866 - 12895), 2014.

[96] K. Kim, P. Pant, and E. Yamashita. Using national household travel survey data for the assessment of transportation system vulnerabilities. *Transportation Research Record: Journal of the Transportation Research Board*, (2376: 71-80):71–80, 2013.

[97] M. Kohlhuber, A. Mielck, S. K. Weiland, and G. Bolte. Social inequality in perceived environmental exposures in relation to housing conditions in germany. *Environ Res*, (101(2): 246-55), 2006.

[98] H. Kromhout, E. Symanski, and S. M. Rappaport. A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Annals of Work Exposures and Health*, (375(3): 253–70), 1993.

[99] H. Kruize, M. Droomers, I. Van Kamp, and A. Ruijsbroek. What causes environmental inequalities and related health effects? an analysis of evolving concepts. *International Journal of Environmental Research and Public Health*, (11(6): 5807-5827), 2014.

[100] A. W. Lance and C.A. Gotway. *Applied Spatial Statistics for Public Health Data*. Wiley, 2004.

[101] T. Larssen. Acid deposition and its effects in china: an overview. *Environmental Science and Policy*, (2(1): 9-24), 1999.

[102] G. M. Laslett. Kriging and splines: an empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, (89(426): 391-400), 1994.

[103] M. D. Lebowitz. Exposure assessment needs in studies of acute health effects. *Science of the Total Environment*, (168(2): 109–117), 1995.

[104] K. W. Lee DY Juang. A comparison of three kriging methods using auxiliary variables in heavy-metal contaminated soils. *Journal of Environmental Quality*, (27(2): 355-363), 1998.

[105] P. Lepom, B. Brown, G. Hanke, R. Loos, P. Quevauviller, and J. Wollgast. Needs for reliable analytical methods for monitoring chemical pollutants in surface water under the european water framework directive. *Journal of Chromatography*, (1216(3): 302-315), 2009.

[106] J. P. LeSage and K.R. Pace. Spatial econometric models. *Handbook of Applied Spatial Analysis*, (355–376), 2009.

[107] L. Letinois. Méthodologie de repartition spatiale de la population. Technical report, 2014.

[108] J. Li and A. Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, (6(3–4): 228-241), 2011.

[109] D. Liam and M. Van Willigen. Environmental stressors: The mental health impacts of living near industrial activity. *Journal of Health and Social Behavior*, (46(3): 289–305), 2005.

[110] Q. Lin HS Zhu. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. *Pedosphere*, (20(5): 594-606), 2010.

[111] J. Lipscomb, S. Haddad, T. Poet, and K. Krishnan. Physiologically-based pharmacokinetic (pbpk) models in toxicity testing and risk assessment. *Advances in experimental medicine and biology*, (745: 76-95), 2012.

[112] Y. Liu, B. Beckingham, H. Ruegner, Z. Li, L. Ma, M. Schwientek, H. Xie, J. Zhao, and P. Grathwohl. Comparison of sedimentary pahs in the rivers of ammer (germany) and liangtan (china): Differences between early- and newly-industrialized countries. *Environmental Science and Technology*, (47(2): 701–709), 2013.

[113] M. Lloyd-Smith and L. Bell. Toxic disputes and the rise of environmental justice in australia. *International Journal of Occupational and Environmental Health*, (9(1): 14-23), 2003.

[114] M. Lodovici, V. Akpan, C. Evangelisti, and P. Dolara. Sidestream tobacco smoke as the main predictor of exposure to polycyclic aromatic hydrocarbons. *Journal of Applied Toxicology*, (24(4): 277-281).

[115] T. C. Long, B. D. Schultz, J. Crooks, M. Breen, J. E. Langstaff, K. K. Isaacs, Y.-M. Tan, R. W. Williams, Y. Cao, A. M. Geller, R. B. Devlin, S. A. Batterman, and M. S. Buckley TJ Breen. Gps-based microenvironment tracker (microtrac) model to estimate time-location of individuals for air pollution exposure assessments: Model evaluation in central north carolina. *Journal of Exposure Science & Environmental Epidemiology*, (24(4): 412–420), 2014.

[116] B. P. Louis TA Carlin. Bayes and empirical bayes methods for data analysis. *Statistics and Computing*, (7(2): 153–154), 1997.

[117] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, (10(4): 325–337), 2000.

[118] Bertrand M. Consommation et lieux d'achat des produits alimentaires en 1991. In *Collection INSEE Resultats*. 1991.

[119] D. Mackay. Dissolution of non-aqueous phase liquids in groundwater. *Journal of Contaminant Hydrology*, (81(1): 23-42), 1991.

[120] M. Maier, D. Maier, and B.J. Lloyd. Factors influencing the mobilisation of polycyclic aromatic hydrocarbons (pahs) from the coal-tar lining of water mains. *Water Research*, (34(3): 773 - 786), 2000.

[121] L. Malherbe, C. Songeur, C. Honoré, A. Ung, and F. Meleux. *Forecasting Urban Air Quality over Cities by Statistical Adaptation of Deterministic Chemistry Transport Model Outputs*. Number 367-371. Springer, 2012.

[122] B. Maliszewska-Kordybach. The effect of temperature on the rate of disappearance of polycyclic aromatic hydrocarbons from soils. *Environmental Pollution*, (79(1): 15 - 20), 1993.

[123] E. Mantseva, A. Malanichev, and N. Vulykh. Polycyclic aromatic hydrocarbons in the environment. Technical report, 2002.

[124] M. Margni, D. W. Pennington, C. Amman, and O. Jolliet. Evaluating multimedia/-multipathway model intake fraction estimates using pop emission and monitoring data. *Environmental Pollution*, (128(1–2): 263-277), 2004.

[125] J. Markus and A.B. McBratney. A review of the contamination of soil with lead. *Environment International*, (27(5): 399-411), 2001.

[126] S. Marsland. *Machine learning: an algorithmic perspective*. Chapman & Hall, 2009.

[127] M. Martuzzi. Inequalities, inequities, environmental justice in waste management and health. *European Journal of Public Health*, (20(1): 21–26), 2010.

[128] A. Martínez-Cob. Multivariate geostatistical analysis of evapotranspiration and precipitation in mountainous terrain. *Journal of Hydrology*, (174(1–2): 19-35), 1996.

[129] D. M. Mason MA Stout. The distribution of chlorpyrifos following a crack and crevice type application in the us epa indoor air quality research house. *Atmos Environ*, (37(39–40): 5539–5549), 2003.

[130] G. Matheron. The theory of regionalised variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique de Fontainebleu 5*, 1971.

[131] B. McBratney A Minasny. Spatial prediction of soil properties using eblup with the matérn covariance function. *Geoderma*, (140(4,15): 324-336), 2007.

[132] T. F. A. McBratney AB Bishop. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma*, (103(1-2): 149-160), 2001.

[133] R. C. McDonald, R. F. Isbell, J. G. Speight, J. Walker, and M.S. Hopkins. *Australian soil and land survey: field handbook*. CSIRO Publishing, 1998.

[134] P. A. McDonnell RA Burrough. *Principles of Geographic Information Systems*. Oxford University Press, 1998.

[135] J. Meliker. Arsenic in drinking water and cerebrovascular disease, diabetes mellitus, and kidney disease in michigan: a standardized mortality ratio analysis. *Environmental Health*, (6–4), 2007.

[136] L. Menut, B. Bessagnet, D. Khvorostyanov, M. Beekmann, N. Blond, A. Colette, I. Coll, G. Curci, G. Foret, A. Hodzic, S. Mailler, F. Meleux, J.-L. Monge, I. Pison, G. Siour, S. Turquety, M. Valari, R. Vautard, and M. G. Vivanco. Chimere 2013: a model for regional atmospheric composition modelling. *Geoscientific Model Development*, 6(4), 2013.

[137] Y. Meshalkina. A brief review of geostatistical methods applied in modern soil science. *Moscow University Soil Science Bulletin*, (62(2): 93–95), 2007.

[138] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, M. N. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, (21(6): 1087–1092), 1953.

[139] R. D. Moore, J. A. Floyer, M. G. AsplinaI, and K. McKendry G Stahl. Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agricultural and Forest Meteorology*, (139(3–4): 224-236), 2006.

[140] R. Morello-Frosch, M. Pastor, C. Porras, and J. Sadd. Environmental justice and regional inequality in southern california: implications for future research. *Environmental Health Perspectives*, (110(2): 149–154.), 2002.

[141] R. Morello-Frosch, L. Zeise, and G. M. Faust JB Solomon. Cumulative environmental impacts: Science and policy to protect communities. *Annual Review of Public Health*, (37: 83-96), 2016.

[142] R. K. Morgan. The evolution of environmental impact assessment in new zealand. *Journal of Environmental Management*, (16: 139-152), 1983.

[143] V. Mugica-Alvarez, S. Hernández, M. Torres, and R. García. Seasonal variation of polycyclic aromatic hydrocarbon exposure levels in mexico city. (60: 548-555), 2010.

[144] V.L. Mulder, M. Lacoste, A. Richer de Forges, and D. Arrouays. Globalsoilmap france: High-resolution spatial modelling the soils of france up to two meter depth. *Science of the Total Environment*, (573: 1352–1369), 2016.

[145] D. J. Mulla, A. G. Journel, and R. E. Franz EH Rossi. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, (62(2): 277-314), 1992.

[146] K. Murasawa, T. Sakurai, K. Nansai, K. Matsuhashi, Y. Moriguchi, K. Tanabe, O. Nakasugi, and N. Morita M Suzuki. Geo-referenced multimedia environmental fate model (g-ciems): Model formulation and comparison to the generic model and monitoring approaches. *Environmental Science and Technology*, (38(21): 5682–5693), 2004.

[147] T. Nakaya. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A*, (32(1): 91–109), 2000.

[148] C. J. Nappo. The workshop on the representativeness of meteorological observations, june 1981, boulder, colo. *Bulletin of the American Meteorological Society*, (63(7): 761-764), 1982.

[149] M. J. Nieuwenhuijsen, V. Putcha, S. Gordon, D. Heederik, K. M. Venables, and P. Cullinan. Exposure-response relations among laboratory animal workers exposed to rats. *Occupational and Environmental Medicine*, (60(2): 104–108), 2003.

[150] M.J. Nieuwenhuijsen. New developments in exposure assessment: The impact on the practice of health risk assessment and epidemiological studies. *Environment International*, (32(8): 996-1009), 2006.

[151] Welch R. M. Wu J. Norvell WA. Kriging on highly skewed data for dtpa-extractable soil zn with auxiliary information for ph and organic carbon. *Geoderma*, (134(1-2): 187-199), 2006.

[152] J. R. Nuckols and Jarup L. Ward, M.H. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives*, (112(9): 1007–1015), 2004.

[153] M. A. Oliver and R. Webster. Kriging, a method of interpolation for geographical information systems. *International Journal of Geographical Information Science*, (4: 313-332), 1990.

[154] A. D. Ord JK Cliff. *Spatial processes: models & applications*. Pion, 1981.

[155] W. R. Ott. Human exposure assessment: the birth of a new science. *Journal of Exposure Analysis and Environmental Epidemiology*, (5(4): 449-472), 1995.

[156] H. Ozkaynak. Modeling population exposures to outdoor sources of hazardous air pollutants. *Journal of Exposure Science and Environmental Epidemiology*, (18: 45–58), 2008.

[157] C. J. Paciorek, J. Schwartz, F. Laden, R. Puett, and J. D. Suh HH Yanosky. Spatio-temporal modeling of chronic pm10 exposure for the nurses' health study. *Atmospheric Environment*, (42(18): 4047-4062), 2008.

[158] V. Palmieri and R. Carvalho. Qual2e model for the corumbatai river. *Ecological Modelling*, (198: 269-275), 2006.

[159] C. Pavlik, A. Ruggles, and D. Armstrong MP Zimmerman. An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, (31(4): 375-390), 1999.

[160] D. N. Pellow. Environmental justice and the political process: movements, corporations, and the state. *The Sociological Quarterly*, (42(1): 47-67), 2001.

[161] C. Peng, M. Wang, Y. Zhao, and W. Chen. Distribution and risks of polycyclic aromatic hydrocarbons in suburban and rural soils of beijing with various land uses. *Environmental monitoring and assessment*, (188(3): 162), 2016.

[162] D. W. Pennington, M. Margni, C. Ammann, and O. Jolliet. Multimedia fate and human intake modelling: spatial versus non spatial insights for chemical emissions in western europe. *Environmental Science and Technology*, (39(4): 1119–1128), 2005.

[163] A. Pistocchi. Rethinking sustainability while pursuing it along the same paths. *Impact Assessment and Project Appraisal*, (167–168), 2011.

[164] T.S. Poet, P.M. Schlosser, C.E. Rodriguez, R.J. Parod, D.E. Rodwell, and C.R. Kirman. Using physiologically based pharmacokinetic modeling and benchmark dose methods to derive an occupational exposure limit for n-methylpyrrolidone. *Regulatory Toxicology and Pharmacology*, (76: 102 - 112), 2016.

[165] J. I. Ponce-Hernandez R Hernandez-Stefanoni. Mapping the spatial variability of plant diversity in a tropical forest: comparison of spatial interpolation methods. *Environmental Monitoring and Assessmen*, (117(1–3): 307-334), 2006.

[166] L. Pulido. Rethinking environmental racism: White privilege and urban development in southern california. *Annals of the Association of American Geographers*, (90(1): 12–40), 2000.

[167] Y. Qiu, B. Fu, J. Wang, and Chen L. Spatiotemporal prediction of soil moisture content using multiple-linear regression in a small catchment of the loess plateau, china. *CATENA*, (54(1–2): 173-195), 2003.

[168] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[169] S. M. Rappaport. Biomarkers intersect with the exposome. *Biomarkers*, (17(6): 483-489), 2012.

[170] Thomson A. Best N. Richardson S. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, (14(1): 35–59), 2005.

[171] J. Rivoirard. *Course on multivariate geostatistics*. Centre de Géostatistique, 2003.

[172] C. Ross, J. Reynolds, and K. Geis. The contingent meaning of neighborhood stability for residents' psychological well-being. *American Sociological Review*, (65(4): 581-597), 2000.

[173] L. J. S. Rossini AJ Liu. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan toronto - a pilot study. *Environment International*, (22(6): 677–692), 1996.

[174] J. L. Sadd. Every breath you take ...: The demographics of toxic air releases in southern california. *Economic Development Quarterly*, (13(2): 107–123), 1999.

[175] D. Sarigiannis, S. Karakitsios, A. Gotti, G. Loizou, J. Cherrie, R. Smolders, K. De Brouwere, K. Galea, K. Jones, E. Handakas, K. Papadaki, and A. Sleeuwenhoek. *Integra: From global scale contamination to tissue dose.* Number 2: 1001-1008. International Environmental Modelling and Software Society, 2014.

[176] D. Sarigiannis, S. Karakitsios, E. Handakas, and A. Gotti. Exposome analysis of polyaromatic hydrocarbons. *Toxicology Letters*, (258: S54–S61), 2016.

[177] C. A. Schloeder, N.E. Zimmermann, and M. J. Jakobs. Comparison of methods for interpolating soil properties using limited data. *Soil Science Society of American Journal*, (65(2): 470-479), 2001.

[178] D. Schlosberg. *Defining Environmental Justice*. Oxford, 2007.

[179] J. Schneider, L. Moosmann, and W. Nagl C Spangl. Representativeness and classification of air. Technical report, 2007.

[180] G. Serre ML Christakos. Spatiotemporal analysis of environmental exposure-health effect associations. *Journal of Exposure Science and Environmental Epidemiology*, (10: 168–187), 2000.

[181] K. Sexton. Cumulative risk assessment: An overview of methodological approaches for evaluating combined health effects from exposure to multiple environmental stressors. *International Journal of Environmental Research and Public Health*, (9(2): 370–390), 2012.

[182] G. Shaddick, J. C. Wakefield, C. De Hoogh, and P. Briggs DJ Elliott. Long-term associations of outdoor air pollution with mortality in great britain. *Thorax*, (62(12): 1088–1094), 2007.

[183] G. Shen, Y. Chen, S. Wei, X. Fu, A. Ding, H. Wu, and S. Tao. Can coronene and/or benzo(a)pyrene/coronene ratio act as unique markers for vehicle emission? *Environmental Pollution*, (184: 650 - 653), 2014.

[184] R. Shukla, Y. Li, and G. K. Lockey JE Lemasters. Balancing cost and precision in exposure assessment studies. *Journal of Occupational and Environmental Medicine*, (38(1): 39–45), 1996.

[185] E. H. Srivastava MR Isaaks. *An introduction to applied geostatistics.* Oxford University Press, 1989.

[186] D. G. Steel, M. Tranmer, and D. Wrigley N Holt. Aggregation and ecological effects in geographically based data. *Geographical Analysis*, (28(3): 244-261), 1996.

[187] Von B. Steiger, K. Nowack, and R. Schulin. Spatialvariation of urease activitymeasured in soil monitoring. *.Journal of Environmental Quality*, (25: 1285-1290), 1996.

[188] A. Stein, M. Hoogerwerf, and J. Bouma. Use of soil map delineations to improve (co-)kriging of point data on moisture deficits. *Geoderma*, (43(2–3): 163-177), 1988.

[189] P. Stretesky and M.J. Hogan. Environmental justice: An analysis of superfund sites in florida. *Social Problems*, (45(2): 268–287), 1998.

[190] H. Strosnider, C. Kennedy, M. Monti, and F. Yip. Rural and urban differences in air quality, 2008–2012, and community drinking water quality, 2010–2015- united states. *MMWR Surveillance Summit*, (66(SS-13):1–10), 2017.

[191] N. Suciu, A. Tediosi, P. Ciffroy, A. Altenpohl, C. Brochot, F. Verdonck, F. Ferrari, E. Giubilato, E. Capri, and G. Fait. Potential for merlin-expo, an advanced tool for higher tier exposure assessment, within the eu chemical legislative frameworks. (562: 474-479), 2016.

[192] J.H. Syed, M. Iqbal, G. Zhong, A. Katsoyiannis, I.C. Yadav, J. Li, and G. Zhang. Polycyclic aromatic hydrocarbons (pahs) in chinese forest soils: profile composition, spatial variations and source apportionment. (7: 2692), 2017.

[193] A. Szasz and M. Meuser. Environmental inequalities: Literature review and proposals for new directions in research and theory. *Current Sociology*, (45(3): 99–120), 1997.

[194] H. Tabari, A.A. Sabziparvar, and M. Ahmadi. Comparison of artificial neural network and multivariate linear regression methods for estimation of daily soil temperature in an arid region. *Meteorology and Atmospheric Physics*, (110(3–4): 135–142), 2011.

[195] S. Taylor P Openshaw. A million or so correlation coefficients: three experiments on the modifiable area unit problem. *Statistical Applications in the Spatial Sciences (Wrigley N, ed). London: Pion*, (127–144), 1979.

[196] O. Tchepel and M. Penedo A, Gomez. Assessment of population exposure to air pollution by benzene. *International Journal of Hygiene and Environmental Health*, (210(3-4): 407-410), 2007.

[197] J. G. Teeguarden, Y. M. Tan, S. W. Edwards, J. A. Leonard, K. A. Andersonv, R. A. Corley, M. L. Kile, S. M. Simonich, D. Stone, R. L. Tanguay, K. M. Waters, S. L. Harper, and D.E. Williams. Completing the link between exposure science and toxicology for improved environmental health decision making: The aggregate exposure pathway framework. *Environmental Science and Technology*, (50(9): 4579–4586), 2016.

[198] T. W. Tesche, R. Morris, G. Tonnesen, D. McNally, J. Boylan, and P. Brewer. Cmaq/camx annual 2002 performance evaluation over the eastern u.s. *Atmospheric Environment*, (40: 4906–4919), 2006.

[199] E. P. A. US. Final guidelines for exposure assessment. *Federal Register, 104: 22888-22938. Washington, DC, United States*, 1992.

[200] E. P. A. US. Exposure factors handbook. *Washington, DC, United States Environmental Protection Agency*, 1997.

[201] E. P. A. US. A review of the reference dose and reference concentration processes. *Risk Assessment Forum, Washington, DC*, 2003.

[202] B.L. van Drooge. *Human Exposure to Polycyclic Aromatic Hydrocarbons in Urban and Rural Ambient Air.* Number 4: 59-82. 2013.

[203] C. J. Van Leeuwen and T.G. Vermeire. *Risk assessment of chemicals: an introduction.* Springer, 2007.

[204] K. Van Veldhoven, M. Chadeau-Hyam, Athersuch T.J., and P. Vineis. Advancing the application of omics-based biomarkers in environmental epidemiology. *Special number on Application of Omics Techniques to Epidemiological Studies*, (54(7): 461–467), 2013.

[205] A. Vervoot and A. Govaerts. Geostatistical interpolation of soil properties in boom clay in flanders. *geoENV VII: Geostatistics for Environmental Applications*, (219-230), 2010.

[206] J. F. Viel, E. Fournier, and Danzon A. Age-period-cohort modelling of non-hodgkin's lymphoma incidence in a french region: a period effect compatible with an environmental exposure. *Environmental Health*, (9–47), 2010.

[207] D. Vienneau, K De Hoogh, and D. Briggs. A gis-based method for modelling air pollution exposures across europe. *Science of the Total Environment*, (408(2): 255-266), 2009.

[208] E. Villanneau, N. Saby, T. Orton, C. Jolivet, L. Boulonne, G. Caria, E. Barriuso, A. Bispo, O. Briand, and D. Arrouays. First evidence of large-scale pah trends in french soils. (11), 2013.

[209] J.L. Volatier. Enquete inca individuelle et nationale sur les consommations alimentaires. Technical report, 2000.

[210] V. Vuksanovic, F. De Smedt, and S. Van Meerbeeck. Transport of polychlorinated biphenyls (pcb) in the scheldt estuary simulated with the water quality model wasp. *Journal of Hydrology*, (174: 1-18), 1996.

[211] H. Wackernagel, C. Lajaunie, N. Blond, C. Roth, and R. Vautard. Geostatistical risk mapping with chemical transport model output and ozone station data. *Ecological Modelling*, (179(2): 177-185), 2004.

[212] G. Walker, G. Mitchell, J. Fairburn, and G. Smith. Industrial pollution and social deprivation: Evidence and complexity in evaluating and responding to environmental inequality. *Local Environment*, (10(4): 361–377), 2005.

[213] C. Wang, S. Wu, S. Zhou, H. Wang, B. Li, H. Chen, Y. Yu, and Y. Shi. Polycyclic aromatic hydrocarbons in soils from urban to rural areas in nanjing: Concentration, source, spatial distribution, and potential human health risk. *Science of The Total Environment*, (527-528: 375 - 383), 2015.

[214] H. Wang, G. Liu, and P. Gong. Use of cokriging to improve estimates of soil salt solute spatial distribution in the yellow river delta. *Acta Geographica Sinica*, (60(3): 511-518), 2005.

[215] Z. Wang, J. Chen, X. Qiao, P. Yang, F. Tian, and L Huang. Distribution and sources of polycyclic aromatic hydrocarbons from urban to rural soils: A case study in dalian, china. *Chemosphere*, (68(5): 965 - 971), 2007.

[216] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics, 2003.

[217] D. Weber and E. Englund. Evaluation and comparison of spatial interpolators. *Mathematical Geology*, (24(4): 381-391), 1992.

[218] R. Webster and M.A. Oliver. *Geostatistics for Environmental Scientists, 2nd Edition*. Wiley, 2001.

[219] M. A. Webster R Oliver. Kriging: a method of interpolation for gis. *International Journal of Geographical Information Systems*, (4(3): 313–32), 1990.

[220] M. Webster R Voltz. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Sciences*, (41(3): 473-490), 1990.

[221] B. Wheeler. Health-related environmental indices and environmental equity in england and wales. *Environment and Planning A*, (36(5): 803–822), 2004.

[222] WHO. Air quality guidelines for europe, second ed. world health organisation regional publications. *European series, No. 91*, 2000.

[223] WHO. Annual report. Technical report, 2000.

[224] WHO. ealth risks of persistent organic pollutants from long-range transboundary air pollution. *Joint WHO/Convention Task Force on the Health Aspects of Air Pollution. World Health Organization, Regional Office for Europe, Copenhagen, Denmark*, 2003.

[225] WHO. Iarc monographs on the evaluation of carcinogenic risks to humans. Technical Report 92, 2010.

[226] WHO. Health impact assessment of air pollution in megacity of tehran, iran. *Iranian Journal of Environmental Health Science & Engineering*, (9:28), 2012.

[227] C. P. Wild. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention*, (14(8): 1847-50), 2005.

[228] D. Williams and C. Collins. Us socioeconomic and racial differences in health: Patterns and explanations. *Annual Review of Sociology*, (21: 349-386), 1995.

[229] A. S. Wong, D.W.S.and Otheringham. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, (23(7): 1025-1044), 1991.

[230] J. Wright, C. L. Wienburg, H. I. J. Black, S. M. Long, D. Osborn, and E. Spurgeon DJ Heywood. Factors influencing the national distribution of polycyclic aromatic hydrocarbons and polychlorinated biphenyls in british soils. *Environment Science Technology*, (40(24): 7629–7635), 2006.

[231] Z. Xia, X. Duan, W. Qiu, D. Liu, B. Wang, S. Tao, Q. Jiang, B. Lu, Y. Song, and X. Hu. Health risk assessment on dietary exposure to polycyclic aromatic hydrocarbons (pahs) in taiyuan, china. *Science of The Total Environment*, (408(22): 5331 - 5337), 2010.

[232] Y. Xiao, F. Tong, Y. Kuang, and B. Chen. Distribution and source apportionment of polycyclic aromatic hydrocarbons (pahs) in forest soils from urban to rural areas in the pearl river delta of southern china. (11: 2642-2656), 2014.

[233] Y. Yang, N. Zhang, M. Xue, S.T. Lu, and S. Tao. Effects of soil organic matter on the development of the microbial polycyclic aromatic hydrocarbons (pahs) degradation potentials. *Environmental Pollution*, (159(2): 591 - 595), 2011.

[234] J. Zhang, P. Hua, and P. Krebs. Potential source contributions and risk assessment of particulate-associated polycyclic aromatic hydrocarbons in size-fractionated road-deposited sediments. *Water Practice and Technology*, (8(2)), 2013.

[235] D. Zimmerman. Another look at anisotropy in geostatistics. *Mathematical Geology*, (25(4): 453–470), 1993.

Annexes

Appendix A

Work Evaluation

A.1 Publications

- Caudeville, J., Ioannidou, D., Boulvert, E., & Bonnard, R. (2017). Cumulative Risk Assessment in the Lorraine Region: A Framework to Characterize Environmental Health Inequalities. International Journal of Environmental Research and Public Health, 14(3), 291. http://doi.org/10.3390/ijerph14030291

- Ioannidou, D., Malherbe, L., Beauchamp, M., Saby, N.P.A., Bonnard, R., Latouche, A. & Caudeville, J. (Submitted). Developing environmental data processing methods to map exposure indicators for Polycyclic Aromatic Hydrocarbon substances. Risk Analysis

A.2 Communications

A.2.1 ISEE 2015, Sao Paulo, Brazil

Characterizing Environmental Health Inequalities With Spatial Environmental Database, Exposure Model And Biomonitoring Measures

Abstract

Environmental health inequalities have become a major preoccupation for public health as evidenced by the emergence of the French national plans for health and environment. The third plan (2015-2019) has highlighted that the development of a methodology to

derive indicators for identifying and characterizing environmental health inequalities is a priority. At a regional scale, environmental monitoring networks are not sufficient to capture the multidimensionality of the exposure at finer resolution. In order to increase representativeness, exposure composite indicators could be built using additional available databases and statistical modeling. To tackle these objectives, the proposed methodology is to combine contaminant source and environmental concentration databases, exposure model and population biological impregnation measurements with a spatial approach. To elicit the proposed methodology we will consider a pesticide case study developed in the context of the MecoExpo study conducted in Picardy (France). Notably in this framework, topsoil concentrations are estimated using a spatial interpolation, method able to integrate topsoil concentration and pesticide deposit data. For water, concentrations are temporally interpolated with methodology that allows us to overcome the problem of under the detection limit concentrations. Then, an exposure multimedia model developed by INERIS is employed to estimate exposure dose in comparison with biological impregnations of cohort participants (neonatal meconium assays - 700 participants). Analysis' outputs enable to build cumulative exposure indicator, identify pollutant sources and determinants of exposure and vulnerable populations. As a perspective, spatial relationships of exposure indicators, biological impregnation and health data will be estimated to identify potential factors influencing variability in disease spatial pattern in the French coordinated integrated environment and health inequality platform project.

A.2.2 Spatial Statistics 2017: One World, One Health, Lancaster, UK

Developing environmental data processing methods to spatialize exposure indicators for three polycyclic aromatic hydrocarbon substances

Abstract

Analyzing the relationship between the environment and health has become a major focus of public health efforts in France, as evidenced by the national action plans for health and the environment. These plans have identified the following two priorities:

- identify and manage geographic areas where hotspot exposures are a potential risk to human health; and

- reduce exposure inequalities.

The aim of this study is to build spatial exposure indicators using a spatial stochastic

multimedia exposure model for detecting vulnerable populations and analyzing exposure determinants at a fine resolution and national scale. A multimedia exposure model was developed by INERIS to assess the transfer of substances from the environment to humans through inhalation and ingestion pathways. The PLAINE platform adds a spatial dimension by linking GIS (Geographic Information System) to the model. A spatial database needs to be constructed for the model application. Data is derived from national databases, and then spatialized with respect to their specificities. Tools are developed using modeling, spatial analysis and geostatistical methods to build and discretize interesting variables and indicators from different supports and resolutions on a 9-km2 regular grid. We applied this model to the risk assessment of exposure to PAH (Polycyclic Aromatic Hydrocarbons) using national databases. The results will permit to estimate ingestion exposure pathway contributions based on exposure scenarios defined for two different referent groups (age, dietary properties, and the fraction of food produced locally). As a perspective, spatial relationships between exposure indicators and health data can be estimated to identify potential factors influencing variability in disease spatial pattern.

Despoina IOANNIDOU

Environmental inequality characterization of Polyaromatic Hydrocarbons in France.

Résumé : La réduction des inégalités d'exposition environnementale constitue un axe majeur en santé publique en France comme en témoignent les priorités des différents Plan Nationaux Santé Environnement (PNSE). L'objectif de cette thèse est de développer une approche intégrée pour la caractérisation des inégalités environnementales et l'évaluation de l'exposition spatialisée de la population aux HAP en France.
Les données produites dans le cadre des réseaux de surveillance de la qualités des milieux environnementaux sont le reflet de la contamination réelle des milieux et de l'exposition globale des populations. Toutefois, elles ne présentent généralement pas une représentativité spatiale suffisante pour caractériser finement les expositions environnementales, ces réseaux n'ayant pas été initialement conçus dans cet objectif. Des méthodes statistiques sont développées pour traiter les bases de données d'entrée (concentrations environnementales dans l'eau, l'air et le sol) et les rendre pertinentes vis à vis des objectifs définis de caractérisation de l'exposition. Un modèle multimédia d'exposition, interfacé avec un Système d'Information Géographique pour intégrer les variables environnementales, est développé pour estimer les doses d'exposition liées à l'ingestion d'aliments, d'eau de consommation, de sol et à l'inhalation de contaminants atmosphériques. La méthodologie a été appliquée pour trois Hydrocarbures Aromatiques Polycycliques (benzo[a]pyrène, benzo[ghi]pérylène et indéno[1,2,3-cd]pyrène) sur l'ensemble du territoire français. Les résultats permettent de cartographier des indicateurs d'exposition, d'identifier les zones de surexposition et de caractériser les déterminants environnementaux. Dans une logique de caractérisation de l'exposition, la spatialisation des données issues des mesures environnementales pose un certain nombre de questions méthodologiques qui confèrent aux cartes réalisées de nombreuses incertitudes et limites relatives à l'échantillonnage et aux représentativités spatiales et temporelles des données. Celles-ci peuvent être réduites par l'acquisition de données supplémentaires et par la construction de variables prédictives des phénomènes spatiaux et temporels considérés.

Les outils de traitement statistique de données développés dans le cadre de ces travaux seront intégrés dans la plateforme PLAINE pour être déclinés sur d'autres polluants en vue de prioriser les mesures de gestion à mettre en œuvre.

Mots clés : statistiques, exposition, environnement, inégalités, modelisation spatiale, kriging, Hydrocarbure Polyaromatique

Abstract : Reducing environmental exposure inequalities has become a major focus of public health efforts in France, as evidenced by the French action plans for health and the environment. The aim of this thesis is to develop an integrated approach to characterize environmental inequalities and evaluate the spatialized exposure to PAH in France.
The data produced as part of the monitoring quality networks of environmental media reflect the actual contamination of the environment and the overall exposure of the populations. However they do not always provide an adequate spatial resolution to characterize environmental exposures as they are usually not assembled for this specific purpose. Statistical methods are employed to process input databases (environmental concentrations in water, air and soil) in the objective of characterizing the exposure. A multimedia model interfaced with a GIS, allows the integration of environmental variables in order to yield exposure doses related to ingestion of food, water and soil as well as atmospheric contaminants' inhalation.The methodology was applied to three Polycyclic Aromatic Hydrocarbon substances, (benzo[a]pyrene, benzo[ghi]perylene and indeno[1,2,3-cd]pyrene), in France. The results obtained, allowed to map exposure indicators and to identify areas of overexposure and characterize environmental determinants. In the context of exposure characterization, the direct spatialization of available data from environmental measurement datasets poses a certain number of methodological questions which lead to uncertainties related to the sampling and the spatial and temporal representativeness of data. These could be reduced by acquiring additional data or by constructing predictive variables for the spatial and temporal phenomena considered.

Data processing algorithms and calculation of exposure carried out in this work, will be integrated in the French coordinated integrated environment and health platform-PLAINE in order to be applied on other pollutants and prioritize preventative actions.

Keywords : statistics, exposure, environment, inequalities, spatial modeling, kriging, Polyaromatic Hydrocarbons