Econométrie des données imparfaites : méthodes et applications

Enora Belz

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1

ECOLE DOCTORALE N° 597
Sciences Economiques et sciences De Gestion
Spécialité : Sciences Economiques

Par
Enora BELZ

Économétrie des données imparfaites : méthodes et applications

Thèse présentée et soutenue à Rennes, le 21 juin 2021
Unité de recherche : CNRS UMR6211 CREM

Rapporteurs avant soutenance :

Emmanuel Flachaire	Professeur, Aix-Marseille Université
Montserrat Guillén	Professeure, Universitat de Barcelona

Composition du Jury :

Président :	Olivier L'Haridon	Professeur, Université de Rennes 1
Examinateurs :	Emmanuel Flachaire	Professeur, Aix-Marseille Université
	Montserrat Guillén	Professeure, Universitat de Barcelona
	Brice Magdalou	Professeur, Université de Montpellier
Dir. de thèse :	Arthur Charpentier	Professeur, Université de Rennes 1

Remerciements

Ces années de thèses constituent pour moi une expérience inoubliable, tant humaine qu'intellectuelle. Quatre années au cours desquelles j'ai pu découvrir le métier de chercheur, rencontrer de nouvelles personnes, mais surtout, reçu le soutien indispensable d'un très grand nombre de personnes sans qui ce travail de thèse n'aurait pu aboutir. Ces quelques paragraphes leur sont dédiés et j'espère, à travers ces quelques lignes, leur transmettre ma profonde reconnaissance.

La première personne que je tiens à remercier est Arthur Charpentier, mon directeur de thèse. Merci à vous, de m'avoir fait confiance et d'avoir accepté d'encadrer mon travail. Vous avez su m'accompagner durant ces quatre années de près (puis de loin). Aucun de mes questionnements et interrogations n'est resté sans réponse et vous avez toujours su être disponible pour m'aider à y répondre. Merci, Arthur, pour vos conseils, votre disponibilité, votre soutien et de m'avoir permis de découvrir Montréal et le Canada.

Merci aux membres de mon Jury de Thèse, Emmanuel Flachaire, Montserrat Guillén, Olivier L'Haridon et Brice Magdalou, d'avoir accepté d'en faire partie. C'est un honneur de présenter mon travail devant vous.

Je remercie, Olivier L'Haridon et Benoît Tarroux, d'avoir formé mon comité de suivi individuel. Vous avez vu évoluer mon travail et m'avez à chaque fois fourni des remarques et des conseils pertinents pour le bon déroulement de la thèse.

Merci à la Région Bretagne et la Chaire Act'info d'avoir financé ce projet de thèse.

Je tiens également à remercier l'École Doctorale EDGE, notamment son directeur, Éric Malin et son directeur passé, Thierry Pénard pour leur attention portée aux doctorants. Je remercie Hélène Jean, sa gestionnaire, pour son efficacité et sa bienveillance.

Merci au CREM de m'avoir accueilli au sein de son laboratoire. Je tiens à remercier son directeur, Franck Moraux et son directeur adjoint, David Masclet pour leur disponibilité et leur soutien. Le CREM m'a permis de travailler dans de bonnes conditions, dans une atmosphère inspirante et stimulante, à travers les conférences, les séminaires, les cafés CREM (et e-c@fés CREM). Je remercie les chercheurs et collègues du CREM avec qui j'ai pu échanger avec bienveillance de mes projets de recherche. Merci également aux personnels administratifs : Hèle Coda-Poirey, Lucie Germain, Naïla Louise-Rose et Anne L'Azou. Un merci particulier à Cécile Madoulet pour sa présence, son aide et son soutien.

J'aimerais aussi remercier l'Université de Rennes 1 et, plus particulièrement la Faculté de Sciences Économiques. Elle m'a vu grandir durant neuf années, en licence et master, puis en doctorat et enfin en contrat d'ATER. Je remercie Auréline Fargeas, Chantal Gueguen, Thierry Karcher et Jean-Christophe Poutineau pour les différents travaux dirigés que j'ai pu donner avec eux. Je remercie également l'équipe administrative, en particulier, Caroline Lemoine, Soizic Masson, Françoise Mazzolini, Géraldine Perrodin et Marina Siraudeau.

Je tiens également à remercier les participants des différents séminaires, conférences et workshops auxquels j'ai participé pour leurs conseils et leurs retours toujours très pertinents. Ainsi, merci aux participants des écoles thématiques TEPP-CNRS de 2018 et 2019, de la journée des chaires "Chairs Days : Insurance, Acturial Science Data and Models" de 2018, du Congrès Annuel de la SCSE au Québec en 2019, de R à Québec 2019 et des Déjeuners du Dommage Corporel.

Mes remerciements vont également à l'association PROJECT et aux collègues doctorant(e)s et post-doctorant(e)s que j'ai côtoyé au cours de ces quatre dernières années. Merci à Amaury, Étienne, Ewen, Joao, Matthieu, Nassay, Nathalie et Roberto, avec qui j'ai partagé le bureau 309. Je remercie, bien entendu, également les doctorants de l'"autre côté" et de SMART-LERECO, notamment, Alejandra, Esther, Jacques, Jimmy, Louise, Lucile, Madeg, Maëva, Martina, May, Romain, Sebastian, Thao, Thibaut, Vincent et Xuan.

Je souhaite adresser mes remerciements à toutes celles et tous ceux qui ont été présents à mes côtés durant ces années en m'accompagnant et m'épaulant au quotidien. Merci les amis. Les personnes extraordinaires que j'ai rencontré lors de cette thèse, merci à Alejandra, Martina,

Sebastian, Roberto et Thao pour la découverte de votre culture, merci à Ewen, Esther, Jimmy, Maëva, Romain et Vincent pour les moments inoubliables. Merci à toi, Nathalie, d'avoir toujours été présente pour moi, et pour nos discussions et rires souvent interminables. Je voudrais également souligner le soutien indéfinissable de tous mes amis. Merci à Émilie, d'avoir fait qui je suis aujourd'hui. Merci à Antoine, Camille, Charles, Florian et Tifenn. Merci à Antoine, Aurélie, Claire, Corinne, Damien, Élise, Étienne et Maxime. Merci à Marième, Océane et Pauline. Merci à Safiah, Thomas et Joris. Merci à Amélie, Gilone et Myriam.

Enfin, je veux remercier du fond du cœur ma famille pour leur soutien sans faille et leurs encouragements constants. Vous êtes mon *trèfle à quatre feuilles*. Merci à mes parents Didier et Brigitte, ma sœur Audrey, mon frère Esteban, mon beau-frère Aurélien, mes grands-mères, mes oncles et tantes et mes cousins. Merci également à ma belle-famille, mes beaux-parents Daniel et Sophie, mon beau-frère Romain et ma belle-sœur Caroline. En dernier lieu, comment faire des remerciements sans te remercier toi, Corentin. Sans toi, rien n'aurait pu être possible. Tu as su, au quotidien, être présent par tes mots, tes conseils et tes blagues. Tu as été pour moi un repère durant ce chemin de thèse.

À vous tous, merci.

Résumé en français[1]

1 Motivation

Née dans les années 1930, l'économétrie vise à combler le fossé entre les modèles économiques et les données économiques. Les modèles statistiques sont initialement créés pour modéliser des données indépendantes et identiquement distribuées. Mais, dans la nature, rares sont les données parfaites et les données sont soumises à diverses perturbations. Les économètres ont, au fil des années, proposés des méthodes statistiques appropriées aux diverses données rencontrées afin d'estimer des relations économiques, de tester des théories économiques ou d'évaluer le lien entre des variables économiques (Wooldridge et al., 2018).

D'après Griliches (1985), les données et les économètres représentent une alliance fragile. Selon ses propres mots :

> Les économètres ont une attitude ambivalente face aux données économiques. À un premier niveau, les "données" sont le monde que nous voulons expliquer, les faits de base que les économistes prétendent élucider. À un second niveau, elles sont la source de tous nos problèmes.[2]

[1] Les références du résumé sont présentes à la page 15.

[2] La citation en anglais dans le papier est la suivante : "Econometricians have an ambivalent attitude towards economic data. At one level, the "data" are the world that we want to explain, the basic facts that economists purport to elucidate. At the other level, they are the source of all our troubles."

Dans cet extrait, il explique l'ambiguïté face aux données qui sont à la fois le problème et la solution. Les données ne sont pas parfaites et pas toujours adaptées au problème posé. La principale raison réside dans la distance entre les économistes et les "producteurs" de données. Les données sont récoltées en majorité par des organismes et instituts de sondages. Les économistes, et de façon plus générale en sciences sociales, les utilisent *a posteriori* pour traiter une question. Cette question n'est pas nécessairement envisagée *a priori* à la création de l'enquête. Dans certains cas, les données ne coïncident pas alors entièrement avec la question de recherche. Néanmoins, si les données étaient parfaites, l'économétrie n'existerait pas (Griliches, 1985). La présence d'imperfections dans les données rend l'analyse plus difficile, voire impossible dans certains cas. L'imperfection est donc un des enjeux clés de l'économétrie. Les recherches sur les problèmes des données imparfaites, qu'elles soient imprécises, ambiguës ou incomplètes sont nombreuses.

Les données rencontrées diffèrent par leur nature et leur spécificité. En premier lieu, il existe une disparité entre les individus statistiques étudiés. On parlera notamment de séries temporelles, de coupe transversale ou de données de panel. Dans la définition même de ces catégories, il existe de multiples formes d'individus. En termes de séries chronologiques, la temporalité peut varier d'une année pour les modèles macroéconomiques à quelques secondes pour les modèles financiers. Pour les données transversales, les individus étudiés peuvent être des ménages, des entreprises, mais aussi des pays ou des villes.

Par ailleurs, la qualité de l'information rencontrée varie d'une étude à une autre. Les données sont sujettes à des altérations. Elles peuvent être incomplètes du fait de problème de données manquantes, de troncature ou de censure. Elles sont parfois imprécises en raison d'erreurs de mesure ou lorsqu'elles sont agrégées. En effet, une donnée agrégée ne donne qu'une information diluée de la réalité. Le choix des catégories de discrétisation est susceptible de générer des irrégularités. Un exemple simple est la disparité dans le choix des classes d'âge (Colvez and Villebrun, 2003). Les seuils retenus pour catégoriser les individus ne sont pas harmonisés et homogènes. La classification peut être établie par seuils sociaux : jusqu'à 16 ans, l'âge scolaire, de 16 à 25 ans, l'âge étudiant, de 26 à 64 ans, la vie active et après 65 ans, la retraite. Les individus peuvent également être classifiés par dizaines ou par vingtaines (Figure 1). Cette décision est subjective, tant en termes de nombre de catégories que de choix des âges pivots. La transition entre les catégories est arbitraire. À quel âge, sommes-nous considérés comme "jeunes" ou "vieux" ? L'évolution de la scolarité ou de l'âge de départ à la retraite entraînent également des variations dans la sélection des limites d'âge. Ces différents problèmes sont à prendre en

compte dans la modélisation et dans l'analyse.

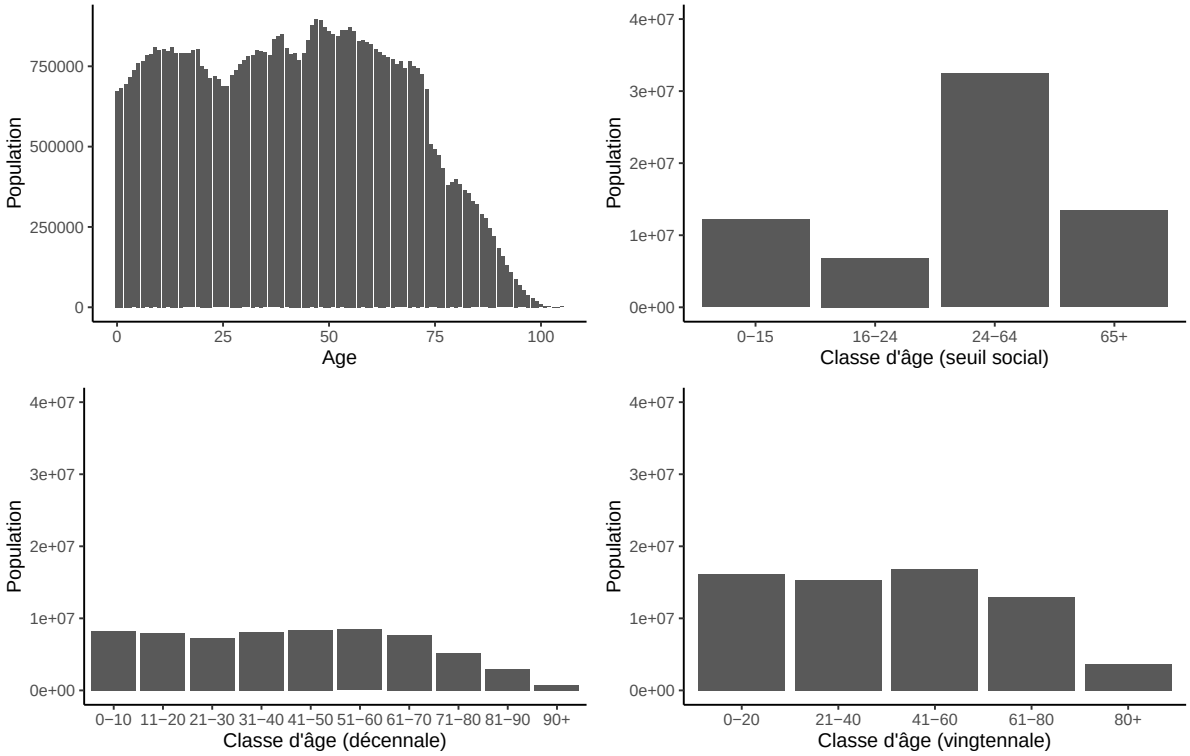


Figure 1. Population totale au 1er janvier 2020, France métropolitaine (source : INSEE)

L'économétrie est en lien avec l'économie et permet de formaliser et styliser un phénomène économique de la réalité. L'économie évolue, par conséquent, l'économétrie doit s'adapter et proposer de nouvelles méthodes statistiques adéquates avec les phénomènes économiques mais aussi avec la diversité des données. Au fil du temps, les économètres font face à de nouveaux problèmes liés à l'évolution des données. À titre d'exemple, aujourd'hui, avec l'émergence du Big Data et les avancées technologiques importantes, un nouveau défi consiste à trouver des méthodes pertinentes de manipulation et d'analyse des données massives. En parallèle, les différentes réformes du droit sur les données personnelles rendent difficile l'accès aux données individuelles. Que ce soit pour des questions de facilités de stockage, de collecte ou de besoin d'anonymat, l'agrégation des données est souvent prédominante et les données agrégées sont disponibles sous n'importe quelles circonstances et éliminent plusieurs problèmes de mesures des enquêtes (Shively, 1969).

L'étude des inégalités, et principalement le revenu, est rendue complexe par la prédominance de données agrégées. Le revenu des individus est une donnée jugée sensible, comme dans une plus grande mesure la santé, l'origine ethnique ou encore les opinions politiques ou convictions religieuses. Ces restrictions sur les données personnelles contraignent les autorités à ne fournir seulement des données agrégées à la place de données individuelles. Ces conditions sur la divul-

gation des données individuelles compliquent l'étude et l'analyse des mesures d'inégalités et du revenu. L'agrégation dominante, notamment pour des données de recensement, est l'agrégation spatiale. La disponibilité des données dépend, en outre, de la taille de l'unité géographique visée. Il existe plusieurs niveaux d'agrégation allant des pays à des zones de 200 mètres de côté. À grande échelle, les données peuvent être plus détaillées qu'au niveau du quartier car elles concernent un plus grand nombre d'individus. De surcroît, les limites de ces unités géographiques posent un problème d'agrégation spatiale (MAUP - Modifiable Areal Unit Problem) (Openshaw, 1984a,b, Openshaw and Taylor, 1979). Le découpage spatial choisi (ou imposé) influence les diverses résultats des analyses. Le MAUP couvre deux effets : l'effet de zonage et l'effet d'échelle. L'effet d'échelle correspond à l'impact de la modification du niveaux d'agrégation. L'effet de zonage se réfère lui à l'impact des délimitations choisies à une même échelle. À titre d'exemple, si l'on s'intéresse au nombre de ménages propriétaires à Paris selon plusieurs niveaux de découpage, les conclusions peuvent être diverses. La Figure 2 représente un découpage par iris à Paris. La Figure 3 quant à elle représente l'information par carreaux de 200 mètres de côté. Dans ces deux cas, la finesse du découpage est approximativement la même mais les limites des polygones diffèrent, on parlera ici d'effet de zonage. La Figure 4 correspond à la même variable mais par arrondissements. Si l'on compare à la répartition par iris, il y a un effet d'échelle car le niveau de découpage est supérieur.

Figure 2. Proportion de ménages propriétaires par iris (Paris, 2014)

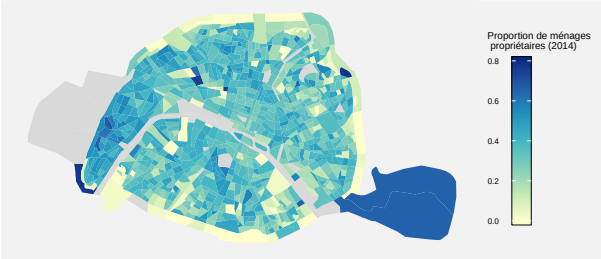
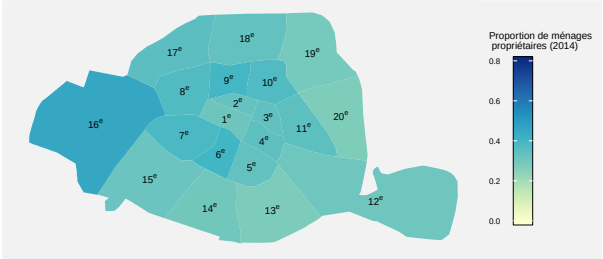


Figure 4. Proportion de ménages propriétaires par arrondissements (Paris, 2014)



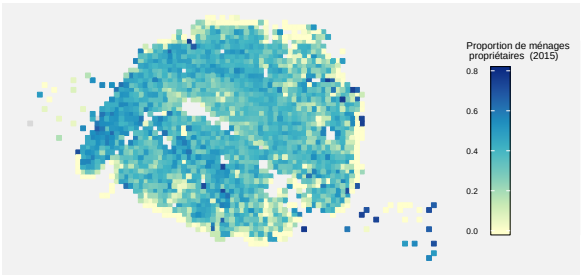


Figure 3. Proportion de ménages propriétaires par carreaux de 200 mètres de côté (Paris, 2015)

Par ailleurs, en Sciences Économiques, nombreuses variables sont également catégorielles, de

comptage, censurées ou encore tronquées. Il est alors nécessaire de sélectionner le modèle le plus adapté afin d'éviter des conséquences indésirables (biais, inefficience, incohérence) (Orme and Buehler, 2001). Dans son livre, Long (1997) mène une revue des modèles utilisables pour chacune de ces variables qu'il nomme "variables catégorielles et dépendantes limitées". L'information que contiennent ces variables est par définition limitée, qu'elles soient catégorielles (binaires, ordinales ou multinomiales), de comptage ou limitées (censurées ou tronquées). Les contraintes affectant ces variables découlent de plusieurs facteurs, entre autres, la confidentialité, la durée des enquêtes, le choix des individus ou des instituts ou encore la nature inhérente aux données peuvent altérer la nature des données. Les instituts statistiques peuvent retranscrire des données continues en variable catégorielle par choix ou nécessité de confidentialité. Cette retranscription provient en amont par catégorisation de la question posée durant l'enquête (catégories de revenu) ou par la suite par discrétisation d'une variable continue (création de catégories d'âge). Le choix des individus produit lui des données tronquées ou censurées (choix de ne pas répondre à une enquête ou question). La troncature peut également provenir du choix des organismes de sondage (choix d'interroger seulement une partie de la population) ou d'un élément de fait. Enfin, certaines données sont par nature qualitative et donc non mesurables (choix d'un moyen de transports, couleurs des yeux).

L'objectif de cette thèse est de proposer et d'étudier des méthodologies adaptées aux données imparfaites. Cette thèse abordera deux types de données imparfaites : les variables agrégées et les variables dépendantes limitées. Dans de nombreux phénomènes économiques, les données collectées amènent des problèmes pratiques. Les éventuelles imperfections doivent être examinées lors de l'utilisation des données, de la modélisation ou de l'interprétation des résultats. Nous nous intéressons ici à deux phénomènes économiques : premièrement, les mesures d'inégalités et, deuxièmement, l'indemnisation des dommages corporels. Dans chaque contexte, un ensemble de méthodes pertinentes est fourni pour appréhender les données disponibles et expliquer les phénomènes.

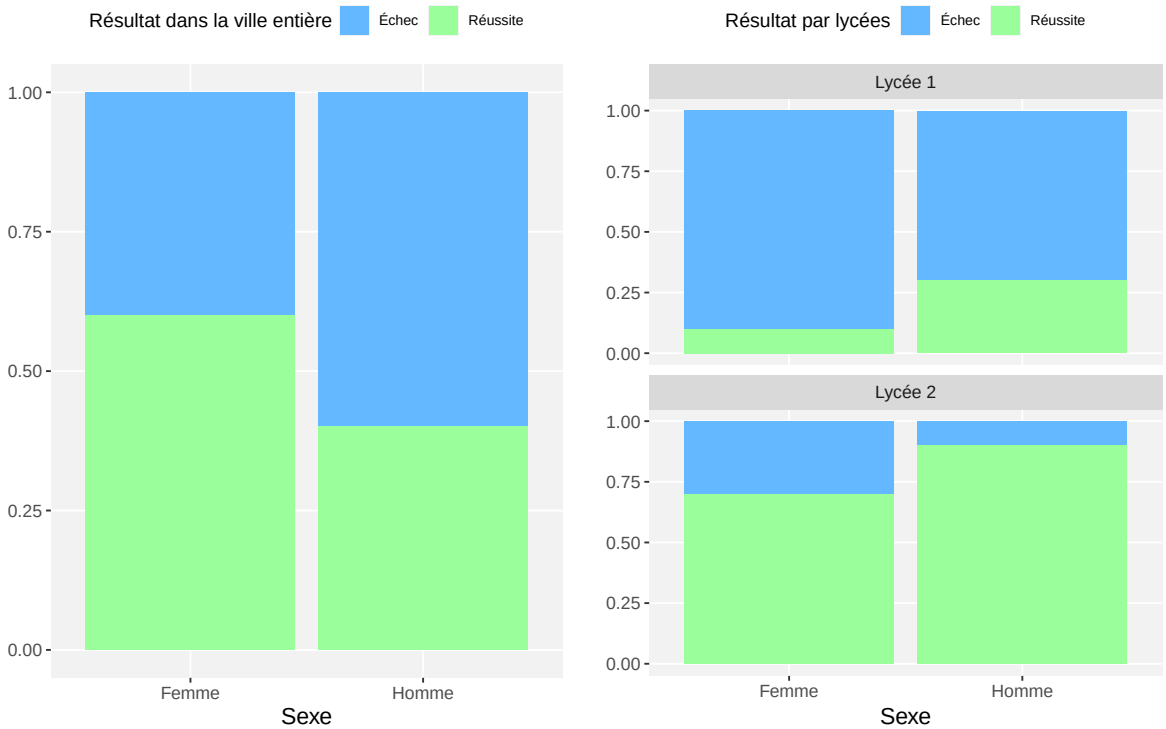
La section 2 décrit les principes de l'erreur écologique et des données compositionnelles, et apporte un aperçu des solutions envisagées dans la littérature. La section 3 présente les variables dépendantes limitées, et revient sur le concept de troncature, de censure et de biais de sélection en détaillant les modèles économétriques appropriés. Enfin, la section 4 décrit le plan de thèse et propose un bref résumé des trois chapitres associés.

2 Données agrégées

Les données agrégées sont obtenues à partir de données individuelles en groupant selon un lieu géographique, une durée ou une caractéristique et peuvent être de plusieurs natures : nombre d'individus, moyennes, quantiles, etc. Une mesure centrale est calculée en regroupant ou classant les données individuelles garantissant la préservation de l'identité des individus.

Nombreux chercheurs ont étudiés la perte d'information liée à la substitution des données individuelles par des données agrégées. Robinson (1950) affirme que les corrélations agrégées (dites *écologiques*) ne peuvent être utilisées comme substituts aux corrélations individuelles. Il met en évidence cette affirmation en étudiant le taux d'illettrisme et l'origine nationale. À l'échelle agrégée, il observe une corrélation négative entre l'illettrisme et l'immigration, plus un État compte d'immigrés, plus le taux d'illettrisme est faible. Néanmoins, à l'échelle individuelle, il constate qu'un immigré est, en moyenne, plus illettré. Ce phénomène, appelé *Ecological Fallacy*, est liée au paradoxe de Simpson. Ce paradoxe évoque la confusion entre des corrélations agrégées et décomposées en groupe. On constate une différence entre les résultats établies en synthétisant sur toute la population et les résultats établies en segmentant avec d'autres critères. La figure 5 décrit le paradoxe de Simpson en se basant sur l'exemple proposé par Rouanet et al. (2002). Cette exemple illustre la réussite scolaire au baccalauréat selon le sexe. Considérons une ville où se trouve deux lycées. À l'échelle de la ville entière, les femmes ont une réussite plus élevée (60% contre 40% pour les hommes). Mais, à l'échelle des lycées, les hommes ont une réussite plus élevée dans les deux cas (30% contre 10% dans le lycée 1 et 90% contre 70% pour le lycée 2). L'effet du sexe sur la réussite est contradictoire entre le niveau global et le niveau conditionnel. Ce constat provient d'un biais de variable omise (ici, le lycée) fortement corrélée à l'exposition (ici, le sexe).

L'erreur écologique a suscité un vive intérêt en écologie, sociologie, sciences politiques mais relativement peu en économie et de nombreuses méthodes ont été proposées afin de pouvoir inférer un comportement individuel à partir de ces données *écologiques*. Duncan and Davis (1953) suggèrent une méthode afin des borner les corrélations individuelles à partir de données agrégées et Goodman (1953) propose une régression en supposant que les coefficients ne dépendent pas de la zone. Freedman et al. (1991) présente lui aussi une alternative avec son *neighborhood model*. Il suppose qu'au sein d'une zone, il n'y a pas de différence systématique de résultats entre deux groupes. King (1997) propose une avancé dans le problème écologique en combinant



Ville	Réussite	Échec
Homme	24	36
Femme	36	24

Lycée 1	Réussite	Échec
Homme	15	35
Femme	1	9

Lycée 2	Réussite	Échec
Homme	9	1
Femme	35	15

Figure 5. Paradoxe de Simpson (Paradoxe des lycées, Rouanet et al. (2002))

l'information de la régression de Goodman et celle de la méthode des bornes. De nombreux auteurs cherchent également à mesurer ce biais d'agrégation dans l'analyse de régression et de corrélation. Clark and Avery (1976) s'intéressent à l'occurrence de ce biais sur la régression dans le cas de l'agrégation par proximité spatiale. Orcutt et al. (1968) montre que ce biais dépend de quatre phénomènes : (i) la nature sous-jacente de la réalité, (ii) le type de données agrégées, (iii) la relation entre les données et la réalité et (iv) le choix des procédures statistiques.

L'utilisation de données agrégées, afin de mesurer un comportement individuel ou agrégé, rend difficile l'inférence. L'anonymisation ou la pseudonymisation des données personnelles est aujourd'hui indispensable, notamment afin de respecter les différentes réformes des données personnelles. La pseudonymisation est réversible contrairement à l'anonymisation. L'anonymisation des données peut être effectuée selon plusieurs techniques : i) la randomisation qui consiste à altérer les données pour les rendre non identifiantes et ii) la généralisation qui consiste à diluer les données de telle sorte qu'elles ne soient plus spécifiques à un individu.

Pour de nombreuses études, les données mises à disposition sont seulement agrégées. L'idée n'est pas forcément d'utiliser ces données agrégées pour mesurer un comportement individuel mais aussi de les inclure dans des études à échelle agrégée de manière appropriée. Un problème

intéressant est l'utilisation des variables factorielles ou catégorielles qui deviennent à l'échelle agrégée des variables compositionnelles. En effet, si une variable est catégorielle au niveau individuel, en agrégeant, on obtient les différentes proportions par classe. Par exemple, lorsque l'on s'intéresse au niveau d'études, à l'échelle individuelle, on connait le niveau d'étude de chaque individu mais lorsque l'on agrège, on se retrouve avec les proportions des différents niveaux d'études dans la zone géographique ou plus largement zone d'agrégation.

Définition 1 *Une variable aléatoire compositionnelle x à D parts est un vecteur de dimension $D \times 1$ avec des composantes positives $x_1, \ldots, x_D$ sommant à 1. (Aitchison, 1986)*

Les données compositionnelles sont classiques en économie, et plus particulièrement l'étude des parts (notion de "shares"), initiée par Woodland (1979) et complétée par Ronning (1992). Récemment, Morais (2017) propose une analyse de séries chronologies compositionnelles et Fry (2011) un panorama d'applications des données compositionnelles en économie.

Les variables compositionnelles ne fournissent qu'une information de manière relative et sont limitées au simplexe. Les éléments d'une composition ne sont pas libres de varier indépendamment les uns des autres et ne sont pas libres de se situer en dehors de 0 et 1. Aitchison (1986) et ses travaux sur la géométrie du simplexe[3] introduit les opérations standards dans le simplexe. Il propose l'utilisation de log-ratios (la transformation additive et la transformation centrée) pour ramener les compositions dans l'espace réel et analyser convenablement les compositions. Egozcue et al. (2003) ont cherché une transformation alternative et proposé la transformation isométrique du log-ratio. Ces différents transformations permettent de travailler en coordonnées et d'appliquer les méthodes statistiques usuelles en tenant compte de l'information relative d'une composition (Van den Boogaart and Tolosana-Delgado, 2013).

3 Variables dépendantes limitées

Les variables dépendantes limitées sont aussi très répandues en économie (Maddala, 1986). Wooldridge (2002) définit une variable dépendante limitée comme étant une variable "dont la portée est limitée d'une manière importante"[4]. De nombreux auteurs ont fournit une revue des

[3] Le simplexe de dimension D est l'ensemble défini par $\tilde{\mathcal{S}}_D = \{(x_1, \ldots, x_D) : x_1 > 0, \ldots, x_D > 0; x_1 + \cdots + x_D = 1\}$ (Aitchison, 1986).

[4] La citation en anglais dans le papier est la suivante : "Roughly, a limited dependent variable is a variable whose range is restricted in some important way."

différents modèles de régression appropriés aux données *catégorielles et dépendantes limitées*, tels que Long (1997) et DeMaris (2004). Greene (2005) complète cette théorie en se centrant sur les variables dépendantes censurées et les distributions sous-jacentes tronquées. Les variables dépendantes limitées regroupent les variables discrètes, censurées et tronquées.

Les économistes sont souvent confrontés à des données de choix entre plusieurs alternatives. Les réponses entraînent alors des données discrètes binaires (emploi/non-emploi d'un individu), ordinales (satisfaction d'un individu) ou encore multinomiales (moyen de transport d'un individu). Parmi les variables discrètes dépendantes limitées, il existe également les données de comptage (nombre d'événements). Les variables étant discrètes, elles ne peuvent être traitées comme continues. Des modèles statistiques ont alors été développés afin de les étudier : les modèles linéaire généralisés (GLM) tels que les modèles de régression à choix binaire (modèle probit ou logit), à choix ordonné (modèle probit ou logit ordonnée), à choix multinomial (modèle logit multinomial) ou de comptage (modèle de Poisson ou binomial négatif).

Les variables limitées comprennent également les problèmes de censure et troncature. Comme le note Greene (2005), il est important de distinguer le processus de censure et de troncature. Les données tronquées proviennent d'une caractéristique de la population étudiée et de la méthode de collecte, on observe seulement une partie de la population dans l'échantillon et la perte d'information est donc complète.

Définition 2 *Une variable aléatoire x est tronquée à gauche (respectivement à droite) si x n'est observable seulement si $x > T$ (resp. $x < T$), où T est une variable aléatoire de troncature.*

Au contraire, les données censurées proviennent d'une caractéristique du processus de collecte ou d'un élément naturel (durée des enquêtes, perte de vue). Pour une partie de l'échantillon, on ne connait pas avec précision la valeur de la variable.

Définition 3 *Une variable aléatoire x est censurée à gauche (respectivement à droite) si on observe x lorsque $x \geq C$ (resp. $x \leq C$) et on observe seulement C si $x < C$ (resp. $x > C$) , où C est une variable aléatoire de censure.*

De nombreuses variables économiques sont limitées mais un traitement n'est pas forcément nécessaire (Wooldridge, 2002). Les variables telles que le salaire ou le prix d'un bien sont positives mais continues sur leur intervalle. D'autres variables entraînent des solutions en coins liées à un

processus d'optimisation telle que la consommation d'un bien où une fraction de l'échantillon aura une consommation égale à zero. Dans ce cas-là, la consommation d'un bien est continue sur les valeurs positives mais a une masse de probabilité non nulle en un point de la distribution (en l'occurrence ici en zéro) et un traitement est nécessaire afin de prendre en compte cette censure en zéro. Par ailleurs, la censure peut aussi apparaître pour des problèmes d'observations des données. À titre d'exemple, la demande de billets pour un spectacle est censurée si le spectacle est complet. La demande observée est le nombre de places vendues (donc la capacité maximale) mais la demande sous-jacente est peut-être supérieure.

Le modèle Tobit simple est utilisé pour les données censurées et les solutions en coins. En 1958, Tobin propose un modèle d'estimation par maximum de vraisemblance afin d'étudier les dépenses des ménages en biens durables et de prendre en compte la non-négativité des dépenses. Goldberger et al. (1964) donnera, par la suite, le nom de Tobit faisant référence à la proximité avec le modèle Probit. Amemiya (1973) met en évidence la cohérence et la normalité asymptotique de l'estimateur du maximum de vraisemblance. Pour les données tronquées, le modèle de régression tronquée est utilisé (Greene, 2005).

La censure est également présente lorsque l'on s'intéresse à des données de survie. On peut observer des données censurées en médecine comme pour le test d'un médicament ou une étude de cohorte, en assurance sur le temps de survenu d'un sinistre ou dans le monde industriel en étudiant la panne ou le taux de défaillance d'une machine. Dans ces cas-là, les données recueillies le sont souvent de manière partielle à cause de la censure ou troncature (Greene, 2005). Pour ce type de données, des modèles proches mathématiquement du modèle Tobit appelés modèles de survie seront privilégiés (Buckley and James, 1979, Cox, 1972, Koul et al., 1981, Miller and Halpern, 1982, Ritov, 1990).

Le modèle Tobit s'intéresse particulièrement à la censure exogène. Néanmoins, la censure ou la troncature peuvent également provenir d'un processus endogène. Les informations disponibles résultent directement des décisions prises par les individus eux mêmes et les individus s'"auto-sélectionnent". Prenons l'exemple des dommages corporels, l'information disponible sur les compensations des individus quelle soit sur le montant à l'amiable ou le montant au procès découle directement de leur décision d'aller ou non au procès. La troncature est alors liée à une autre variable auxiliaire. On parlera généralement de troncature accidentelle.

Pour répondre à ce problème, Heckman (1976) et Heckman (1979) propose un modèle de

sélection, nommé Tobit II ou Heckit, pour étudier l'offre de travail et le salaire des femmes. En effet, on observe le salaire de marché uniquement pour les femmes qui travaillent. La connaissance de leur salaire découle de leur décision d'entrer ou non sur le marché du travail. Il considère un modèle à deux équations prenant en compte le biais de sélection de l'échantillon. L'idée est de prendre en compte ce biais en introduisant une variable explicative supplémentaire : l'inverse du ratio de Mills. Une estimation en deux étapes y est proposé avec un modèle Probit permettant de déterminer le ratio de Mills puis un modèle linéaire en incluant l'inverse du ratio de Mills en variable explicative. Lee (1978) propose un modèle similaire afin d'étudier le salaire des travailleurs dans le cas où ils adhèrent ou non au syndicat. Amemiya (1984) classe les diverses modèles liés au modèle Tobit en cinq types de base selon la forme de la fonction de vraisemblance.

4 Plan de thèse

Cette thèse se constitue de trois chapitres sur l'utilisation et l'application des méthodes économétriques et statistiques lorsque les données s'avèrent incomplètes. Elle explore le problème d'inférence et d'estimation en présence de deux types de données imparfaites : les données agrégées et les données censurées. Dans le cadre des données agrégées, elle fournit une étude des variables compositionnelles liée à l'agrégation des données catégorielles (Chapitre I) et une méthode d'estimation des mesures d'inégalités à partir des données quantiles (Chapitre II). Dans le cadre des données censurées, elle propose une méthode d'analyse des montants d'indemnisation pour les préjudices corporels (Chapitre III).

Le **premier chapitre**, intitulé "Données agrégées et variables compositionnelles", examine l'utilisation des données agrégées dans un cadre statistique, en particulier, les données catégorielles qui deviennent des variables compositionnelles. Dans l'analyse des données, les données de composition sont couramment utilisées dans plusieurs domaines comme la géologie, l'économie ou le domaine social. La réforme du droit sur les données personnelles en Europe a notamment rendu difficile l'accès aux données individuelles surtout quand on cherche à étudier des données jugées sensibles. La solution souvent envisagée est la mise à disposition de données agrégées spatialement. Les données compositionnelles sont des vecteurs représentant les différentes parties d'un ensemble et sont soumis à une contrainte unitaire. Ce type de données apparaît généralement lorsqu'elles sont normalisées comme la composition de chaque minéral dans une roche, la proportion des groupes d'âge dans une ville ou les données de recensement.

L'objectif est de développer des méthodes qui permettent d'appliquer les procédures statistiques classiques sur ces données. Dans ce chapitre, l'analyse de régression compositionnelle est étudiée en s'intéressant particulièrement au cas des compositions en variables explicatives. Les méthodes classiques s'avèrent incohérentes avec les compositions qui, en raison de l'information relative qu'elles contiennent, conduisent à une interprétation trompeuse des coefficients. Cette interprétation fallacieuse est notamment induite par la clause *ceteris paribus* et par des problèmes de multicolinéarité. La solution proposée ici afin d'appliquer les méthodes de régression sur des covariables compositionnelles est de travailler en coordonnées. Une application sur le revenu médian à Paris en fonction des niveaux de diplômes et tailles de logement y est présentée.

Le **deuxième chapitre**, intitulé "Estimating Inequality Measures from Quantile Data" ("Estimation des mesures d'inégalité à partir de données quantiles"), se concentre sur le problème d'utilisation des données agrégées dans le cadre de mesures d'inégalités. Une approche appropriée et largement utilisée pour décrire les inégalités économiques consiste à fournir des indices permettant de mesurer le degré d'inégalité. Ces mesures permettent de comparer les niveaux de vie entre différents pays, régions ou dans le temps. Néanmoins, les données de revenus au niveau individuel sont rarement disponibles, notamment lorsque l'on s'intéresse à de petites zones géographiques. En raison de restrictions en matière de vie privée et de confidentialité, les données individuelles ou sur la part des revenus ne sont souvent pas diffusées et seuls des quantiles sont reportés. L'objectif de ce chapitre est de déterminer une méthodologie pour estimer une courbe de Lorenz et les indices associés avec uniquement des données quantiles. Les méthodes doivent ainsi être ajustées afin d'obtenir une courbe de Lorenz à partir de ces quantiles.

Ce deuxième chapitre propose une méthode innovante pour modéliser les courbes de Lorenz et estimer les indices d'inégalité sur de petites populations, lorsque seulement des quantiles sont disponibles. La méthode est basée sur l'espérance conditionnelle afin de trouver les différentes parts de revenu et ainsi modéliser une courbe de Lorenz avec les formes fonctionnelles déjà proposées dans la littérature. Les indices d'inégalité peuvent ensuite être dérivés à l'aide de cette courbe de Lorenz (indices de Gini, Pietra, Theil). Des simulations sont réalisées pour évaluer cette méthode et la comparer aux autres méthodes utilisées. Un exemple basé sur des données parisiennes réelles de revenus est également présenté pour illustrer la méthode.

Le **troisième chapitre**, intitulé "Bodily Injury Claims in France : Negotiation or Court ?" ("La réparation des dommages corporels en France : négociation ou tribunal ?"), examine la procédure d'indemnisation des dommages corporels en France. Le système français de compen-

sation, comme dans un très grand nombre de pays développés, est un régime sans égard à la faute. Ces systèmes sont basés sur l'indemnisation de la victime sans faute et un assureur se substitue au défendeur. En France, la loi du 5 juillet 1985 (Loi *Badinter*) vise à améliorer le processus d'indemnisation des victimes d'accidents de la circulation. Les victimes ont le droit d'être indemnisées pour tous leurs dommages. L'indemnisation doit alors refléter à la fois les pertes financières, telles que les frais médicaux ou la perte de revenus, et les pertes non financières, telles que la douleur et la souffrance. Le système français est un système en trois étapes. La compagnie d'assurance propose une indemnisation extrajudiciaire fondée sur l'expertise médicale. Suite à cette première proposition, la victime est confrontée à un choix simple : soit accepter cette indemnisation, soit saisir le tribunal. Si la victime décide d'aller en justice, un deuxième montant lui sera proposé. Une dernière étape peut avoir lieu en cas d'appel.

Cette procédure peut être vue comme un problème de décision. Le choix entre accepter le montant proposé à l'amiable ou recourir aux tribunaux est basée sur la différence entre le montant à l'amiable et le montant espéré au procès. L'aversion au risque et les préférences temporelles de la victime influencent également la décision de négocier ou de poursuivre. D'un point de vue économétrique, la seule information disponible est le montant final obtenu, la durée de la procédure totale et la procédure choisie. En effet, les données sont malheureusement incomplètes. Pour les montants acceptés à l'amiable, la victime n'étant pas allée au procès, le montant judiciaire est inconnu. Pour les montants décidés au procès, cela se justifie d'un point de vue légal. Le juge doit proposer un montant au procès sans se baser sur la valeur qui a été proposée à l'amiable. La disponibilité de l'information dépend donc de la décision et les variables sont dépendantes limitées. Basé sur le modèle de probit structurel de Maddala (1986) et Lee (1979), un modèle à cinq équations est développé : une équation de décision pour modéliser le choix entre négociation et procès, deux équations de montant pour modéliser l'indemnisation perçue dans le cadre de procédures extrajudiciaires et judiciaires et deux équations pour modéliser la durée des procédures extrajudiciaires et judiciaires.

Contents

General Introduction

1 Motivation

Econometrics, which emerged in the 1930s, aims to reduce the gap between economic models and economic data. Statistical models are initially created to model independent and identically distributed data. However, there are rarely perfect data in the real world, and data are subject to various disturbances. Econometricians have, over the years, suggested appropriate statistical methods for the different kinds of data collected, allowing the estimation of economic relationships, the testing of economic theories, or the evaluation of the relationship between economic variables (Wooldridge et al., 2018).

According to Griliches (1985), data and econometricians constitute an uneasy alliance. In his own words:

> Econometricians have an ambivalent attitude towards economic data. At one level, the "data" are the world that we want to explain, the basic facts that economists purport to elucidate. At the other level, they are the source of all our troubles.

In this quote, the author explains the ambiguity of the data that are both the problem and the solution. The data are not perfect and are sometimes not adapted to the problem encountered. The main reason resides in the distance between economists and the "producers" of data. Many data are collected by survey organisms and institutes. Economists, and more generally in the

social sciences, use them *a posteriori* to deal with a question. This question is not necessarily considered *a priori* at the survey's conception. Hence, the data collected may not entirely match with the research question in some cases. Nevertheless, if the data were perfect, econometrics would not exist (Griliches, 1985). Imperfect data complicate or even inhibit the analysis. Data imperfections constitute therefore one of the key issues in econometrics. A great deal of research examines the problems associated with imperfect data, whether they are imprecise, ambiguous or incomplete.

The data encountered differ in their nature and specificity. Firstly, a disparity exists between the statistical individuals studied. These include time series, cross-sectional or panel data. Multiple types of individuals can be considered even within these categories. Time-series periods can range from a year for macroeconomic models to a few seconds for financial models. For cross-sectional data, the individuals studied can be households, firms, or even countries or cities.

Moreover, the quality of the information observed differs between studies. Some data are subject to alterations. They may be incomplete due to problems with missing data, truncation or censorship. They are sometimes inaccurate due to measurement errors or aggregation. Indeed, an aggregated data provides only a diluted picture of reality. The choice of categorical classifications may generate irregularities. A simple example is the disparity among age groups (Colvez and Villebrun, 2003). The thresholds used to categorize individuals are not standardized and consistent. The classification can be established by social thresholds: up to 16, school age, from 16 to 25, student age, from 26 to 64, working life and after 65, retirement. Individuals can also be classified by tens or twenties (Figure 6). This decision is subjective, both in terms of the choice of the number of categories and of the pivotal ages. The switch between categories is arbitrary. At which age are we considered "young" or "old"? Changes in education or retirement age also lead to variations in age limit selection. All of these issues should be carefully considered in the modeling and analysis.

Econometrics is related to economics and enables to formalize and stylize an economic phenomenon from reality. Economy evolves, therefore, econometrics needs to adapt and propose new statistical methods suitable to economic phenomena and to the diversity of data. Over past decades, econometricians have been facing new problems related to data evolution. For instance, nowadays, the emergence of Big Data and significant technological advances bring a new challenge to find relevant methods for manipulating and analyzing massive data. In parallel, the several reforms of the personal data law restrict access to individual data. Whether for

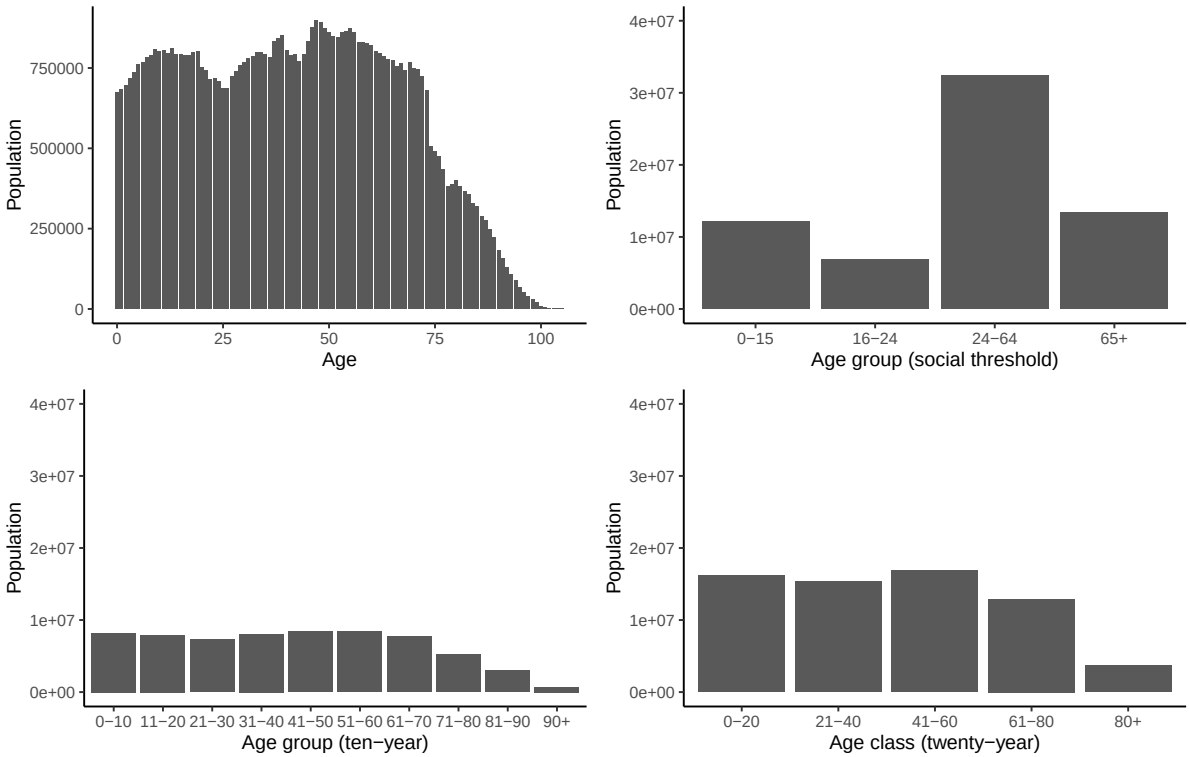


Figure 6. Total population on January 1, 2020, France (source: INSEE).

storage or collection considerations or anonymity requirements, aggregation of data is usually predominant, and aggregated data is often accessible in all circumstances and eliminates many measurement problems in surveys (Shively, 1969).

Exploring inequalities, especially income inequality, is complicated due to the prevalence of aggregated data. For individuals, income is a sensitive factor, such as health, ethnic origin, political opinions and religious beliefs. These restrictions on personal data obligate public authorities to only provide aggregated data instead of individual data. As a consequence, the requirements for individual data reporting constrain analysis and study of inequality and income measures. The predominant aggregation, especially for census data, is the spatial one. Data availability is further related to the size of the target geographic unit. Several levels of aggregation exist, ranging from countries to 200-meter zones. At the macro level, the data are potentially more detailed than at the neighbourhood level as they involve a larger number of individuals. In addition, the boundaries of these geographical units create a spatial aggregation problem (MAUP - Modifiable Areal Unit Problem) (Openshaw, 1984a,b, Openshaw and Taylor, 1979). The spatial division selected (or imposed) affects each result of the analyses. The MAUP has two effects: the zoning effect and the scale effect. The scale effect refers to the impact of the chosen aggregation scale. The zoning effect refers to the impact of zone boundaries at the same aggregation scale. For example, if we consider the number of owner-occupied households

in Paris according to different scales, we can reach several conclusions. Figure 7 depicts an iris cutting in Paris. Figure 8 represents a tile cut of 200 side meters. In both cases, the cutting accuracy is approximately the same, but the polygon boundaries differ, referred as the zoning effect. Figure 9 corresponds to the same variable by arrondissements. By comparison with the iris pattern, there is a scaling effect because the cut-off level is higher.

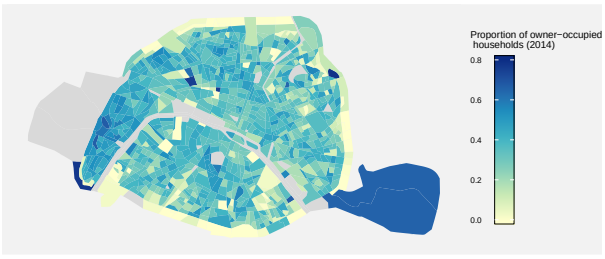


Figure 7. Proportion of owner-occupied households per iris
(Paris, 2014)

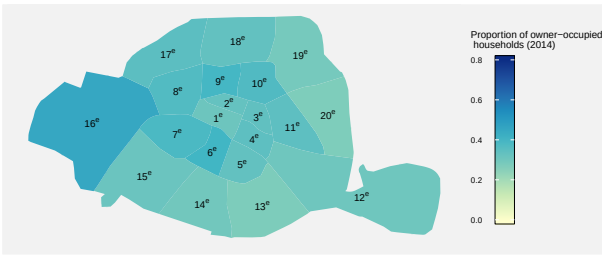


Figure 9. Proportion of owner-occupied households per arrondissements
(Paris, 2014)

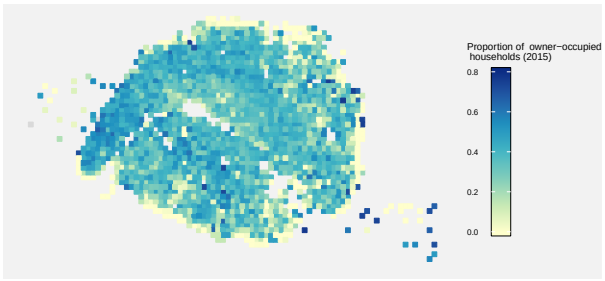


Figure 8. Proportion of owner-occupied households per 200 side meters tile
(Paris, 2015)

Moreover, in Economics, many variables are also categorical, counted, censored or truncated. The most suitable model must then be selected to prevent undesirable consequences (bias, inefficiency, inconsistency) (Orme and Buehler, 2001). In his book, Long (1997) reviews the models applicable to each of these variables, called "Categorical and Limited Dependent Variables". Information contained in these variables is by definition limited, whether categorical (binary, ordinal or multinomial), counted or limited (censored or truncated). Constraints on these variables result from several factors, including privacy, survey duration, choices of individuals or institutes, and the inherent nature of the data. Statistical institutes can transcribe continuous data into categorical variables for choice or confidentiality reasons. This transcription is done upstream by categorizing survey questions (income categories) or subsequently by discretizing a continuous variable (creation of age categories). The individuals' choice generates truncated or censored data (not answering a survey or question). Truncation may also occur from the choice of survey agency (polling only a part of the population). Finally, some data are inherently

qualitative and therefore not measurable (means of transport, eye colors).

The aim of this thesis is to propose and explore suitable methods to deal with imperfect data. This thesis will discuss two kinds of imperfect data: aggregated variables and limited dependent variables. In many economic phenomena, the data collected lead to practical problems. Potential imperfections should be considered when using the data, modeling or interpreting the results. Two economic phenomena are examined here: first, inequality measures and, second, bodily injury compensations. In each context, a set of relevant methods is provided to apprehend the available data and explain the phenomena.

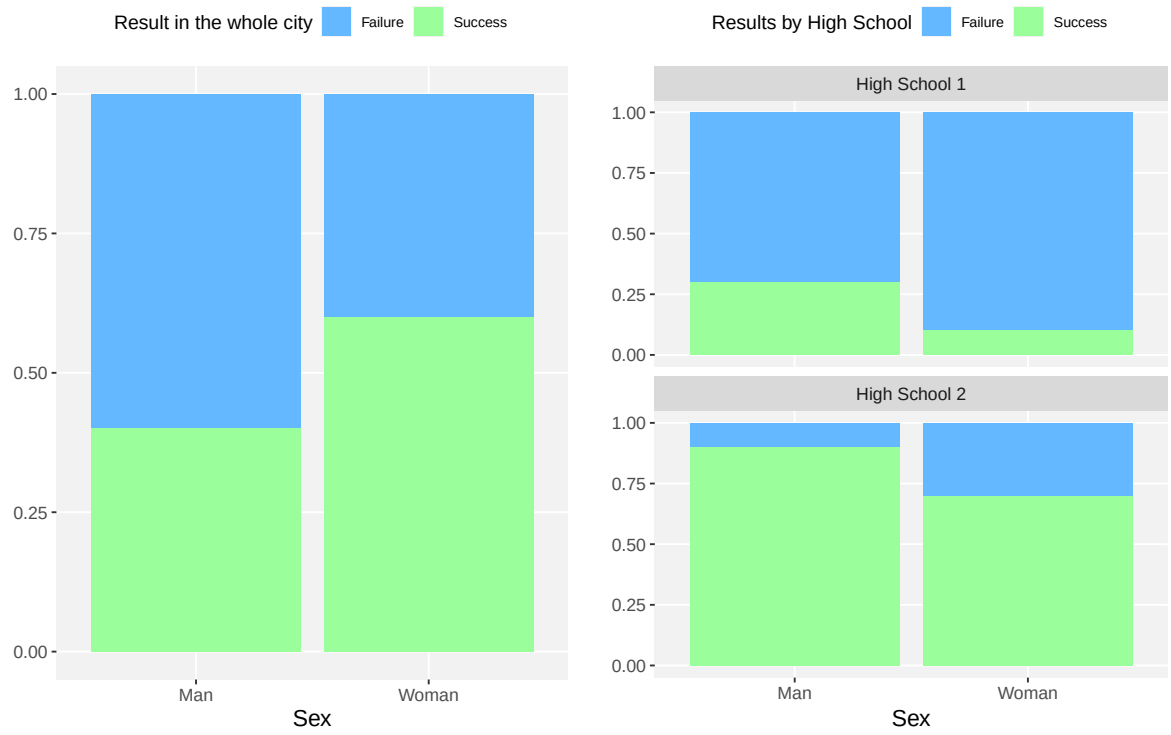
Section 2 describes ecological fallacy and compositional data concepts, and provides an overview of the solutions discussed in the literature. Section 3 presents the limited dependent variables, and reviews truncation, censorship and selection bias by detailing relevant econometric models. Finally, section 4 outlines the thesis plan and briefly summarizes the three chapters.

2 Aggregated data

Aggregated data are obtained by grouping individual data according to a geographic location, a time period or a characteristic. Such data can be of various types: number of individuals, averages, quantiles, etc. A measure of central tendency is computed by grouping or classifying individual data to ensure the preservation of individual identities.

A significant number of scientific contributions documents the loss of information related to the substitution of individual data by aggregated data. The pioneer Robinson (1950) asserts that aggregated (so-called *ecological*) correlations can not be used as substitutes for individual correlations. He highlights this assertion with the study of literacy rates and national origin. At the aggregate level, he notes a negative correlation between illiteracy and immigration: the more immigrants there are in a state, the lower the illiteracy rate. However, at the individual level, he finds that an immigrant is, on average, more illiterate. This phenomenon, called *Ecological Fallacy*, is linked to Simpson's paradox. This paradox refers to the confounding of aggregated and decomposed group correlations. The results obtained by synthesis on the whole population differ from the results obtained by segmentation with other criteria. Figure 10 depicts the Simpson's paradox based on the example proposed by Rouanet et al. (2002). This example illustrates educational success in the baccalaureate by gender. Consider a city with two high

schools. In the whole city, women have a higher success rate (60% compared to 40% for men). By contrast, at high school level, men have a higher success rate in both cases (30% against 10% in high school 1 and 90% against 70% for high school 2). The effect of gender on success rate is contradictory between the overall level and the conditional level. This finding is due to an omitted variable bias (here, high school) that is highly correlated with exposure (here, gender).



City	Success	Failure
Man	24	36
Woman	36	24

High Sch. 1	Success	Failure
Man	15	35
Woman	1	9

High Sch. 2	Success	Failure
Man	9	1
Woman	35	15

Figure 10. The Simpson's paradox (The high school paradox, Rouanet et al. (2002))

The Ecological Fallacy appears a major research topic in ecology, sociology, political science and relatively little in economics, and numerous methods have been proposed to infer individual behavior from these *ecological* data. As the information is relative, the number of parameters to be estimated is larger than the data available. Duncan and Davis (1953) suggest a method for bounding individual correlations from aggregate data and Goodman (1953) proposes a regression assuming that the coefficients do not depend on the area. Freedman et al. (1991) also gives an alternative with its *neighborhood model*. He assumes that, within an area, there is no systematic difference in outcomes between two groups. King (1997) proposes an advance in the ecological problem by combining the Goodman regression and the boundary method. Several authors also aim to measure the aggregation bias in regression and correlation analysis. Clark and Avery (1976) focus on the appearance of this bias in regression for spatial proximity aggregation. Orcutt et al. (1968) suggests that this bias lies in four phenomena: (i) the underlying nature

of reality, (ii) the aggregated data type, (iii) the relationship between the data and reality, and (iv) the statistical procedures used.

Working with aggregated data as a measure of either individual or aggregated behavior makes inference difficult. The anonymization or pseudonymization of personal data is nowadays required, especially to comply with the respective personal data reforms. Pseudonymization is reversible as opposed to anonymization. Anonymization of data can be conducted using several techniques: i) randomization, which consists of altering the data to make it non-identifying, and ii) generalization, which consists of diluting the data to make it unspecific to an individual.

Unfortunately, the data available for many studies are only aggregated. The idea is not necessarily to use these data to measure individual behaviours, but also to include them in aggregated scale studies in an appropriate way. An important concern is the use of factorial or categorical variables which become compositional variables at the aggregate level. Indeed, when a variable is categorical at the individual level, aggregating it provides the different proportions per class. For example, focusing on the level of education, at the individual scale, the level of education of each individual is known, but at the aggregate scale, only the proportions of different levels of education within a geographic area are known.

Definition 1 *A compositional random variable x of D shares is a $D \times 1$ vector with positive components $x_1, \ldots, x_D$ summing to 1. (Aitchison, 1986)*

Compositional data are common in economics, and more particularly the study of shares, initiated by Woodland (1979) and completed by Ronning (1992). Recently, Morais (2017) proposes an analysis of compositional time series and Fry (2011) a panorama of compositional data applications in economics.

Compositional variables provide only relative information and are limited to the simplex. The components of a composition are not free to fluctuate independently from each other and are not allowed to fall outside 0 and 1. Aitchison (1986) and his research on simplex geometry[1] introduces standard operations in simplex. He suggests the use of log-ratios (the additive transformation and the centered transformation) to bring the compositions back into real space and analyze them properly. Egozcue et al. (2003) seek an alternative transformation and propose the isometric log-ratio transformation. These transformations enable to work in coordinates

[1]The D-dimensional simplex is the set defined by $\tilde{\mathcal{S}}_D = \{(x_1, \ldots, x_D) : x_1 > 0, \ldots, x_D > 0; x_1 + \cdots + x_D = 1\}$ (Aitchison, 1986).

and to apply the usual statistical methods, considering the relative information of a composition (Van den Boogaart and Tolosana-Delgado, 2013).

3 Limited dependent variables

Limited dependent variables are also very common in economics (Maddala, 1986). Wooldridge (2002) defines a limited dependent variable as a variable "whose range is restricted in some important way". Many authors provide a review of the alternative regression models suitable for *categorical and limited dependent* data, such as Long (1997) and DeMaris (2004). Greene (2005) extends this theory by focusing on censored dependent variables and truncated underlying distributions. Limited dependent variables include discrete, censored and truncated variables.

Economists are frequently confronted with choice data between several alternatives. Answers then lead to discrete binary data (an individual's employment/non-employment), ordinal data (an individual's satisfaction) or multinomial data (an individual's means of transport). Counted data (number of events) are also included among the discrete limited dependent variables. Variables being discrete, they can not be manipulated like continuous variables. Statistical models have been developed in order to analyse them: generalized linear models (GLM) such as regression models with binary choice (probit or logit model), ordered choice (ordered probit or logit model), multinomial choice (multinomial logit model) or counting (Poisson or negative binomial model).

Limited variables also include censorship and truncation problems. As noted in Greene (2005), a distinction must be drawn between the censorship and truncation processes. Truncated data is a result of the population feature or the collection method, only a portion of the population is observed in the sample and thus the information loss is complete.

Definition 2 *A random variable x is left (respectively right) truncated if x is observable only if $x > T$ (resp. $x < T$), where T is a truncation random variable.*

On the contrary, censored data derive from a feature of the collection process or a natural element (survey duration, loss of sight). The value of the variable is not precisely known for a part of the sample.

Definition 3 *A random variable x is censored to the left (respectively to the right) if x is*

observed when $x \geq C$ (resp. $x \leq C$) and only C is observed if $x < C$ (resp. $x > C$) , where C is a censoring random variable.

Many economic variables are limited, but a treatment is not always necessary (Wooldridge, 2002). Variables such as wages or prices of goods are positive but continuous over the range. Further variables lead to corner solutions linked to an optimisation process. One example is the consumption of a good, where a part of the sample has a consumption equal to zero. In this case, the consumption of a good is continuous on positive values but has a non-zero probability at one point in the distribution (in this case, zero) and a treatment is required to seize the zero censorship. Furthermore, censorship may also occur because of data observation problems. For example, the entrance demand for a concert is censored if the event is sold out. The observed demand is the number of tickets sold (thus, the maximum capacity) but the underlying demand may be higher.

The simple Tobit model is used for censored data and corner solutions. In 1958, Tobin proposed a maximum likelihood estimation model to study the household expenditure of durable goods and to take into account the non-negativity of expenditure. Goldberger et al. (1964) later called it the Tobit model in reference to its proximity to the Probit model. Amemiya (1973) highlights the coherence and asymptotic normality of the maximum likelihood estimator. For truncated data, the truncated regression model is used (Greene, 2005).

Furthermore, censorship occurs when the data is survival. Censored data can be observed in medicine such as a pharmaceutical test or a cohort study, in insurance by studying the occurrence time of a claim or in the industrial sector by examining the breakdown or failure rate of a machine. In these cases, the data collected is frequently only partially collected because of censorship or truncation (Greene, 2005). Several models, which are mathematically close to the Tobit model and are called survival models, are preferred when dealing with this type of data: Buckley and James (1979), Cox (1972), Koul et al. (1981), Miller and Halpern (1982), Ritov (1990).

The Tobit model is particularly relevant for exogenous censorship. However, censorship or truncation might derive from an endogenous process. The information available directly results from decisions made by individuals themselves and individuals are self-selected. Consider the example of personal injury: the information available on the compensation amounts of individuals, whether the amount is settled out of court or at trial, is a direct result of their

decision whether or not to go to trial. The truncation is then related to another auxiliary variable. This is usually called accidental truncation.

In order to deal with this problem, Heckman (1976) and Heckman (1979) proposes a selection model, called Tobit II or Heckit, for studying women's labour supply and wages. Market wages are only observed for working women. The availability of their wages is a result from their decision to enter the labour market. He considers a two-equation model taking into account the sample selection bias. The idea is to correct this bias by introducing an additional explanatory variable: the inverse of the Mills ratio. A two-step estimation method is proposed using a Probit model to determine the Mills ratio and then a linear model including the inverse of the Mills ratio as an explanatory variable. Lee (1978) suggests a similar model for studying the wages of workers, whether or not they are union members. Amemiya (1984) classifies the various models related to the Tobit model into five basic types according to the form of the likelihood function.

4 Thesis outline

This thesis consists of three chapters about the use and the application of econometric and statistical methods when data are incomplete. It explores the inference and estimation problem in two kinds of imperfect data: aggregated data and censored data. In the context of aggregated data, a study of compositional variables related to the aggregation of categorical data (Chapter I) and a method for estimating inequality measures from quantile data (Chapter II) are provided. In the context of the censored data, a method for analysing the compensation amounts for bodily injury claims (Chapter III) is proposed.

The **first chapter**, entitled "Aggregate data and compositional variables" ("Données agrégées et variables compositionnelles"), examines the use of aggregated data in a statistical framework, in particular, categorical data that become compositional variables. In data analysis, compositional data are commonly used in several fields such as geology, economics or social sciences. In particular, the reform of personal data legislation in Europe caused difficulties in obtaining individual data, especially when such data is deemed sensitive. The commonly used solution is the release of spatially aggregated data. Compositional data are vectors which represent the several parts of a whole and are constrained by a unitary sum. Such data generally appears when they are standardized such as the composition of each mineral in a rock, the proportion of age groups in a town or census data.

The purpose is to provide methodological tools to apply standard statistical methods to these data. In this chapter, compositional regression analysis is examined, focusing on the case of compositions as explanatory variables. Traditional methods seem to be inconsistent with compositions which, because of the relative information they contain, lead to a misleading interpretation of the coefficients. This spurious interpretation is notably induced by the *ceteris paribus* clause and by problems of multicollinearity. Therefore, the proposed solution to apply regression methods on composition covariates is to work in coordinates. An application on median income in Paris according to degree levels and housing size is presented.

The **second chapter**, entitled "Estimating Inequality Measures from Quantile Data", focuses on the problem of using aggregated data in the context of inequality measures. An appropriate and widely used approach to depicting economic inequality is to provide indices to measure the degree of inequality. These measures enable comparisons of living standards between different countries, regions or among time. However, individual-level income data are rarely provided, especially for narrow geographic areas. Due to privacy and confidentiality restrictions, individual or income share data are usually not reported and only quantiles are available. The aim of this chapter is to determine a methodology for estimating a Lorenz curve and the associated indices with only quantile data. The methods should therefore be adapted in order to obtain a Lorenz curve from these quantiles.

This second chapter proposes an innovative method for modelling Lorenz curves and estimating inequality indices over small-scale populations where only quantiles are available. The method is based on the conditional expectation to find the different income shares and thus to model a Lorenz curve using the functional forms suggested in the literature. Inequality indices can subsequently be derived using this Lorenz curve (Gini, Pietra, Theil indexes). Simulations are performed to evaluate this method and compare it with other approaches used. An example based on real Parisian income data is also provided as an illustration of the method.

The **third chapter**, entitled "Bodily Injury Claims in France: Negotiation or Court?", examines the bodily injury compensation procedure in France. The French compensation system, as in many developed countries, is a no-fault system. These systems are based on compensation to the victim without fault and an insurer then substitutes for the defendant. In France, the law of 5 July 1985 (*Badinter* law) aims to improve the compensation process for victims of traffic accidents. Victims are entitled to be compensated for all their damages. Compensation must then reflect both financial losses, such as medical expenses or loss of income, and non-financial

losses, such as pain and suffering. The French system is a three-step system. The insurance company proposes an out-of-court compensation on the basis of medical expertise. After this first proposal, the victim is faced with a simple choice: either accept this compensation or go to court. If the victim decides to go to court, a second amount will be proposed. A final step may be taken in the case of an appeal.

This procedure can be considered as a decision-making problem. The choice of whether or not to go to court is based on the difference between the amount settled out of court and the amount expected at trial. The victim's risk aversion and time preferences also affect the decision to negotiate or pursue. From an econometric perspective, the only information available is the ultimate amount obtained, the duration of the total procedure and the procedure chosen. Unfortunately, the data is incomplete. For out-of-court settlements, the victim did not go to court and therefore the judicial amount is unknown. For trial cases, a legal argument is provided to justify this. The judge must propose an amount regardless of the extrajudicial value previously proposed. The information availability is therefore decision-dependent and the variables are limited dependent. Based on the structural probit model of Maddala (1986) and Lee (1979), a five-equation model is developed: a decision equation to model the choice between negotiation and trial, two amount equations to model the compensation received in extrajudicial and judicial procedures and two equations to model the duration of extrajudicial and judicial procedures.

Bibliography

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall London, 1986.

T. Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016, 1973.

T. Amemiya. Tobit models: A survey. *Journal of econometrics*, 24(1-2):3–61, 1984.

J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

W. A. Clark and K. L. Avery. The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4):428–438, 1976.

A. Colvez and D. Villebrun. La question des catégories d'âge et des «charnières» entre les différents types de population. *Revue française des affaires sociales*, (1):255–266, 2003.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

A. DeMaris. *Regression with social data: Modeling continuous and limited response variables*, volume 417. John Wiley & Sons, 2004.

O. D. Duncan and B. Davis. An alternative to ecological correlation. *American sociological review*, 1953.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

D. A. Freedman, S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. Ecological regression and voting rights. *Evaluation Review*, 15(6):673–711, 1991.

T. Fry. Applications in economics. *V. Pawlowsky-Glahn, and A. Buccianti*, 2011.

A. S. Goldberger et al. Econometric theory. *Econometric theory.*, 1964.

L. A. Goodman. Ecological regressions and behavior of individuals. *American sociological review*, 1953.

W. H. Greene. Censored data and truncated distributions. *Available at SSRN 825845*, 2005.

Z. Griliches. Data and econometricians–the uneasy alliance. *The American Economic Review*, 75(2):196–200, 1985.

J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

G. King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press, 1997.

H. Koul, V. v. Susarla, J. Van Ryzin, et al. Regression analysis with randomly right-censored data. *The Annals of statistics*, 9(6):1276–1288, 1981.

L.-F. Lee. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International economic review*, pages 415–433, 1978.

L.-F. Lee. Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica: Journal of the Econometric Society*, pages 977–996, 1979.

J. S. Long. Regression models for categorical and limited dependent variables (vol. 7). *Advanced quantitative techniques in the social sciences*, page 219, 1997.

G. S. Maddala. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.

R. Miller and J. Halpern. Regression with censored data. *Biometrika*, 69(3):521–531, 1982.

J. Morais. *Impact of media investments on brands' market shares: a compositional data analysis approach*. PhD thesis, 2017.

S. Openshaw. Ecological fallacies and the analysis of areal census data. *Environment and planning A*, 16(1):17–31, 1984a.

S. Openshaw. The modifiable areal unit problem. *CATMOG - Concepts and Techniques in Modern Geography*, 38, 1984b.

S. Openshaw and P. Taylor. Statistical applications in the spatial sciences, chapter a million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Wrigley N. Publishers, London, Pion*, pages 127–144, 1979.

G. H. Orcutt, H. W. Watts, and J. B. Edwards. Data aggregation and information loss. *The American Economic Review*, 58(4):773–787, 1968.

J. G. Orme and C. Buehler. Introduction to multiple regression for categorical and limited dependent variables. *Social Work Research*, 25(1):49, 2001.

Y. Ritov. Estimation in a linear regression model with censored data. *The Annals of Statistics*, pages 303–328, 1990.

W. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 1950.

G. Ronning. Share equations in econometrics: A story of repression, trustration and dead ends. *Statistical papers*, 33(1):307, 1992.

H. Rouanet, F. Lebaron, V. Le Hay, W. Ackermann, and B. Le Roux. Régression et analyse géométrique des données: réflexions et suggestions. *Mathématiques et sciences humaines. Mathematics and social sciences*, (160), 2002.

W. P. Shively. "Ecological" inference: the use of aggregate data to study individuals. *The American Political Science Review*, 63(4):1183–1196, 1969.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

K. G. Van den Boogaart and R. Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

A. D. Woodland. Stochastic specification and the estimation of share equations. *Journal of Econometrics*, 10(3):361–383, 1979.

J. Wooldridge, P. André, M. Beine, S. Béreau, M. de la Rupelle, A. Durré, J.-Y. Gnabo, C. Heuchenne, M. Leturcq, and M. Petitjean. *Introduction à l'économétrie: une approche moderne*. De Boeck Supérieur, 2018.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. Cambridge and London: MIT Press, 2002.

CHAPITRE I

Données Agrégées et Variables Compositionnelles[1]

[1] Chapitre co-écrit avec Arthur Charpentier

1 Introduction et Motivation

La microéconométrie est un outil indispensable afin d'étudier les phénomènes économiques à partir de données microéconomiques. Ces données se réfèrent aux unités décisionnelles au niveau microéconomique : entreprises, ménages, individus, etc. On peut alors être confrontés à deux type de données : les données à caractère personnel ou les données anonymes. Pour partager des données personnelles, il est aujourd'hui indispensable d'anonymiser, de manière réversible ou non (l'anonymisation réversible est aussi appelée pseudonymisation). L'anonymisation est une stratégie intéressante, en effet, comme le note le règlement européen du 27 avril 2016 "il n'y a pas lieu d'appliquer les principes de protection aux données qui ont été rendues suffisamment anonymes pour que la personne concernée ne soit plus identifiable". Parmi les techniques classiques, la randomisation propose d'altérer les données pour les rendre non identifiantes et la généralisation propose de diluer les données personnelles de telle sorte qu'elles perdent en précision et qu'elles ne soient plus spécifiques à une personne, mais communes à un ensemble de personnes (par exemple avec la notion de k-anonymat). C'est cette dernière technique que nous étudierons ici.

Dans le cadre européen, le Règlement Général sur la Protection des Données référence et renforce les dispositions relatives à la protection des données personnelles pour les individus de l'Union Européenne. Par conséquent, les diverses instituts statistiques doivent s'adapter pour répondre aux directives mises en place par ce règlement. Au lieu de mettre à disposition des données individuelles pour faire des études (revenus fiscaux des contribuables, résultats scolaires d'élèves scolarisés, etc.), il est devenu fréquent de publier des données agrégées (revenus fiscaux par zone IRIS, résultats par établissements scolaires, etc.). Cependant, deux problèmes se posent alors très rapidement : peut-on utiliser ces données agrégées pour inférer des comportements individuels (on parlera d'*inférence écologique*) et comment manipuler les variables catégorielles (que deviennent des variables compositionnelles une fois agrégées). L'objectif de ce chapitre est de répondre à ces questions en proposant un panorama des méthodes applicables aux données compositionnelles et au problème de l'inférence écologique.

Nous allons présenter dans la section 2 les données agrégées, et plus particulièrement l'agrégation spatiale. La section 3 reviendra sur la difficulté d'agréger les variables catégorielles et de les utiliser en régression. Finalement, dans le section 4, nous reviendrons sur le problème de l'inférence écologique, qui, d'un point de vue formel, dit que si on cherche le lien entre deux

variables x et y mais que les données individuelles ne sont pas disponibles, et qu'on dispose simplement d'agrégation basée sur une variable z (ici spatiale), il convient de tenir compte des corrélations croisées entre x, y et z.

1.1 Inférence écologique et Agrégation

Pour de nombreuses applications économiques, les données individuelles sont difficiles à obtenir car il est compliqué de les anonymiser : par exemple, si on dispose du revenu fiscal, de l'âge et du code postal, les données ne sont plus anonymes, au sens du Règlement Général sur la Protection des Données (RGPD, ou règlement 2016/679). Une des dispositions du RGPD est l'encouragement à anonymiser les données par agrégation, avec k individus au moins (pour obtenir le k-anonymat).

Considérons des observations individuelles $(y_i, \boldsymbol{x}_i, z_i)$, avec $z \in \mathcal{Z} = \{\zeta_1, \zeta_2, \cdots, \zeta_m\}$ un ensemble fini (correspondant aux différents groupes au sein desquels on va agréger les données) et un modèle de régression

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} \tag{I.1}$$

Une autre écriture usuelle est

$$y_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{I.2}$$

où ε désigne un bruit imprévisible. Supposons que ces données ne sont pas observables directement et que seules des grandeurs agrégées suivant la variable z sont disponibles. On parlera éventuellement de "binning". Classiquement, z pourra désigner une zone géographique (Openshaw (1984b) pour une revue exhaustive ou Clark and Avery (1976) une analyse des quartiers de Los Angeles), un bureau de vote (Klima et al. (2019) ou King (1997) pour des analyses politiques), un hôpital (Schwartz (1994), Greenland (2001) ou Berlin et al. (2002) pour des analyses de santé publique). Soit $n_j = \#\{i : z_i \in \zeta_j\}$ le nombre d'observations dans la classe j. On note alors la version agrégée de y et des variables x_u,

$$\overline{y}_j = \frac{1}{n_j} \sum_{z_i \in \zeta_j} y_i \text{ et } \overline{x}_{u,j} = \frac{1}{n_j} \sum_{z_i \in \zeta_j} x_{u,i} \text{ pour } u = 1, 2, \cdots, k.$$

Avec ces notations, on dispose d'observations agrégées $(\overline{y}_j, \overline{\boldsymbol{x}}_j, z_j)$ et on considère la régression

$$\overline{y}_j = b_0 + \overline{\boldsymbol{x}}_j^\top \boldsymbol{b} + \eta_j. \tag{I.3}$$

La question centrale en inférence écologique est de savoir si $\widehat{\boldsymbol{b}}$, estimateur par moindres carrés de $\boldsymbol{b}$, est aussi un bon estimateur de $\boldsymbol{\beta}$. Autrement dit, on se demande si un effet significatif observé au niveau agrégé (sur une des variables) existe encore au niveau individuel. Et, réciproquement, si un effet significatif au niveau individuel sera observable au niveau agrégé.

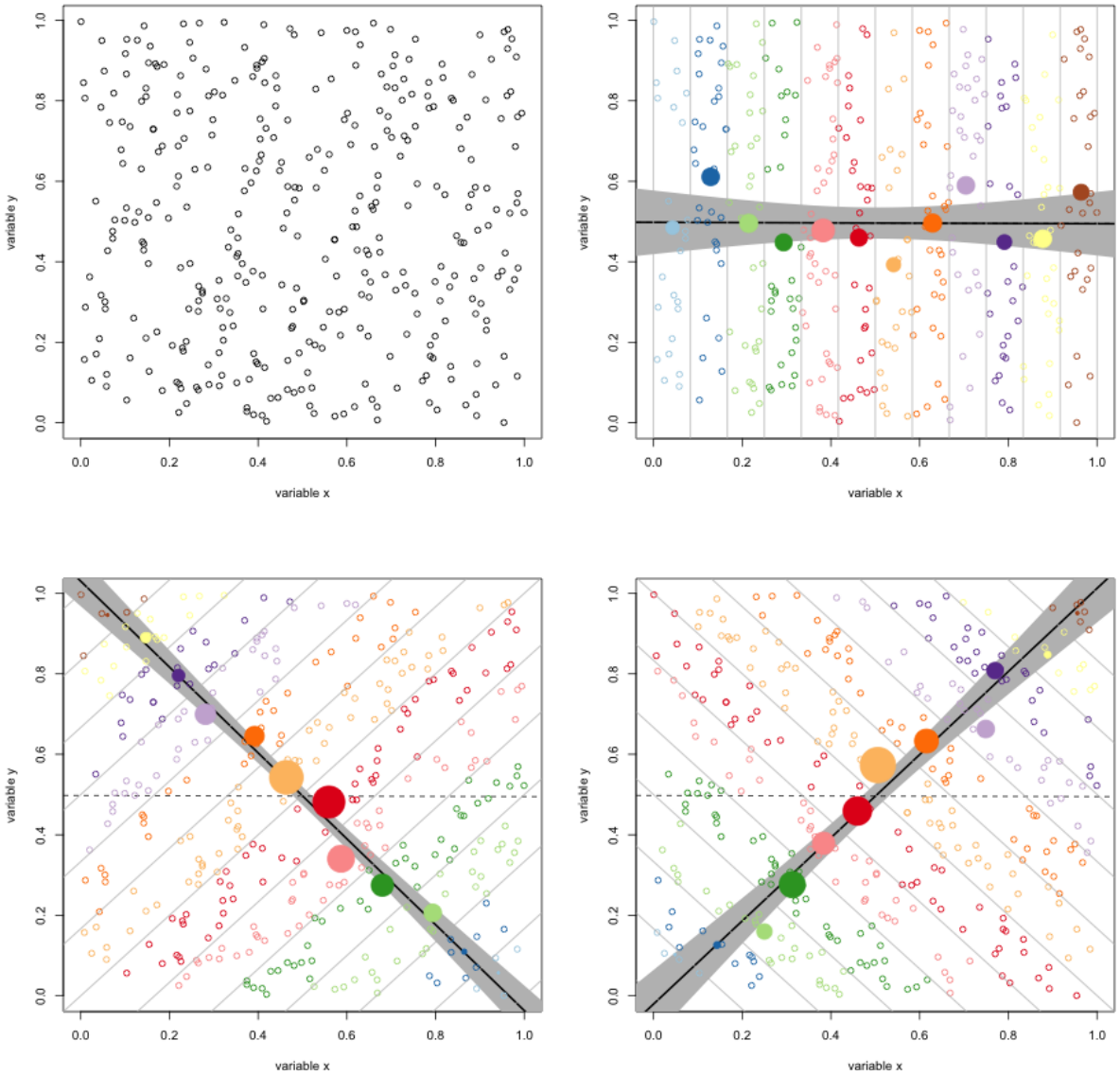


Figure I.1. Paradoxe écologique

Note : Corrélation entre $\overline{y}_j$ et $\overline{\boldsymbol{x}}_j$ en fonction du regroupement par classe, sur un même jeu de données (en haut à gauche). Les trois autres graphiques présentent (en ronds pleins) différentes agrégations. En haut à droite, les classes sont verticales (et donc suivant x), les ronds pleins sont les moyennes des variables x et y, par classe. En bas, les classes sont respectivement à $+45°$ (suivant $x + y$) et $-45°$ (suivant $y - x$). En haut à droite, $\overline{y}_j$ et $\overline{\boldsymbol{x}}_j$ (comme sur les données initiales, x_i et y_i étant ici indépendantes), alors qu'en bas, $\overline{y}_j$ et $\overline{\boldsymbol{x}}_j$ sont respectivement corrélés négativement et positivement.

Sur la Figure I.1, des variables x et y indépendantes sont simulées, de telle sorte que $\beta = 0$. Dans le cas en haut à droite, la variable d'agrégation z est (parfaitement) positivement corrélée avec x (et indépendante de y, en fait ici $z = x$ - ou plutôt un découpage en classes de x) et

$\widehat{b} \approx 0$. En revanche, en bas à gauche, si z est positivement corrélée avec x, et négativement avec y (en fait ici $z = y - x$), $\widehat{b} < 0$. Et en bas à droite, si z est positivement corrélée avec x, et aussi avec y (ici $z = y + x$), $\widehat{b} > 0$. Dans les deux cas, les signes sont opposés, et largement significatifs. Travailler sur des variables agrégées est complexe et dépend très fortement de la corrélation entre la variable d'agrégation z et les variables $\boldsymbol{x}$ et y. Nous reviendrons sur ce point dans la Section 4.

Plus formellement, le problème avec l'inférence écologique est que le processus d'agrégation réduit l'information, et cette perte d'information empêche - habituellement - l'identification des paramètres d'intérêt dans le modèle au niveau individuel sous-jacent. Spécifier le processus d'agrégation est utile pour mieux comprendre, même si dans l'article fondateur de Robinson (1950), la corrélation entre le lettrisme et la race a été calculée à divers niveaux d'agrégation géographique, comparée à la corrélation au niveau individuel, mais il n'est jamais fait explicitement référence à un modèle structurel d'agrégation. Comme on l'a vu sur la Figure I.1, le biais écologique est dû à la variabilité intra-zone.

Si les problèmes d'agrégation sont apparus assez tôt en économie - avec Theil (1954) et les travaux de la commission Cowles à la même époque - avec une discussion sur la cohérence entre les modèles microéconomiques et macroéconomiques, la littérature économétrique autour de l'inférence écologique s'est essentiellement développée en dehors du champs de l'économie.

1.2 Des variables qualitatives aux compositions

Un autre problème surgit rapidement lorsque l'on agrège les données, plus particulièrement les variables factorielles ou catégorielles. En effet, si x est une variable catégorielle, prenant d modalités $\{\xi_1, \xi_2, \cdots, \xi_d\}$ (par exemple le niveau d'étude, la catégorie professionnelle du chef de ménage, les classes d'âge, etc.), en agrégeant les données, on obtient souvent les proportions par classe. On aura ainsi la proportion de personnes de plus de 65 ans par département ou la proportion de cadres par commune. Formellement, on notera

$$\overline{x}_{u,j} = \frac{1}{n_j} \sum_{z_i \in \zeta_j} \mathbb{1}_{x_i \in \xi_u}, \ u = 1, 2, \cdots, d,$$

et $\overline{\boldsymbol{x}}_j = (\overline{x}_{1,j}, \overline{x}_{2,j}, \cdots, \overline{x}_{d,j})$ quand la variable x prend d modalités. On peut ainsi avoir $d = 2$, avec deux modalités, par exemple le genre (masculin ou féminin) au niveau individuel : par zone

j, on a le couple $(\overline{x}_{\mathsf{F},j}, \overline{x}_{\mathsf{H},j})$ indiquant les proportions de femmes et d'hommes, respectivement, en notant que $\overline{x}_{\mathsf{F},j} = 1 - \overline{x}_{\mathsf{H},j}$.

Les données compositionnelles sont classiques en économie, et plus particulièrement l'étude des parts (notion de "*shares*"), initiée par Woodland (1979) et complétée par Ronning (1992). Récemment, Morais (2017) propose une analyse de séries chronologies compositionnelles, $(\overline{x}_t)$ (correspondant à des parts d'investissements). Fry (2011) propose aussi un panorama d'applications des données compositionnelles en économie.

On pourrait aussi parler de données floues[2]. Une variable catégorielle agrégée est alors un vecteur de proportion, dont les composantes somment à 1. On parlera alors de compositions. Pour des raisons de simplification ultérieure, on demande aussi à ce qu'aucune composante ne soit nulle[3].

Définition 1 *Une variable compositionnelle $\overline{\boldsymbol{x}}$ à d composantes est un vecteur du sous-simplexe*[4] $\tilde{\mathcal{S}}_d \subset \mathbb{R}^d$,

$$\tilde{\mathcal{S}}_d = \left\{ \boldsymbol{u} = (u_1, \cdots, u_d) \in (0,\infty)^d : \sum_{i=1}^{d} u_i = 1 \right\}.$$

Soit x une variable qualitative prenant des valeurs dans $\mathcal{X}_d = \{\xi_1, \cdots, \xi_d\}$. En agrégeant les données suivant un critère z, on obtient une variable compositionnelle $\overline{x}$ à valeurs dans $\mathcal{X}_d$.

Définition 2 *L'opérateur de clôture $\mathcal{C}$ permet de passer de $\mathbb{R}_+^d$ à $\mathcal{S}_d \subset \mathbb{R}^d$, tout simplement en considérant*

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{n_1}{n_1 + \cdots + n_d}, \cdots, \frac{n_d}{n_1 + \cdots + n_d} \right)$$

Aussi, si on dispose d'observations $\{x_1, \cdots, x_n\}$, si on pose n_j la variable de comptage

$$n_j = \sum_{i=1}^{n} \mathbb{1}_{x_i = j}, \text{ pour } j \in \{1, 2, \cdots, d\},$$

[2] Au sens de la logique floue (ou *fuzzy logic*) : en logique booléenne, les valeurs de vérité des variables peuvent seulement être les valeurs entières 0 ou 1, mais en logique floue, n'importe quel nombre réel entre 0 et 1 peut indiquer la part de vérité. Ici au lieu d'avoir soit les catégories A, B ou C (et donc par exemple un vecteur $(0,1,0)$) on a des proportions pour chaque catégorie (par exemple $(5\%, 85\%, 10\%)$).

[3] Numériquement, la raison est qu'on transformera les composantes par un passage au logarithme. Intuitivement, on demande juste que toutes les modalités aient du sens : si on considère la couleur (naturelle) des cheveux, on mettra comme modalité 'roux' ou 'blond' mais pas 'vert' ou 'bleu' (pour lesquels $\overline{x}_{\mathsf{vert},j}$ ou $\overline{x}_{\mathsf{bleu},j}$ seront nuls).

[4] Le simplexe traditionnel est une portion d'hyperplan de $\mathbb{R}^d$, correspondant à l'enveloppe convexe de d points extrémaux, correspondants aux points $(0, \cdots, 0, 1, 0, \cdots, 0)$. Ici, on enlève les d espaces de dimension $d-2$ correspondant aux enveloppes convexes de $d-1$ points extrémaux. Pour $d = 3$, le simplexe est un triangle et le sous-simplexe est l'intérieur du triangle, auquel le contour a été enlevé.

la variable $\overline{\boldsymbol{x}} = \mathcal{C}(\boldsymbol{n}) = \mathcal{C}(n_1, \cdots, n_d)$ est une variable compositionnelle, donnant les proportions relatives dans chacune des classes pour la population totale.

Formellement, $\boldsymbol{n} = (n_1, \cdots, n_d)$ est vu comme la réalisation d'une loi multinomiale $\mathcal{M}(n, \boldsymbol{p})$ où $\boldsymbol{p} \in \mathcal{S}_d$ est le vecteur de probabilité inconnu. $\overline{\boldsymbol{x}} = \mathcal{C}(\boldsymbol{n})$ correspond aux fréquences empiriques dans chacunes des classes, et correspond à l'estimateur du maximum de vraisemblance du paramètre $\boldsymbol{p}$. La contrainte forte $\overline{x}_1 + \cdots + \overline{x}_d = 1$ impose une corrélation négative entre les composantes : si $\boldsymbol{N} \sim \mathcal{M}(n, \boldsymbol{p})$, et $\overline{\boldsymbol{X}} = \mathcal{C}(\boldsymbol{N})$, alors $\text{Cov}[\bar{X}_j, \bar{X}_{j'}] = -n^{-1} p_j p_{j'} < 0$. Cette multicolinéarité des composantes pose (potentiellement) des problèmes importants d'interprétation.

2 Des microdonnées aux données agrégées

Deux grands principes de la diffusion de statistiques publiques s'opposent de manière assez fondamentale, comme le rappelle VanWey et al. (2005). D'un côté, les Instituts de Statistique ont vocation à diffuser des données les plus utiles possibles, et de l'autre, ils doivent respecter de fortes contraintes de confidentialité pour les enquêtes, mais aussi pour bon nombre de données dites administratives (par exemple les données fiscales). Et comme le notent Loonis and Bellefon (2018), "*le plus souvent, la règle de protection des données n'est autre qu'un seuil, en deçà duquel on interdit la diffusion de statistiques*".

2.1 Agrégation par Zone Géographique et Anonymat

Les données, mises sous forme agrégées dans des tableaux, ont pendant longtemps constitué les sorties traditionnelles des organismes nationaux de statistique. En particulier, les statisticiens travaillant sur le contrôle de divulgation statistique - *Statistical Disclosure Control* (SDC), *Statistical Disclosure Limitation* (SDL) - ont pu établir des règles garantissant l'anonymat des données, important dans le cas de données sensibles. En prenant la terminologie imposée par Sweeney (2002), le k-anonymat fait en sorte que les enregistrements communiqués correspondent à des groupes d'au moins k individus. Pour satisfaire la confidentialité des données, les valeurs originales sont remplacées par une mesure centrale (généralement la moyenne ou la médiane). On va alors procéder en deux temps : partitionner puis agréger. Smith (1985) et Matthews et al. (2011) mentionnent ainsi que l'on va chercher à constituer des groupes aussi homogènes que possible - sans toutefois expliquer comment. Une approche classique consiste à grouper spatia-

lement les individus. Openshaw (1984a) rappelle ainsi que dans nombre de pays, les données de recensement sont reportées par unité géographique. Au Royaume-Uni par exemple, les informations sur les personnes et les ménages ne sont disponibles que sous forme agrégée, par zone géographique arbitraire. C'est d'ailleurs la méthode la plus simple évoquée dans Loth (2005).

Aux États-Unis, les regroupements classiques vont du comté (3077 comtés en 2000, soit 24544 habitants en moyenne, allant de 82 habitants - Loving au Texas - à plus de 10 millions) au code postal à 5 chiffres (41666 ZIP codes, avec en moyenne 7498 habitant en 2000, allant de 1 habitant à plus de 125000) ou celui à 9 chiffres (appelé ZIP+4). Ces derniers ne sont pas techniquement des régions (des polygônes) mais des rues. Au niveau européen, GeoStat 2011 a été le premier exemple de grille de population de l'Union Européenne[5]. Le projet du Système de Statistique Européen (ESSnet), a été lancé en coopération avec le Forum Européen de Géographie et de Statistique (EFGS), avec comme objectif de recenser la population et l'habitat dans un ensemble de données maillées de 1 km^2. Il y a un peu de moins de 2 millions de carrés, avec en moyenne 114 habitants par carré (densité de population en Europe). Toutefois, 50% des carrés sont peuplés par moins de 50 personnes (ce qui pose des soucis d'anonymat).

2.2 Le carroyage INSEE

L'INSEE est allé plus loin en proposant des carrés de 200m de côté[6]. L'INSEE a ainsi mis en ligne un grand nombre de données (notamment fiscales). Un article du Canard Enchaîné (le 27 février 2013) notait que 270 mille de ces carreaux ne comptait qu'une seule résidence, et sur 62 mille parcelles, un unique individu était rattaché. Le débat sur l'anonymat des données fiscales qui a suivi a poussé l'INSEE à suspendre temporairement la diffusion de ces données. Depuis, comme l'expliquent Loonis and Bellefon (2018) dans le dernier chapitre sur la 'confidentialité des données spatiales', ces carreaux ont été agrégés en rectangles quand ils comprenaient moins de 11 ménages fiscaux, seuil réglementaire à respecter. Classiquement, deux algorithmes peuvent être utilisés : une approche *backward* consiste à partir d'un découpage sommaire en quelques carrés, puis à couper en rectangles, à l'intérieur, tant que c'est possible (Encadré 14.2.1 page 367, méthode utilisée par l'INSEE) ou une approche *forward* partant d'une granularité très fine, puis à regrouper avec des voisins si nécessaire (Encadré 14.2.2 page 367, utilisé en Allemagne). Une étude sur les données fiscales mentionnée dans Loonis and Bellefon (2018) évoque trois

[5] https ://ec.europa.eu/eurostat/statistics-explained/index.php/Population_grids
[6] https ://www.insee.fr/fr/statistiques ?debut=0&categorie=1&geo=ICQ-1

niveaux d'agrégation. Le niveau le plus agrégé (niveau 3) correspond aux départements ; ensuite (niveau 2), on considère des 'gros rectangles' contenant au moins 5000 individus (intersectés avec le niveau 3) ; enfin (niveau 1), des 'petits rectangles' contenant au moins 100 individus (imbriqués dans les rectangles de niveau 2).

Ces carreaux, carrés ou rectangulaires, présentent l'avantage (théorique) de s'affranchir de tout zonage administratif, de ne tenir compte d'aucune réalité socio-économique, d'aucune contrainte naturelle, comme le rappelle Deichmann et al. (2001). Si les carreaux ont la même taille, les données carroyées permettent de comparer des territoires dans le temps.

2.3 Ilots Regroupés pour l'Information Statistique (IRIS)

En France, on peut rester au niveau des départements (une centaine - 101 - soit 664 500 habitants en moyenne, allant de 77000 habitants à près de 2,6 millions), les cantons (passés de 4035 à 2054) et les communes (353 570). La taille très hétérogène de ces dernières a poussé l'INSEE à introduire l'IRIS[7], pour *Ilots Regroupés pour l'Information Statistique* (16100 zones). Comme le rappelle l'INSEE, les IRIS d'habitat ont une population qui se situe en général entre 1800 et 5000 habitants. Ils sont (par construction) homogènes quant au type d'habitat et leurs limites s'appuient sur les grandes coupures du tissu urbain (voies principales, voies ferrées, cours d'eau). Sur la Figure I.2, on peut voir la distribution de la population par IRIS à gauche, et sur une carte à droite.

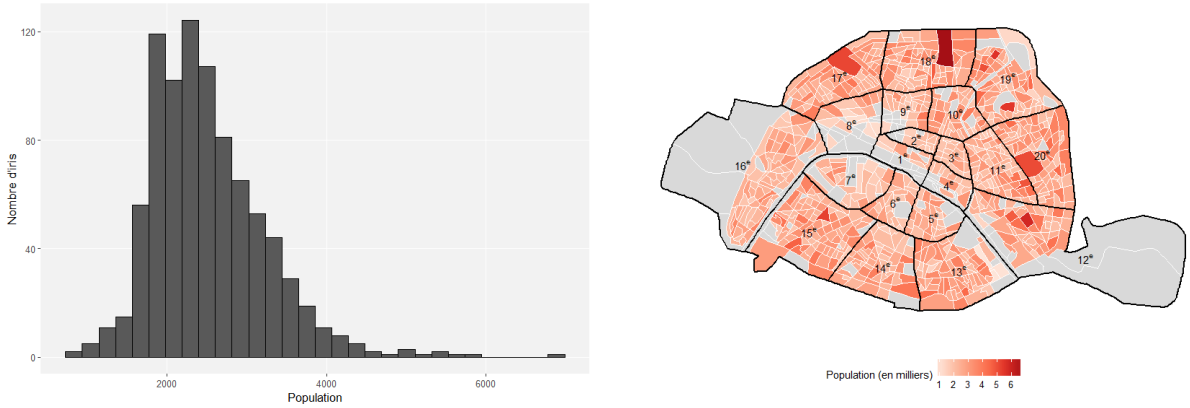


Figure I.2. Distribution de la population par IRIS dans Paris.

[7] https ://www.insee.fr/fr/metadonnees/definition/c1523

2.4 Agrégation spatiale

Même si Gehlke and Biehl (1934) ont découvert certains aspects de l'influence du découpage spatial (effets d'échelle et effets de zonage), le terme MAUP (*Modifiable Areal Unit Problem*) n'a été introduit officiellement qu'à la fin des années 70, avec Openshaw and Taylor (1979), qui proposait d'évaluer systématiquement la variabilité des corrélations lorsque différents regroupements spatiaux étaient considérés. En particulier, Openshaw and Taylor (1979) ont constaté que les corrélations (dans l'Iowa) entre le vote républicain et le pourcentage de personnes âgées pouvaient varier de -0,97 et +0,99 selon la façon dont les comtés étaient agrégés. Openshaw and Rao (1995) ont établi que la corrélation entre le chômage et l'absence de voiture pouvait varier entre -1,00 et +1,00, à partir de plus de 2500 ED (*enumeration districts*) dans le comté de Merseyside. Le rapport ESPON (Grasland et al., 2006) présente des agrégats fréquemment utilisées en Irlande, en Suède et en Allemagne.

Fotheringham and Wong (1991) reviennent sur l'impact du MAUP dans le contexte de régressions (linéaires et logistiques). L'étude note en particulier que lorsque de plus petites unités sont fusionnées pour former de plus grandes unités, en considérant les valeurs moyennes par unité, les corrélations entre les variables des unités agrégées sont souvent plus élevées que celles du niveau désagrégé (ou agrégées à un niveau plus granulaire). Comme le note Wong (2004), le MAUP est une des explications classiques du paradoxe écologique. En effet, bien souvent, les données spatiales sont des agrégats d'individus, mais le but de l'analyse est bien souvent d'identifier des comportements concernant les individus. Cependant, le MAUP fait que les données agrégées (peu importe l'échelle retenue) ne peuvent fournir une image cohérente de la situation individuelle. On retrouve ici le paradoxe écologique évoqué en introduction : il peut être erroné de déduire des situations individuelles à partir de données agrégées. Pearl and Mackenzie (2018) reviennent longuement sur ce point dans le chapitre 6.

Notons que s'il est plus facile d'avoir des données agrégées que des données individuelles, certaines données (jugées "sensibles") ne sont pas accessibles. Par exemple, la Direction générale des finances publiques (DGFiP) ne donnait le taux de personnes assujetties à l'ISF (Impôt sur la Fortune) qu'au niveau agrégé, pour des villes de plus de 20000 habitants (à condition qu'au moins 50 personnes dans la ville payent cet impôt). En 2017, on pouvait obtenir cette information pour 382 villes.

3 Données Compositionnelles et Géométrie du Simplexe

Pour introduire la régression sur des données compositionnelles, nous présenterons dans un premier temps la régression sur une seule variable, catégorielle, d'une variable continue y.

3.1 Régresser sur les composantes ?

Considérons un ensemble d'observations $(y_i, \boldsymbol{x}_i)$, où y est la variable continue que l'on souhaite décrire (par exemple le revenu) et $\boldsymbol{x}$ est une composition associée à une variable qualitative (par exemple le niveau d'étude). On peut écrire la régression linéaire

$$y_i = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon_i$$

tout en gardant en mémoire que pour avoir l'identifiabilité, une modalité devra servir de référence. Mais comme le notait Chayes (1960), la contrainte $\boldsymbol{x}^\top \mathbb{1} = x_1 + \cdots + x_d = 1$ (et donc la corrélation négative entre les composantes[8]) induisait un biais et des problèmes de corrélations fallacieuses. Et Chayes (1960) ne proposait pas de méthode satisfaisante pour les corriger. Il a fallu attendre Aitchison (1986) pour avoir une formalisation satisfaisante du problème. En effet, Aitchison (1986) a expliqué que les compositions ne fournissent des informations utiles que de manière relative. Cette observation a suggéré l'utilisation de ratios - et de log-ratios - pour analyser les compositions et développer des transformations des données, comme nous le verrons dans la section suivante. Mais c'est surtout les travaux sur la géométrie du simplexe qui ont permis de mieux comprendre comment analyser et interpréter les compositions.

3.2 Transformer les composantes (pour supprimer les contraintes)

Dans l'écriture du modèle de régression traditionnel, décrit par l'équation (I.1), $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$ le second terme correspond au produit scalaire $\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ dans $\mathbb{R}^d$. Si les variables explicatives sont dans le simplexe, il est alors naturel d'adapter la géométrie pour se placer dans le bon espace ou de transformer les variables pour se ramener dans $\mathbb{R}^{d-1}$.

Travailler dans le simplexe est très contraignant. Il convient mieux de se ramener dans

[8] Pearson (1897) a étudié les corrélations des rapports des mesures osseuses et a constaté que même si les corrélations entre les mesures initiales étaient faibles, les corrélations entre les rapports et les mesures communes étaient relativement importantes.

l'espace euclidien. Les transformations les plus populaires sont basées sur les log-ratios, dès lors qu'aucune composante n'est nulle. Soit $\boldsymbol{x} \in \mathbb{R}_+^d$, la transformation additive ALR_j avec pour référence la $j$ème composante est la transformation $\widetilde{\mathcal{S}}_d \to \mathbb{R}^{d-1}$

$$\text{ALR}_j(\boldsymbol{x}) = \left(\log \frac{x_1}{x_j}, \cdots, \log \frac{x_{j-1}}{x_j}, \log \frac{x_{j+1}}{x_j}, \cdots, \log \frac{x_d}{x_j} \right).$$

On peut définir la transformation inverse en notant que $\text{ALR}_j^{-1}(\boldsymbol{y}) = \mathcal{C}(\exp[\boldsymbol{y}_0^j])$, où $\boldsymbol{y}_0^j = (y_1, \cdots, y_{j-1}, 0, y_j, \cdots, y_d)$. En dimension $d = 3$, la transformation du log-ratio additif est

$$\text{ALR}_3(\boldsymbol{x}) = \left(\log \frac{x_1}{x_3}, \log \frac{x_2}{x_3} \right)$$

alors que la transformation inverse est

$$\text{ALR}_3^{-1}(\boldsymbol{y}) = \frac{(\exp[y_1], \exp[y_2], 1)}{1 + \exp[y_1] + \exp[y_2]}$$

qui rappelle la transformation logistique pour les variables multinomiales, introduite par Theil (1969).

On note CLR la transformation $\mathcal{S}_d \to \mathbb{R}^d$ dite centrée, où on n'utilise plus une modalité de référence, mais la moyenne géométrique

$$\text{CLR}(\boldsymbol{x}) = \left(\log \frac{x_1}{\tilde{x}}, \cdots, \log \frac{x_n}{\tilde{x}} \right), \text{ avec } \tilde{x} = \left(\prod_{i=1}^{d} x_i \right)^{1/d} \in \mathbb{R}^d.$$

et la transformation inverse est tout simplement $\text{CLR}^{-1}(\boldsymbol{y}) = \mathcal{C}(\exp[\boldsymbol{y}])$. Cette symétrie rend l'interprétation relativement simple, comme le montrent Filzmoser and Hron (2008). En dimension $d = 3$, la transformation du log-ratio centrée est

$$\text{CLR}(\boldsymbol{x}) = \left(\log \frac{x_1}{\tilde{x}}, \log \frac{x_2}{\tilde{x}}, \log \frac{x_3}{\tilde{x}} \right)$$

où $\tilde{x} = \sqrt[3]{x_1 x_2 x_3}$, alors que

$$\text{CLR}^{-1}(\boldsymbol{y}) = \frac{(\exp[y_1], \exp[y_2], \exp[y_3])}{\exp[y_1] + \exp[y_2] + \exp[y_3]}.$$

Enfin pour la transformation dite isométrique ILR, introduite par Egozcue et al. (2003), on

se donne une base orthonormée[9] $\boldsymbol{e} = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{d-1}\}$ de $\mathcal{S}_d$, et on note $\text{ILR}_{\boldsymbol{e}}$ la transformation $\widetilde{\mathcal{S}}_d \to \mathbb{R}^{d-1}$

$$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x}) = (\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle, \cdots, \langle \boldsymbol{x}, \boldsymbol{e}_{d-1}\rangle).$$

Notons que $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})^\top \text{ILR}_{\boldsymbol{e}}(\boldsymbol{y})$. Cette relation est particulièrement intéressante puisque le modèle de régression $y_i = \beta_0 + \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \varepsilon_i$ s'écrit alors simplement

$$y_i = \beta_0 + \text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})^\top \boldsymbol{b} + \varepsilon_i, \text{ où } \boldsymbol{b} = \text{ILR}_{\boldsymbol{e}}(\beta), \tag{ILR}$$

qui redevient une régression linéaire usuelle[10] et l'estimateur naturel de $\boldsymbol{\beta}$ est alors $\widehat{\boldsymbol{\beta}} = \text{ILR}_{\boldsymbol{e}}^{-1}(\widehat{\boldsymbol{b}}^{\text{OLS}})$

$$\widehat{\boldsymbol{b}}^{\text{OLS}} = (\text{ILR}_{\boldsymbol{e}}(\boldsymbol{X})^\top \text{ILR}_{\boldsymbol{e}}(\boldsymbol{X}))^{-1} \text{ILR}_{\boldsymbol{e}}(\boldsymbol{X})^\top \boldsymbol{y}.$$

3.3 Visualisation d'une régression

Si la variable explicative $\boldsymbol{x}$ peut prendre trois modalités, il est usuel de se placer dans la projection dans le plan du simplexe et de repérer un point en fonction des distances aux trois sommets. Le théorème de Viviani (*dans un triangle équilatéral, la somme des distances d'un point intérieur au triangle aux trois côtés est égale à la hauteur du triangle*) permet alors d'introduire la notion de graphiques "ternaires" (West, 2012). Ce graphique permet de visualiser des observations $\boldsymbol{x}_i$, comme sur la Figure I.3, à droite (à gauche, on peut visualiser les deux dernières coordonnées dans $\mathbb{R}^{d-1}$, la première servant ici de "référence").

Soit y_i le revenu dans la zone i, et $\boldsymbol{x}_i$ la variable compositionnelle indiquant la proportion de personnes ayant comme diplôme le plus élevé le brevet des collèges (BEP), le baccalauréat (BAC) ou un diplôme de l'enseignement supérieur (SUP), $\boldsymbol{x}_i = (\boldsymbol{x}_{\mathsf{BEP},i}, \boldsymbol{x}_{\mathsf{BAC},i}, \boldsymbol{x}_{\mathsf{SUP},i})$. La figure I.4 représente les distributions de revenu médian et du niveau de diplômes à Paris. Le premier modèle considéré est la régression classique

$$y_i = \gamma_0 + \gamma_2 \boldsymbol{x}_{\mathsf{BEP},i} + \gamma_3 \boldsymbol{x}_{\mathsf{SUP},i} + \eta_i \tag{OLS-1}$$

[9] Pour le produit scalaire d'Aitchison, comme définies dans Egozcue and Pawlowsky-Glahn (2006). Notons que si $\boldsymbol{x}, \boldsymbol{y} \in \tilde{\mathcal{S}}_d$, on peut définir le produit scalaire d'Aitchison en posant

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \frac{1}{d}\sum_{i<j} \log\frac{x_i}{x_j}\log\frac{y_i}{y_j}$$

[10] Techniquement, on peut doter le simplexe d'opérateurs de sommes et de produits, en posant $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(\boldsymbol{xy})$ et $b \otimes \boldsymbol{x} = C(\boldsymbol{x}^b)$ pour $b \in \mathbb{R}$. Dans ce cas, le modèle de régression multivarié peut formellement s'écrire $(\beta_1 \otimes X_1) \oplus (\beta_2 \otimes X_2) \oplus \cdots \oplus (\beta_k \otimes X_k) \oplus \varepsilon$ où les résidus étant alors dans le simplexe.

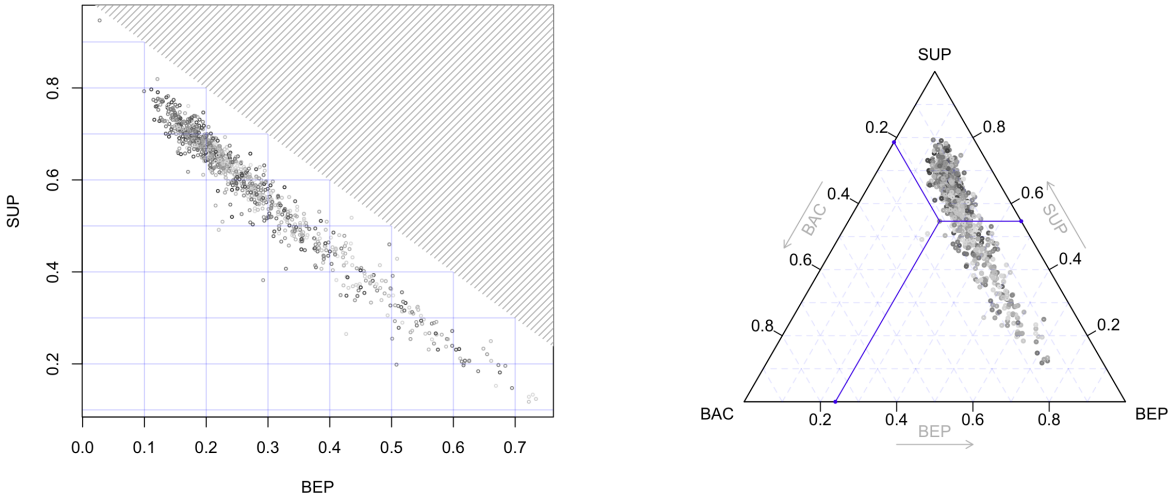


Figure I.3. Nuage des points (x_2, x_3) dans $\mathbb{R}^2$ à gauche, et nuage des points (x_1, x_2, x_3) dans $\widetilde{S}_3$ à droite.

Note : à gauche, la partie supérieure droite est inatteignable, puisqu'alors $x_2 + x_3 \geq 1$. À droite est également représenté le point $x = (21\%, 24\%, 55\%)$ dans l'espace (BAC, BEP, SUP).

On peut aussi considérer une régression sur une transformation de $\boldsymbol{x}$,

$$y_i = b_0 + b_1 \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_i)_1 + b_2 \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_i)_2 + \varepsilon_i \tag{ILR-1}$$

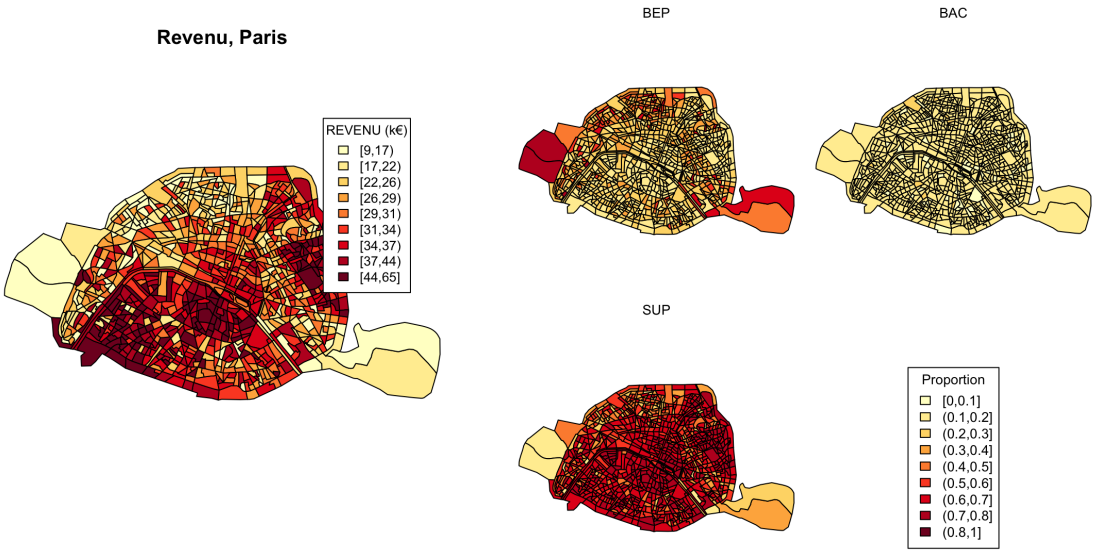


Figure I.4. Distribution du revenu médian à gauche, par quartier, à Paris, et distribution du niveau d'étude à droite, avec la proportion de BEP, BAC et SUP.

Le tableau I.1 fournit les résultats de ces deux régressions. L'interprétation de la partie gauche du tableau pour la régression linéaire (classique) est assez naturelle ici. La constante n'est pas significative, autrement dit, la proportion de personnes BEP n'apporte rien sur la

richesse du quartier. Mais cette dernière sera croissante avec la proportion de BAC et croissante avec la proportion de SUP. Pour la seconde régression, avec la transformation des variables, dont l'estimation est à droite du tableau, on peut transformer le paramètre estimé,

$$\widehat{\boldsymbol{\beta}} = \text{ILR}_{\boldsymbol{e}}^{-1}(\widehat{\boldsymbol{b}}) = (0.0565; 0.4892; 0.4542)$$

et la prévision s'écrit alors $\widehat{y}_i = \widehat{\beta}_0 + \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}} \rangle$. Ici $\boldsymbol{\beta}$ a une interprétation relativement simple, puisqu'il donne la direction dans laquelle $\boldsymbol{x}$ doit être perturbé pour avoir le plus d'effet sur y. En effet, si $\boldsymbol{u}$ est un vecteur unitaire

$$\mathbb{E}[Y|\boldsymbol{x} \oplus \boldsymbol{u}] = \beta_0 + \langle \boldsymbol{x} \oplus \boldsymbol{u}, \boldsymbol{\beta} \rangle = \mathbb{E}[Y|\boldsymbol{x}] + \underbrace{\langle \boldsymbol{u}, \boldsymbol{\beta} \rangle}_{\leq \|\boldsymbol{\beta}\|}$$

le maximum étant obtenu quand $\boldsymbol{u} = \boldsymbol{\beta}$. Mais probablement plus intéressant, on peut visualiser les prévisions, soit dans $\mathbb{R}^2$, soit dans le simplexe $\mathcal{S}_d$, comme sur la Figure I.5 avec les courbes d'iso-niveau du revenu dans le plan ternaire.

Table I.1. Régression linéaire : revenu en fonction du niveau d'étude.

	(OLS-1) Revenu		(ILR-1) Revenu
$\boldsymbol{x}_{\mathsf{BAC}}$	2.777***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1$	1.028***
	(0.671)		(0.122)
$\boldsymbol{x}_{\mathsf{SUP}}$	7.559***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_2$	1.229***
	(0.288)		(0.067)
Constante	-1.636	Constante	1.718
	(0.353)		(0.045)
Observations	868		868
R^2	0.669		0.701
R^2 ajusté	0.668		0.700
$\widehat{\sigma}$	0.594		0.565
Statistique F	873.7***		1009***

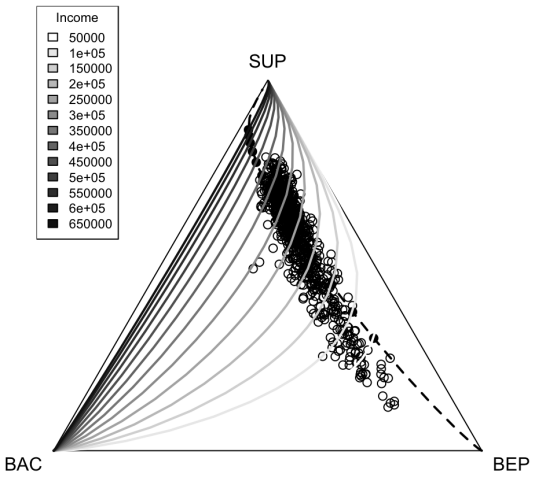


Figure I.5. Courbes d'iso-niveau de y, avec le modèle (ILR-1) dans le sous-simplexe $\widetilde{S}_3$.

3.4 Comparaison des deux régressions, (OLS) et (ILR)

Pour comparer les deux modèles, (OLS) et (ILR) on peut visualiser les prédictions dans le plan $(\boldsymbol{x}_{\mathsf{BEP}}, \boldsymbol{x}_{\mathsf{SUP}})$, comme sur la figure I.6. Sur la partie gauche, on a la régression linéaire standard, et les courbes d'iso-niveau sont parallèles. Sur la partie gauche, on représente les courbes d'iso-niveau de la fonction

$$(\boldsymbol{x}_{\mathsf{BEP}}, \boldsymbol{x}_{\mathsf{SUP}}) \mapsto \widehat{b}_0 + \widehat{b}^\top ILR_e(\boldsymbol{x}_{\mathsf{BEP}}, 1 - \boldsymbol{x}_{\mathsf{BEP}} - \boldsymbol{x}_{\mathsf{SUP}}, \boldsymbol{x}_{\mathsf{SUP}}).$$

À droite de la Figure I.7, on peut voir l'évolution de la prévision en fonction de $\boldsymbol{x} = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%. La valeur de 15% pour les bacheliers a été retenue ici car elle est proche de la valeur moyenne sur toute la ville de Paris. Les points représentés sur le base de la Figure I.7 correspondent au cas où la valeur de BAC est comprise entre 13% et 15%.

Les modèles (OLS-1) et (ILR-1) donnent des prévisions très proches car localement, les deux modèles sont très proches. On peut toutefois s'interroger sur la pertinence d'un modèle linéaire.

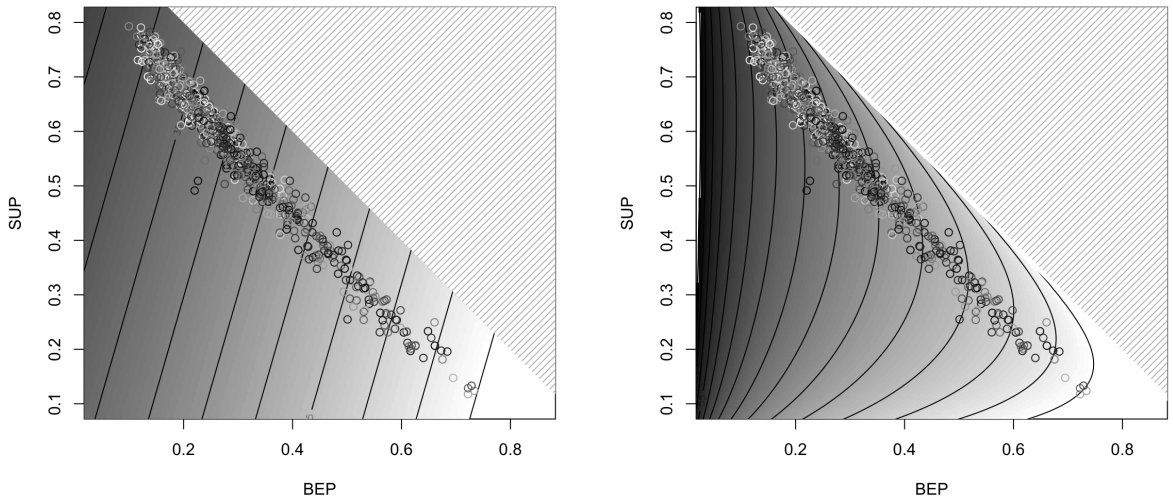


Figure I.6. Courbes d'iso-niveau de y, avec le modèle (OLS-1) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-1) à droite.

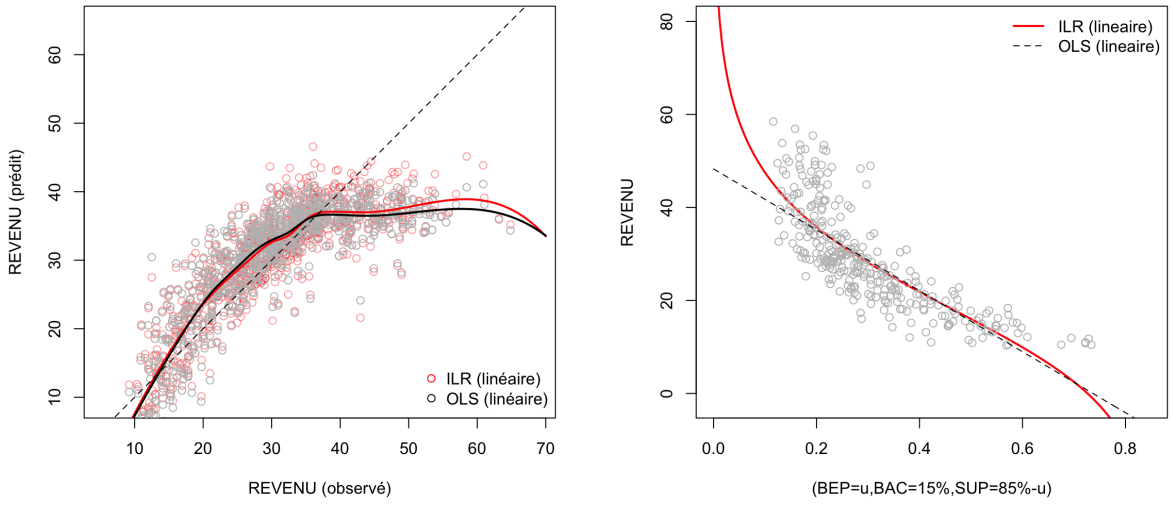


Figure I.7. À gauche, comparaison des prévisions de niveau moyen de revenu, par quartier, avec les modèles (OLS-1) et (ILR-1). À droite, prévision en fonction de $\boldsymbol{x} = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%.

3.5 Version non-linéaire de la régression sur données compositionnelles

Sur notre exemple numérique, comme le montre la partie de gauche de la Figure I.7, le revenu ne semble pas augmenter linéairement avec la proportion de diplômés de l'enseignement supérieur, ou décroitre linéairement avec la proportion de titulaire d'un brevet des collèges. On peut alors naturellement tenter des transformations de type Box-Cox (décrites dans Box and Cox (1964)). En utilisant le test du rapport de vraisemblance, on peut en déduire un intervalle de confiance pour la valeur optimale de λ, comme à droite de la Figure I.8. Pour le linéaire standard, une valeur dans l'intervalle $[-0.028, 0.205]$ est suggérée, alors que pour le modèle (ILR), la valeur optimale serait dans $[0.084; 0.317]$. On notera que ce dernier ne contient pas 0, correspondant à la transformation logarithmique de la variable dépendante. On peut néanmoins regarder ce que donnerait un modèle sur le logarithme du revenu médian.

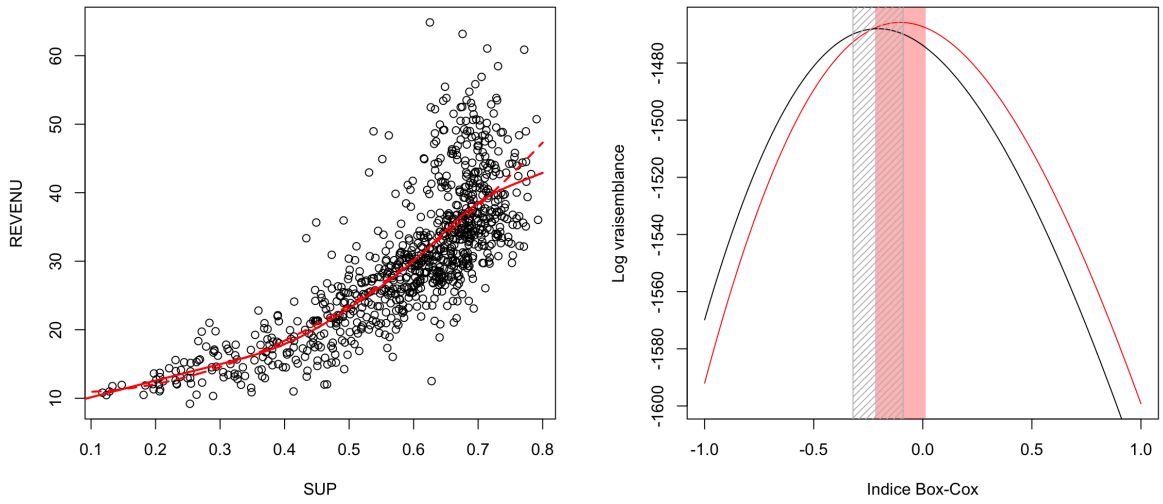


Figure I.8. Évolution du revenu moyen en fonction de la proportion de diplômés de l'enseignement supérieur, par quartier, à gauche, et log-vraisemblance profilée du modèle de Box-Cox à droite.

La version logarithmique de (OLS-1) s'écrit

$$\log y_i = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} + \eta, \tag{OLS-2}$$

avec $\text{Var}[\eta] = \sigma^2$, de telle sorte que la prévision obtenue pour un $\boldsymbol{x}$ donné s'écrit

$$\widehat{y} = \exp\left[\widehat{\beta}_0 + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} + \frac{\widehat{\sigma}^2}{2}\right].$$

La version logarithmique de (ILR-1) sera tout simplement

$$\log y_i = b_0 + \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_i)^\top \boldsymbol{b} + \varepsilon_i \qquad \text{(ILR-2)}$$

Les courbes d'iso-niveau et les prévisions des modèles OLS-2 et ILR-2 sont représentées sur les figures I.9 et I.10 et les résultats dans le tableau I.2.

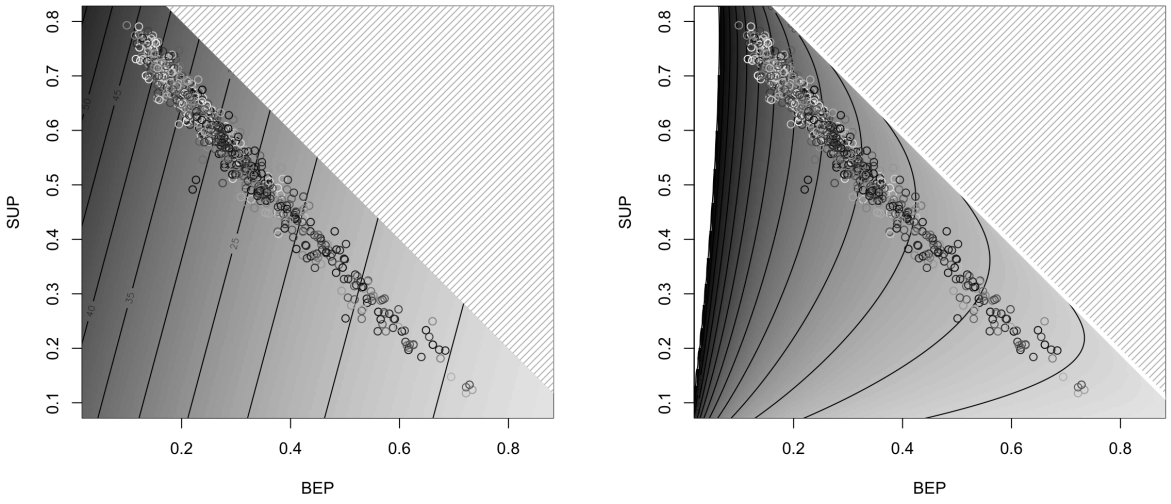


Figure I.9. Courbes d'iso-niveau de y, avec le modèle (OLS-2) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-2) à droite.

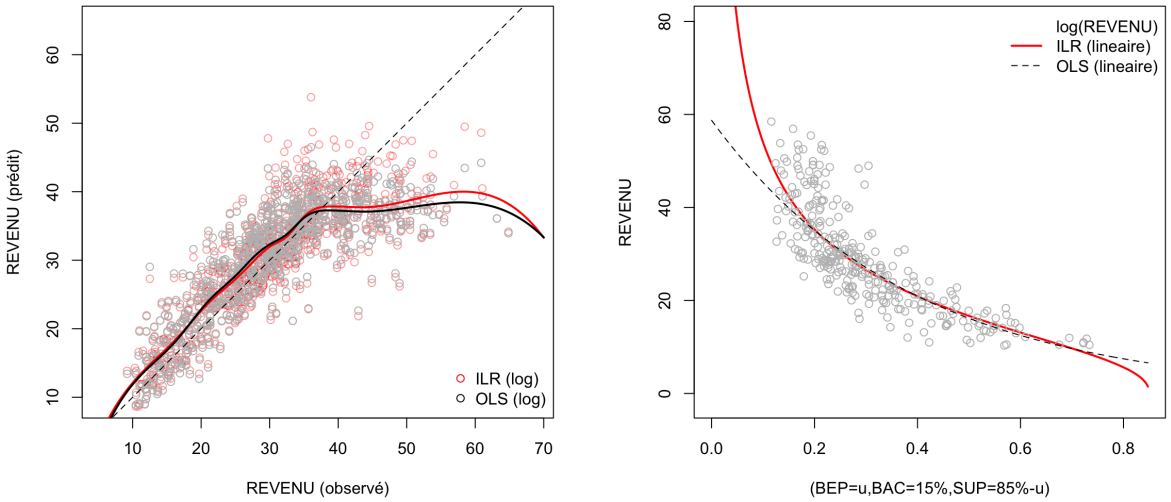


Figure I.10. À droite, prévision en fonction de $\boldsymbol{x} = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%, et à gauche, comparaison des prévisions avec les deux modèles (OLS-2) et (ILR-2) sur le logarithme du revenu.

Table I.2. Régression sur une transformation logarithmique : revenu (en log) en fonction du niveau d'étude.

	(OLS-2) Revenu (log)		(ILR-2) Revenu (log)
$\boldsymbol{x}_{\mathsf{BAC}}$	1.49^{***}	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1$	0.26^{***}
	(0.21)		(0.04)
$\boldsymbol{x}_{\mathsf{SUP}}$	3.02^{***}	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_2$	0.52^{***}
	(0.09)		(0.02)
Constante	0.58^{***}	Constante	-0.94^{***}
	(0.01)		(0.11)
Observations	868		868
R^2	0.749		0.759
R^2 ajusté	0.749		0.759
$\widehat{\sigma}$	0.186		0.182
Statistique F	1284^{***}		1357^{***}

Une alternative usuelle pour introduire un effet non-linéaire est de prendre rajouter une version quadratique de la variable explicative. Ici, la version quadratique de (OLS-1) s'écrit

$$y_i = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{x}^\top \boldsymbol{B}\ \boldsymbol{x} + \eta_i \tag{OLS-3}$$

On peut également transformer (ILR-1), pour introduire une composante quadratique, de la forme $y_i = \beta_0 + \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \langle \boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{x}_i \rangle + \varepsilon_i$, avec une matrice carrée symmétrique $\boldsymbol{B}$, soit

$$y_i = b_0 + \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_i)^\top \boldsymbol{b} + \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_i)^\top \boldsymbol{B}\ \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}) + \varepsilon_i \tag{ILR-3}$$

Les courbes d'iso-niveau et les prévisions des modèles OLS-3 et ILR-3 sont représentées sur les figures I.11 et I.12 et les résultats dans le tableau I.3.

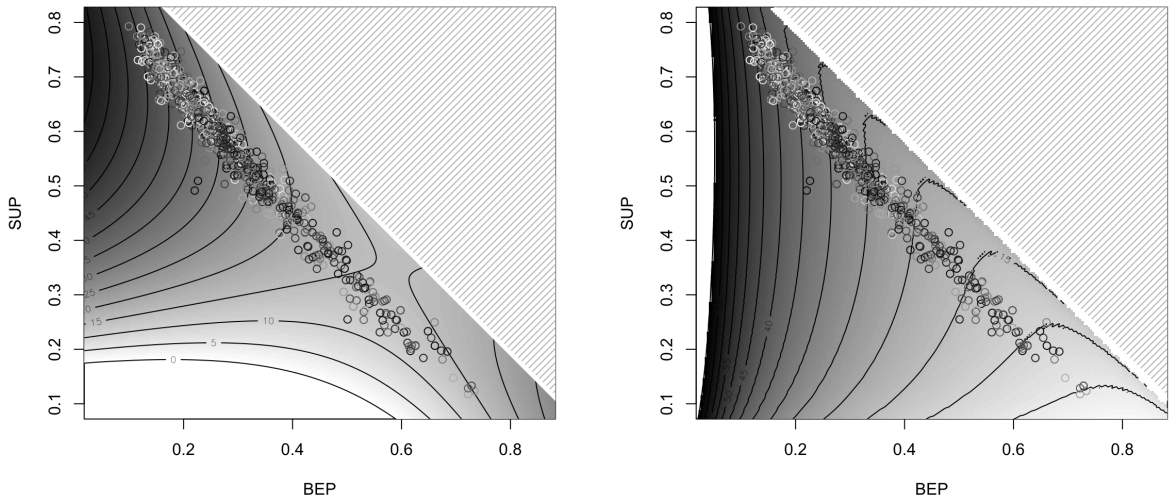


Figure I.11. Courbes d'iso-niveau de y, avec le modèle (OLS-3) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-3) à droite.

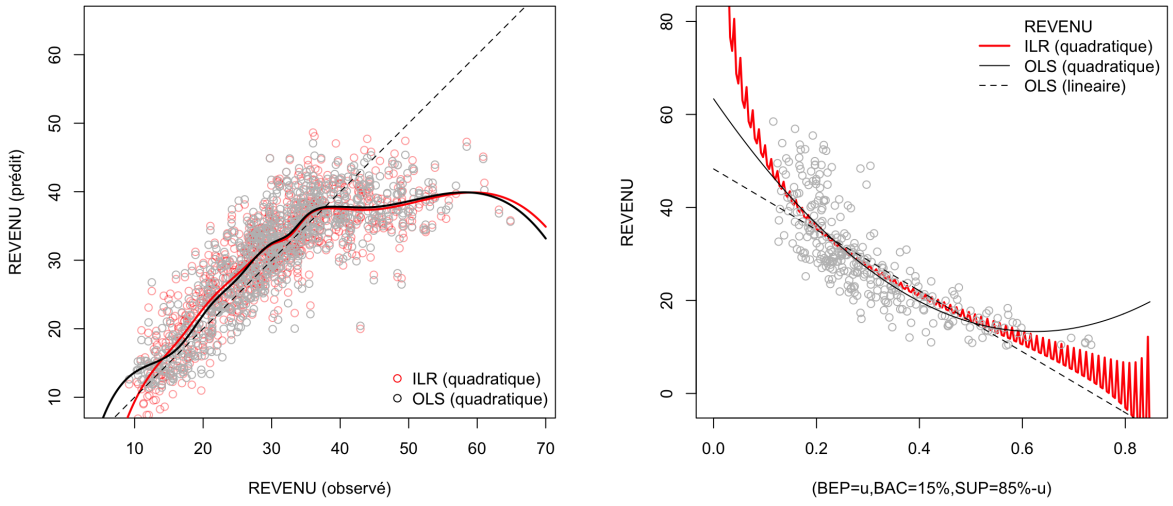


Figure I.12. À droite, prévision en fonction de $\boldsymbol{x} = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%, et à gauche comparaison des prévisions avec les deux modèles (OLS-3) et (ILR-3) avec une transformation quadratique de x.

Table I.3. Régression quadratique : revenu en fonction du niveau d'étude.

	(OLS-3) Revenu		(ILR-3) Revenu
$\boldsymbol{x}_{\mathsf{BEP}}$	19.441***	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1$	0.162
	(3.742)		(0.163)
$\boldsymbol{x}_{\mathsf{SUP}}$	10.800***	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_2$	1.005***
	(2.724)		(0.069)
$\boldsymbol{x}^2_{\mathsf{BEP}}$	-7.621	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1^2$	0.241***
	(5.231)		(0.058)
$\boldsymbol{x}^2_{\mathsf{SUP}}$	6.182**	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_2^2$	0.292***
	(1.908)		(0.051)
$\boldsymbol{x}^2_{\mathsf{BAC}}$	15.890 ***	$\mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1(\boldsymbol{x})_2$	0.239***
	(2.751)		(0.031)
Constante	-10.142***	Constante	1.835***
	(1.364)		(0.048)
Observations	868		868
R^2	0.733		0.740
R^2 ajusté	0.731		0.739
$\widehat{\sigma}$	0.534		0.527
Statistique F	471.8***		490***

3.6 Propriétés des estimateurs

La matrice ILR($\boldsymbol{X}$) est une matrice $n \times (d-1)$, où d est le nombre de modalité du facteur x. En pratique, certaines modalités peuvent être regroupées, car elles ne sont pas significativement différentes, comme le montrent Kaufman and Rousseeuw (2009) ou Bondell and Reich (2009). En fait, si les groupes sont constitués aléatoirement, indépendamment de y et de x, alors le rang[11] de ILR($\boldsymbol{X}$) correspond au nombre de modalités "*réel*". Autrement dit, si x est une variable à 8 modalités mais que le rang de ILR($\boldsymbol{X}$) est 5, c'est que 3 modalités peuvent être fusionnées entre elles, pour constituer au final 5 "*vraies*" catégorie (une étant la référence).

3.7 Revenu et taille du logement

Dans les données de l'INSEE, on peut connaître la proportion de personnes habitant un logement (résidence principale) de moins de 40m^2, entre 40m^2 et 100m^2, et plus de 100m^2 (Figure I.13). Au même titre que pour les classes de diplômes, il est possible de considérer une régression du revenu dans la zone y_i et x_i la variable compositionnelle indiquant la proportion de personnes résidant dans les différents classes de taille de logement (voir Table I.4).

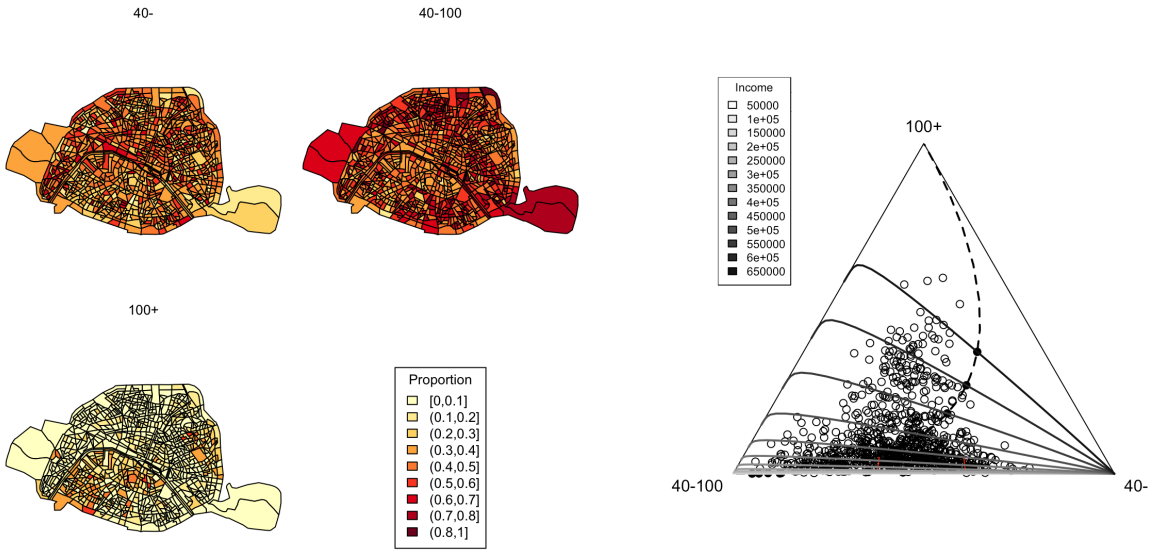


Figure I.13. Variable x_2 correspondant à la proportion de personnes dans le quartier habitant un logement (résidence principale) de moins de 40m^2, entre 40m^2 et 100m^2, et plus de 100m^2. La Figure de droite est la régression linéaire ILR (Table I.4).

[11] Comme classiquement en économétrie, le "rang" de la matrice $\boldsymbol{X}$ est en fait le rang de la matrice $\boldsymbol{X}^\top\boldsymbol{X}$

Table I.4. Régression linéaire : revenu en fonction de la superficie du logement.

	(OLS) Revenu		(ILR) Revenu
$\boldsymbol{x}_{40\text{-}}$	11.056***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_1$	4.790***
	(1.665)		(0.477)
$\boldsymbol{x}_{100+}$	82.309***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x})_2$	9.001***
	(2.088)		(0.238)
Constante	17.669	Constante	45.093
	(0.757)		(0.434)
Observations	868		868
R^2	0.648		0.637
R^2 ajusté	6.132		0.636
$\widehat{\sigma}$	0.594		6.224
statistique F	793.3***		757.1***

La Figure I.14 montre les courbes d'iso-niveau du revenu, pour les deux modèles, dans le plan (moins de 40m^2, plus de 100m^2). Une section de coupe, lorsque la proportion de personnes ayant un logement entre 40m^2 et 100m^2 est de l'ordre de 50%, est présenté sur la Figure I.15.

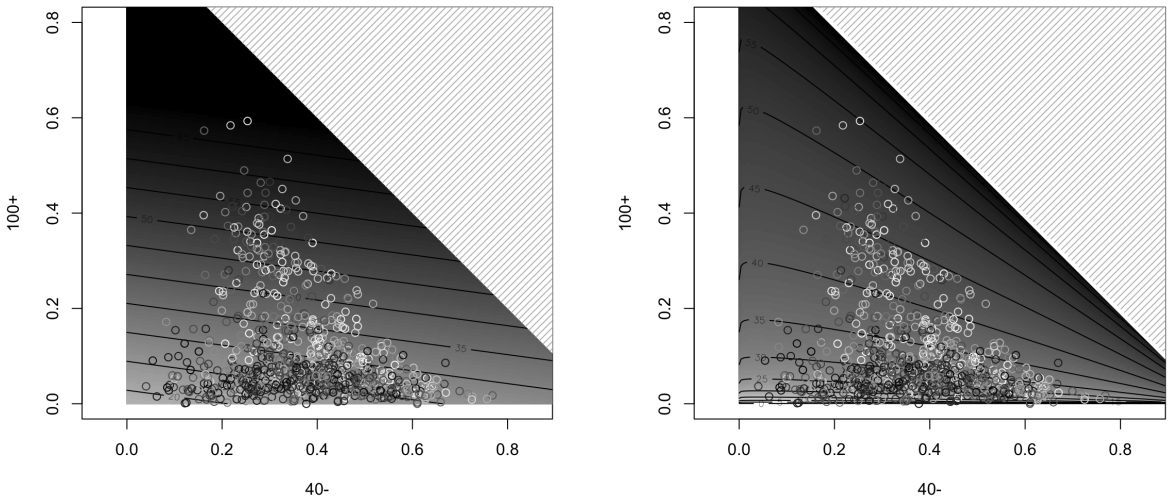


Figure I.14. Courbes d'iso-niveau du revenu y, avec le modèle (OLS) dans le plan (moins de 40m^2, plus de 100m^2), de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR) à droite.

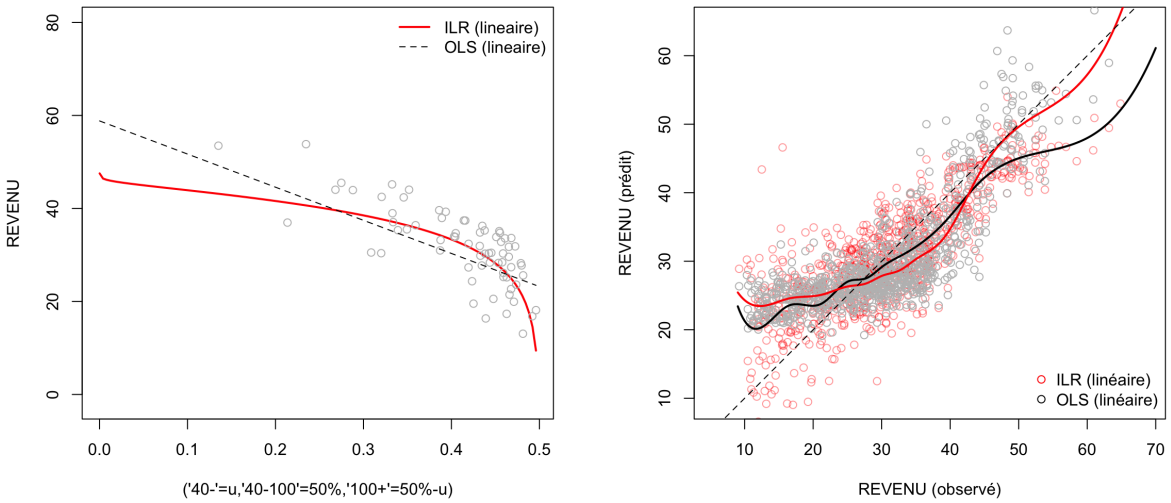


Figure I.15. À gauche, prévision du revenu en fonction de $x = (u, 50\%, 50\% - u)$, pour u variant de 0% à 50%, et à droite comparaison des prévisions avec les deux modèles (OLS) et (ILR).

3.8 Régression sur plusieurs variables compositionnelles

De la même manière qu'il est possible de passer d'une régression simple à la régression multiple, on peut régresser sur plusieurs variables compositionnelles. On posera ainsi

$$y_i = b_0 + \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_{1,i})^\top \boldsymbol{b}_1 + \cdots + \mathrm{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_{k_i})^\top \boldsymbol{b}_k + \varepsilon_i$$

Le tableau I.5 représente les résultats des régressions (OLS) et (ILR) du revenu médian dans la zone i en fonction des deux variables compositionnelles présentées précédemment : le niveau d'étude et la taille de logement.

Table I.5. Régression sur deux variables explicatives : revenu en fonction du niveau d'étude et de la taille du logement.

	(OLS) Revenu		(ILR) Revenu
$\boldsymbol{x}_{1,\text{BEP}}$	-9.978$^{\cdot}$	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_1)_1$	7.013***
	(5.754)		(0.967)
$\boldsymbol{x}_{1,\text{SUP}}$	34.956***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_1)_2$	9.240***
	(5.146)		(0.580)
$\boldsymbol{x}_{2,\text{40-}}$	-6.001***	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_2)_1$	0.203
	(1.096)		(0.411)
$\boldsymbol{x}_{2,\text{100+}}$	51.115**	$\text{ILR}_{\boldsymbol{e}}(\boldsymbol{x}_2)_2$	5.482***
	(1.499)		(0.231)
Constante	9.650*	Constante	33.404***
	(4.573)		(0.976)
Observations	868		868
R^2	0.8728		0.7898
R^2 ajusté	0.8722		0.7888
$\widehat{\sigma}$	3.691		4.744
statistique F	1475***		807.8***

4 Corrélation des Variables x, y et z

Comme nous l'avons mentionné dans l'introduction, il convient de faire très attention lors de l'agrégation, en particulier il est nécessaire de comprendre le rôle joué par la corrélation entre les trois variables y (la variable dépendante), x (la variable explicative) et z (la variable d'agrégation). Shively (1969) a été un des premiers à souligner ce point, et à énumérer quelques solutions et la recherche a beaucoup travaillé sous l'hypothèse où la variable d'intérêt, y est catégorielle. En science politique par exemple, on voudra étudier le vote en fonction de la richesse, mais en disposant de données agrégées par bureau de votes, ou au mieux, par urne. La richesse est alors la richesse moyenne par quartier. Gelman (2009) revient ainsi longuement sur le paradoxe qui fait que les gens votent majoritairement républicain, aux États-Unis, dans les quartiers "pauvres".

En effet, dans le cas où la variable d'intérêt est binaire, de nombreuses techniques ont été développées pour inférer des probabilités au niveau individuel. Considérons un individu $i = 1, \ldots, n_j$ dans une zone $j = 1, \ldots, m$, notons $y_{i:j} \in \{0, 1\}$ la variable d'intérêt et $\boldsymbol{x}_{i:j}$ ses

caractéristiques. Il y a alors $p_{i:j}$ la probabilité (individuelle) d'avoir $y_{i:j} = 1$, qui dépendra des caractéristiques $\boldsymbol{x}_{i:j}$ tel que

$$p_{i:j} = g(\boldsymbol{x}_{i:j}, \alpha)$$

avec g une fonction de lien (exponentielle, linéaire, logit, etc.) et α un paramètre. Par exemple, si la fonction de lien est logit et avec une seule caractéristique $x_{i:j}$, on aura alors $\text{logit}(p_{i:j}) = \alpha_j + \alpha_2 x_{i:j}$.

Au niveau agrégé, nous n'observons que les variables groupées de résultats et de caractéristiques, c'est-à-dire le nombre $\overline{y}_j$ de résultats observés dans la zone j qui suit une binomiale de paramètres n_j et $\overline{p}_j$. La probabilité d'être exposé $\overline{p}_j$ dans la zone j est alors

$$\overline{p}_j = \int p_{i:j}(x) f_j(x) dx$$

où $f_j(x)$ est la distribution jointe de x dans la zone j. S'il existe qu'une seule caractéristique binaire $x_{i:j}$, la variable groupée est alors une proportion ϕ_j d'une caractéristique dans la zone j et la probabilité d'être exposé $\overline{p}_j$ devient une somme

$$\begin{aligned}\overline{p}_j &= \frac{1}{n_j} \sum_i p_{i:j}(x_{i:j} = 1)\mathbb{1}(x_{i:j} = 1) + p_{i:j}(x_{i:j} = 0)\mathbb{1}(x_{i:j} = 0) \\ &= \text{expit}(\alpha_j + \alpha_2)\phi_j + \text{expit}(\alpha_j)(1 - \phi_j) = \beta_j^1 \phi_j + \beta_j^0 (1 - \phi_j)\end{aligned}$$

Un problème survient ici, nous avons m observations et nous cherchons à estimer $2m$ coefficients. Goodman (1953) suggère que les coefficients ne dépendent pas de la zone, $\beta_j^1 = \beta^1$ et $\beta_j^0 = \beta^0$ alors l'équation revient à

$$\overline{p}_j = \beta^1 \phi_j + \beta^0 (1 - \phi_j) \tag{I.4}$$

Freedman et al. (1991) propose une alternative avec son "*neighborhood model*". Il suppose qu'au sein d'une zone, il n'y a pas de différence systématique de résultats entre les deux groupes. L'idée est en effet de supposer que les personnes vivant à proximité les unes des autres sont proches en termes de caractéristiques. Duncan and Davis (1953) proposent la méthode des bornes. L'idée est de borner les coefficients, le maximum de la probabilité d'être exposé sachant que $x_{i:j} = 1$ (β_j^1) est atteint lorsqu'aucun des $x_{i:j} = 0$ n'est exposé ($\beta_j^0 = 0$), soit $\overline{p}_j = \beta_j^1 \phi_j$ et le minimum est atteint quand tous les $x_{i:j} = 0$ sont exposés ($\beta_j^0 = 1$), soit $\overline{p}_j = \beta_j^1 \phi_j + (1 - \phi_j)$. Les bornes

pour β_j^1 et β_j^0 sont alors

$$\max\left\{0; \frac{\overline{p}_j - (1-\phi_j)}{\phi_j}\right\} \leq \beta_j^1 \leq \min\left\{1; \frac{\overline{p}_j}{\phi_j}\right\}$$

$$\max\left\{0; \frac{\overline{p}_j - \phi_j}{1-\phi_j}\right\} \leq \beta_j^0 \leq \min\left\{1; \frac{\overline{p}_j}{1-\phi_j}\right\}$$

Le problème de la régression de Goodman est l'absence de contraintes sur les paramètres. En effet, il n'est pas garanti que les coefficients estimés soient compris entre 0 et 1. King (1997) propose une avancé dans le problème écologique en combinant l'information de la régression de Goodman et celle de la méthode des bornes. Les coefficients β_j^1 et β_j^0 sont liés par une *tomography line* dans la carré unité

$$\beta_j^0 = \frac{\overline{p}_j}{1-\phi_j} - \frac{\phi_j}{1-\phi_j}\beta_j^1$$

Il suppose que les coefficients β_j^1 et β_j^0 se trouvent alors dans un seul cluster généré par une loi normale tronquée bivariée, sous contrainte d'absence d'autocorrélation spatiale et d'absence de biais d'agrégation et estime les coefficients par maximum de vraisemblance.

S'il existe de nombreuses solutions pour inférer des comportements à partir de données individuelles dans le cas de variables catégorielles, l'extension au cas continu n'offre pas de solution satisfaisante. Mais elle est indipensable en économie, par exemple pour analyser des données sensibles comme le revenu.

5 Conclusion

Le Réglement Générale sur la Protection des Données et les différents dispositifs mis en place pour la protection des données personnelles tendent à limiter l'accès aux données individuelles. Les instituts statistiques, pour se conformer aux réglementations, sont encouragés à mettre à disposition des données anonymisées et fournissent le plus souvent des données agrégées. L'information disponible suite à l'agrégation est diluée et des ajustements dans les modèles statistiques et économétriques sont nécessaires afin de tenir compte de leurs particularités.

Une donnée factorielle ou catégorielle, lorsqu'elle est agrégée, est alors un vecteur de proportions sommant à 1. On parle ici de compositions. Ces données compositionnelles ne fournissent qu'une information relative de la réalité. Aitchison (1986), et ses travaux sur la géométrie du

simplexe, propose l'utilisation de log-ratios pour analyser proprement ces données.

Dans cet article, nous nous intéressons plus particulièrement à la régression linéaire avec des variables explicatives compositionnelles. Dans ces modèles, l'idée principale est de travailler en coordonnées, c'est-à-dire de transformer les variables compositionnelles à l'aide de log-ratios avant de les inclure dans le modèle. Plusieurs modèles sont proposés : un modèle en niveau, un modèle logarithmique, un modèle quadratique ainsi qu'un modèle avec plusieurs compositions.

Les modèles classiques et les modèles sur des données transformées donnent des prévisions proches. Néanmoins, on peut s'interroger sur la pertinence d'un modèle linéaire. En effet, comprendre la nature particulière des compositions permet une interprétation plus cohérente et intuitive des résultats. Inclure des données compositionnelles sans tenir compte de leur nature entraîne une interprétation non pertinente des résultats due notamment à la multicolinéarité des composantes et la violation de la clause *ceteris paribus*. Classiquement, on comparera toujours à la valeur de référence et l'utilisation de log-ratios permet de s'affranchir de cette contrainte. L'application de modèles en log-ratios s'avère ainsi plus appropriée dans le cadre des données compositionnelles.

Cette note méthodologique permet d'appréhender les particularités des compositions et de détailler les dispositions et précautions à prendre lors de l'utilisation de telles données dans un modèle linéaire.

6 Annexes : Aspects computationnels

Plusieurs librairies sous R proposent des fonctions[12] pour faire de l'analyse de données compositionelles, y compris des régressions, comme en atteste l'ouvrage de Van den Boogaart and Tolosana-Delgado (2013). On peut par exemple utiliser la librarie `compositions`

```
> library(compositions)
```

Soit une variable compositionnelle X avec comme composantes X_1, X_2 et X_3, la fonction `acomp()` permet de définir la classe de X en composition.

```
> X_composition <- acomp(X)
```

[12] Les codes complets pour reproduire l'analyse sont en ligne sur https://github.com/freakonometrics/compositions.

Dans le cas d'une composition à trois composantes, il est alors possible de la représenter graphiquement sous forme de diagramme ternaire.

```
> plot(X_composition)
```

La librairie de fonction `compositions` permet aussi de calculer les différentes transformations en log-ratios ainsi que leurs transformations inverses.

```
> alr_X <- alr(X_composition)
> alrInv(alr_X)
> clr_X <- clr(X_composition)
> clrInv(clr_X)
> ilr_X <- ilr(X_composition)
> ilrInv(ilr_X)
```

Soit une variable Y qui peut être expliquée par la composition X, on peut réaliser une régression de Y sur la transformation isométrique de X à l'aide de la fonction `lm()`.

```
>  reg_ilr <- lm(Y~ilr(X_composition))
>  summary(reg_ilr)
```

Il ne reste alors plus qu'à transformer les paramètres estimés grâce à la transformation isométrique inverse.

```
b0 = coef(reg_ilr)[1]
b1 = ilrInv(coef(reg_ilr)[-1],orig=X_composition)
```

Bibliographie

J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman and Hall London, 1986.

J. A. Berlin, J. Santanna, C. H. Schmid, L. A. Szczech, and H. I. Feldman. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers : ecological bias rears its ugly head. *Statistics in medicine*, 21(3) :371–387, 2002.

N. Best, S. Cockings, J. Bennett, J. Wakefield, and P. Elliott. Ecological regression analysis of environmental benzene exposure and childhood leukaemia : sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 164(1) :155–174, 2001.

H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in anova. *Biometrics*, 65(1) :169–177, 2009.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society : Series B (Methodological)*, 26(2) :211–243, 1964.

F. Chayes. On correlation between variables of constant sum. *Journal of Geophysical research*, 65(12) :4185–4193, 1960.

W. A. Clark and K. L. Avery. The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4) :428–438, 1976.

C. A. G. Crawford and L. J. Young. A spatial view of the ecological inference problem. *Ecological Inference*, page 233, 2004.

U. Deichmann, D. Balk, and G. Yetman. Transforming population data for interdisciplinary usages : from census to grid. *Washington (DC) : Center for International Earth Science Information Network*, 200(1), 2001.

O. D. Duncan and B. Davis. An alternative to ecological correlation. *American sociological review*, 1953.

J. J. Egozcue and V. Pawlowsky-Glahn. Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, 264(1) :145–159, 2006.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3) :279–300, 2003.

K. El Emam, A. Brown, and P. AbdelMalik. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *Journal of the American Medical Informatics Association*, 16(2) :256–266, 2009.

M. Elff. Social divisions, party positions, and electoral behaviour. *Electoral Studies*, 28(2) : 297–308, 2009.

P. Filzmoser and K. Hron. Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3) :233–248, 2008.

A. S. Fotheringham and D. W. Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7) :1025–1044, 1991.

D. A. Freedman, S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. Ecological regression and voting rights. *Evaluation Review*, 15(6) :673–711, 1991.

T. Fry. Applications in economics. *V. Pawlowsky-Glahn, and A. Buccianti*, 2011.

C. E. Gehlke and K. Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A) :169–170, 1934.

A. Gelman. *Red State, Blue State, Rich State, Poor State : Why Americans Vote the Way They Do-Expanded Edition*. Princeton University Press, 2009.

L. A. Goodman. Ecological regressions and behavior of individuals. *American sociological review*, 1953.

C. Grasland, M. Madelin, et al. The modifiable area unit problem. *Final Report of ESPON*, 3 (3) :2000–2006, 2006.

S. Greenland. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International journal of epidemiology*, 30(6) :1343–1350, 2001.

M. T. Hannan and L. Burstein. Estimation from grouped observations. *American Sociological Review*, pages 374–392, 1974.

D. Holt, D. Steel, M. Tranmer, and N. Wrigley. Aggregation and ecological effects in geographically based data. *Geographical analysis*, 28(3) :244–261, 1996.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

G. King. *A solution to the ecological inference problem : Reconstructing individual behavior from aggregate data*. Princeton University Press, 1997.

A. Klima, T. Schlesinger, P. W. Thurner, and H. Küchenhoff. Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research*, 48(2) : 296–325, 2019.

W. Lin, P. Shi, R. Feng, and H. Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4) :785–797, 2014.

V. Loonis and M.-P. Bellefon. *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R.*, volume 131. Insee, 2018.

A. Loth. Risques de ré-identification dans les bases de données de santé, moyens de s'en prémunir : un projet de loi conciliant ouverture et protection. (64) :7–18, 2005.

G. J. Matthews, O. Harel, et al. Data confidentiality : A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5 :1–29, 2011.

J. Morais. *Impact of media investments on brands' market shares : a compositional data analysis approach*. PhD thesis, 2017.

S. Openshaw. Ecological fallacies and the analysis of areal census data. *Environment and planning A*, 16(1) :17–31, 1984a.

S. Openshaw. The modifiable areal unit problem. *CATMOG - Concepts and Techniques in Modern Geography*, 38, 1984b.

S. Openshaw and L. Rao. Algorithms for reengineering 1991 census geography. *Environment and planning A*, 27(3) :425–446, 1995.

S. Openshaw and P. Taylor. Statistical applications in the spatial sciences, chapter a million or so correlation coefficients : three experiments on the modifiable areal unit problem. *Wrigley N. Publishers, London, Pion*, pages 127–144, 1979.

V. Pawlowsky-Glahn and A. Buccianti. *Compositional data analysis : Theory and applications*. John Wiley & Sons, 2011.

J. Pearl and D. Mackenzie. *The book of why : the new science of cause and effect*. Basic Books, 2018.

K. Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367) :489–498, 1897.

W. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 1950.

G. Ronning. Share equations in econometrics : A story of repression, trustration and dead ends. *Statistical papers*, 33(1) :307, 1992.

S. Schwartz. The fallacy of the ecological fallacy : the potential misuse of a concept and the consequences. *American journal of public health*, 84(5) :819–824, 1994.

W. P. Shively. "Ecological" inference : the use of aggregate data to study individuals. *The American Political Science Review*, 63(4) :1183–1196, 1969.

P. J. Smith. *A Selective Review of Confidentiality Research Published Since 1975*. Bureau of the Census, 1985.

L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :571–588, 2002.

H. Theil. Linear aggregation of economic relations. 1954.

H. Theil. A multinomial extension of the linear logit model. *International economic review*, 10 (3) :251–259, 1969.

K. G. Van den Boogaart and R. Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

L. K. VanWey, R. R. Rindfuss, M. P. Gutmann, B. Entwisle, and D. L. Balk. Confidentiality and spatially explicit data : Concerns and challenges. *Proceedings of the National Academy of Sciences*, 102(43) :15337–15342, 2005.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2) : 158–183, 2007.

J. Wakefield and H. Lyons. Spatial aggregation and the ecological fallacy. *Handbook of spatial statistics*, pages 541–558, 2010.

D. West. *Ternary equilibrium diagrams*. Springer Science & Business Media, 2012.

D. W. Wong. The modifiable areal unit problem (maup). In *WorldMinds : geographical perspectives on 100 problems*, pages 571–575. Springer, 2004.

A. D. Woodland. Stochastic specification and the estimation of share equations. *Journal of Econometrics*, 10(3) :361–383, 1979.

CHAPTER II

Estimating Inequality Measures from Quantile Data

1 Introduction

A suitable and widely used approach to depicting economic inequalities is to provide indices to measure the degree of inequality. These measures enable comparisons of living standards between different countries, regions or among time. Descriptive measures could be useful in understanding economic relationship (Kaplow, 2005). In this regard, several indices have been described in the literature, including the Gini coefficient, the Pietra index or the Theil indexes. Similarly, the Lorenz curve is a relevant indicator of income distribution and most indices are related to it. The modelling of a Lorenz curve then provides a useful tool. Two main strategies have been developed for modelling a Lorenz curve, either by approximating the empirical Lorenz curve or either by modelling an income distribution and deriving the Lorenz curve from it.

The first is focused on the parametric approximation of the size distribution of income. Based on this estimation, a Lorenz curve and the several indices can be derived (see Champernowne and Cowell (1998) for a survey). Several relevant functional forms have been used to describe the income distribution (see Chotikapanich (2008) for an exhaustive list). The most popular forms are the so-called generalized distribution of the second kind (GB2) proposed by McDonald (1984) and its special and limited cases. Special and limited cases include Pareto distributions, Log-normal distributions, Gamma-type size distributions (Gamma, Generalized Gamma, Weibull) and Beta-type size distribution (GB2, Singh-Maddala, Dagum, Fisk, Beta distribution of the first and the second kind) (see Kleiber and Kotz (2003) for a survey). Alternatively, an approach is to apply on shares the generalized non-parametric Pareto interpolation technique developed by Blanchet et al. (2017).

The second literature examines the parametric approximation of a Lorenz curve. In the same way, several functional forms have been developed to fit Lorenz curves. The pioneers are Kakwani and Podder (1973) who proposed a functional form consistent with the distribution of Australian data. Many others authors suggested functional forms, in particular, Rasche et al. (1980), Pakes (1981), Gupta (1984), Arnold (1986), Villaseñor and Arnold (1989), Basmann et al. (1990), Ortega et al. (1991), Chotikapanich (1993), Sarabia (1997) or Rohde (2009). Several methods allow the estimation of the functional form parameters. The earlier models are based on linear or non-linear least squares. Castillo et al. (1998) and Sarabia et al. (1999) proposed an alternative approach by considering the median or least median square. More recently, Chotikapanich and Griffiths (2002) developed a maximum likelihood estimator based on a Dirichlet distribution

to capture the cumulative nature of the shares. The selection of the best functional form is subsequently based on an adjustment criterion. Chotikapanich and Griffiths (2005) suggested an alternative way by averaging functional forms using a Bayesian model averaging approach.

However, problems arise due to the lack of income data at the individual level. Data are usually aggregated and provide less information than individual data. Institutions sometimes report inequality indices with aggregate data, although most only the Gini index. In such cases, an interesting approach is to reconstruct the proposed index and provide alternative indices. Analysis is straightforward when individual data are accessible. The income distribution and the Lorenz curve can be estimated directly from their empirical forms. However, personal income data are not widely available. They are usually aggregated, whether census or survey data. Census data appear to be more relevant to provide an overview over time or across sub-regions. Nevertheless, the institutions provide several different forms of available data. At the national or regional scale, data are mostly reported in the form of income shares. Institutions may provide data as class mean income and deciles or quintiles group shares like on World Income Inequality Database (WIID), World Inequality Database (WID) or even World Bank. The available information thus depicts points on the Lorenz curve. For example, the poorest 10% hold 3.6% of total income in France in 2016 (WIID). Many studies focused on estimating parametric Lorenz curves and income distributions based on grouped data where income and population shares and mean income are given. For example, Hajargasht and Griffiths (2020) suggested a minimum distance estimation of Lorenz curves or Griffiths and Hajargasht (2015) a GMM estimation of income distributions. All these methods involve the use of grouped data.

When dealing with smaller areas, the available data may be only quantiles. The income quartiles, quintiles or deciles can be provided but not income shares or class mean income. The available information does not depicts points on the Lorenz curve. For example, the poorest 10% receive less than €10,860 in France in 2016 (INSEE). Unfortunately, their share in the total income is not available. The information given is therefore nearby but not precisely the same. In other words, when focusing on a high geographical level (region or country), income shares can be found, although when dealing with smaller areas such as cities, the data provided are frequently reported as income quantiles.

The purpose of this paper is to determine a methodology for estimating a Lorenz curve and associated indices with only quantile data. The methods need to be adjusted to obtain a Lorenz curve from these quantiles. Quantiles can easily be transformed into tabulated data.

The classic method with tabulated data is to assume the midpoint of each class as the class mean. This assumption is rather strong. For this reason, we develop an innovative method based on conditional expectations to compute class means. We will illustrate this method by measuring inequalities within the city of Paris. French municipalities can be subdivided into narrower areas called iris (Ilots Regroupés pour l'Information Statistique). The iris generally have between 1800 and 5000 inhabitants and are built in relation to large sections of the urban network. The narrowness of the areas limits the data available due to the privacy of personal data and INSEE (Institut National de la Statistique et des Études Économiques) provides only quartiles and deciles of income. This methodology is also applicable to other quantile data such as ZIP code income data in the United States.

In Section 2, we will present the proposed methodology[1] for modelling a Lorenz curve from quantile data. Section 3 will illustrate the method with an application on simulated income data and Parisian iris data.

2 Methods

2.1 Definition of a Lorenz curve

The Lorenz curve, introduced by Lorenz (1905), is the curve defined by the points $(p, L(p))$ where p is the cumulative proportion of the income-receiving units sorted from the poorest to the richest and $L(p)$ the cumulative proportion of income received by these units. Gastwirth (1971) proposed a general definition of the Lorenz curve such that, if X is the income of a member of the population and assumed to be a random variable with cumulative distribution function $F(x)$, quantile function $F^{-1}(x)$ and mean $\mu = \int x dF(x)$, then the Lorenz curve is the mapping

$$p \mapsto L(p) = \frac{1}{\mu} \int_0^p F^{-1}(t) dt \tag{II.1}$$

A Lorenz curve will have some expected properties: (1) if $p = 0$ then $L(p) = 0$, (2) if $p = 1$ then $L(p) = 1$, (3) $L(p) \leq p$ and (4) $L(p)$ is continuous and differentiable and the slope of the curve increases monotonically. Several authors developed functional forms to fit a Lorenz curve (see Table II.1).

[1] A R vignette with codes is available from https://github.com/EnoraBelz/Inequality.

Table II.1. Lorenz curve functional forms

	Functional form $L(p)$	Parameter constraints
Kakwani and Podder (1973)	$p^{\alpha}e^{-\beta(1-p)}$	$\beta > 0, \alpha \geq 1$
Rasche et al. (1980)	$(1-(1-p)^{\alpha})^{\beta}$	$\beta \geq 1, 0 \leq \alpha \leq 1$
Arnold (1986)	$\frac{p(1+(\alpha-1)p)}{1+(\alpha-1)+\beta(1-p)}$	$\alpha, \beta > 0, \alpha - \beta < 1$
Ortega et al. (1991)	$p^{\alpha}(1-(1-p)^{\beta})$	$\alpha \geq 0, 0 < \beta < 1$
Chotikapanich (1993)	$\frac{e^{kp}-1}{e^{k}-1}$	$k > 0$
Sarabia (1997)	$\pi_1 p + \pi_2 p^{\alpha_1} + (1-\pi_1-\pi_2)(1-(1-p)^{\alpha_2})$	$0 \leq \pi_1, \pi_2 \leq 1, \alpha_1 \geq 1, 0 < \alpha_2 < 1$
Rohde (2009)	$p(\frac{\beta-1}{\beta-p})$	$\beta > 1$

2.2 From quantile data to tabulated data

These parametric forms require grouped or individual data which allow the derivation of an empirical Lorenz curve. Quantiles are not sufficient in their current form. Information on the cumulative proportion of the population is given, but the cumulative proportion of income is unknown. Indeed, a quantile indicates the income level such that $p\%$ is below this threshold. The information concerning the total income received by $p\%$ is not available and therefore the share in the total income of $p\%$ is undefined. Nevertheless, quantiles can be transformed into tabulated data. Quantiles become the boundaries of income ranges and the size of the subset becomes the share of individuals. In the case of deciles, the boundaries are the deciles and the proportion of individuals in each range is 10 percent. The use of both quartiles, deciles or other q-quantiles is also feasible. The proportion of individuals depends on being between two same subdivisions or between two different subdivisions.

Suppose X a random variable of income and different associated quantiles Q_k where $0 < k < K$ is the rank of the quantile and s_k the population share below this quantile so that $\mathbb{P}(X < Q_k) = s_k$. The ranges are therefore $[0, Q_1[$ with proportion s_1 for the first range, $[Q_k, Q_{k+1}[$ with proportion $s_{k+1} - s_k$ from the second to the penultimate range and $[Q_K, +\infty[$ with proportion $1 - s_K$ for the last range (see Table II.2 for an example).

Table II.2. From quantile data to tabulated data with quartiles and deciles

From quantile data...

	D_1	D_2	Q_1	D_3	...	D_6	D_7	Q_3	D_8	D_9
	10%	20%	25%	30%		60%	70%	75%	80%	90%

...to tabulated data

$[0; D_1[$	$[D_1; D_2[$	$[D_2; Q_1[$	$[Q_1; D_3[$	$[D_3; D_4[$	...	$[D_6; D_7[$	$[D_7; Q_3[$	$[Q_3; D_8[$	$[D_8; D_9[$	$[D_9; \infty[$
10%	10%	5%	5%	10%		10%	5%	5%	10%	10%

2.3 Income shares

The first step is to determine the total income received by each income bin and thereby the shares of income received. This requires determining the mean income within each income bin. In many studies, the mean income of each class is assumed to be the midpoint of the interval and a Pareto-tail for the open-ended interval (*Midpoint method*). In other words, it means that, somehow, we assume that the income is uniformly distributed in the ranges and consequently the center of mass (i.e. the mean) equals the center of the interval. This assumption is rather strong. Indeed, this assumes that each individual in a group has the same income level and therefore neglects intra-group variations. In addition, some income levels are more predominant in a population (e.g. minimum income, agreement income).

It can be noticed that quantiles enable to fit an income distribution function. Several different functional forms have been developed to model the distribution of income. McDonald (1984) suggested the Generalized Beta distribution of the second kind (GB2) which includes most of the previous functional forms as limited or special cases. For this reason, a GB2 distribution is assumed to describe the size of income and the four parameters are optimizied on quantile data. Alternatively, other forms of income distribution may be used. Suppose X a random variable of income with known density $f(x)$, the conditional expectation of being in a range $[a, b[$ can be determined as

$$E(X|x \in [a,b[) = \frac{\int_a^b xf(x)dx}{\int_a^b f(x)dx} \tag{II.2}$$

where $f(x)$ is the density of X. Therefore, knowing the size distribution of income allows to determine the means of the income per bins. The conditional means between each range can be computed by plugging-in the (four) estimated parameters of the GB2 distribution (*Conditional Expectation method*). Using the class mean income and the class population share , the income share of each bin in the total income can then be determined and consequently the cumulative income shares.

2.4 Functional form optimization

Quantile data become similar to grouped data and classical functional forms of Lorenz curves can be applied. In this way, seven different forms are used : Kakwani and Podder (1973),

Rasche et al. (1980), Arnold (1986), Ortega et al. (1991), Chotikapanich (1993), Sarabia (1997) and Rohde (2009). The parameters are optimized by non-linear least squares (NLS) estimator as

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} (F(p_k, \theta) - s_k)^2 \tag{II.3}$$

where $F(., \theta)$ is the functional form to be optimized according to the parameter(s) θ, p_k is the population share of the k^{th} income range and s_k is the income share of the k^{th} income range. Thereafter, the different functional forms are compared according to a goodness-of-fit measure, especially the value of the Chi-squared statistic

$$\chi^2 = \sum_{k=1}^{K} \frac{(s_k - F(p_k, \hat{\theta}))^2}{F(p_k, \hat{\theta})} \tag{II.4}$$

where $F(., \hat{\theta})$ is the functional form with the optimal parameter(s) $\hat{\theta}$, p_k is the population share of the k^{th} income range and s_k is the income share of the k^{th} income range.

The algorithm is therefore as follows:

- Convert quantiles into tabulated data
- Estimate a GB2 family distribution on quantiles using ML estimation techniques
- Compute conditional expectations of each bin using the fitted distribution
- Calculate income shares and cumulative income shares
- Estimate a functional form on shares using a NLS estimator
- Compare the functional forms with goodness-of-fit measures.

2.5 Inequality measures

The Lorenz curve enables to determine the inequality indices. Gini coefficient G (Gini, 1914) can be evaluated as

$$G = \frac{E(|Y - X|)}{2E(X)} \tag{II.5}$$

where X and Y are i.i.d with common distribution $F(x)$. Alternatively, the Gini coefficient can be determined in terms of the Lorenz curve, corresponding to two times the area between $L(p)$ and the egalitarian line as

$$G = 2 \int_0^1 (p - L(p)) dp \tag{II.6}$$

As a result, Gini coefficients can be derived from the functional forms as a function of parameters (Table II.3).

Table II.3. Gini coefficient of the functional forms

	Gini Index
Kakwani and Podder (1973)	$1 - \frac{2e^{-\beta}}{1+\alpha} {}_1F_1(1+\alpha; 2+\alpha; \beta)$
Rasche et al. (1980)	$1 - \frac{2}{\alpha} B(\frac{1}{\alpha}; \beta + 1)$
Arnold (1986)	$\begin{cases} \frac{\beta}{\beta-\alpha+1} + \frac{2\alpha\beta}{(\beta-\alpha+1)^2}[1 + \frac{\beta+1}{\beta-\alpha+1}\log(\frac{\alpha}{\beta+1})] & \text{if } \beta-\alpha+1 \neq 0 \\ \frac{\beta}{3(1+\beta)} & \text{if } \beta-\alpha+1 = 0 \end{cases}$
Ortega et al. (1991)	$\frac{\alpha-1}{\alpha+1} + 2B(\alpha+1; \beta+1)$
Chotikapanich (1993)	$\frac{(k-2)e^k + (k+2)}{k(e^k-1)}$
Sarabia (1997)	$\pi_2(1 - \frac{2}{1+\alpha_1}) - (1-\pi_1-\pi_2)(1-\frac{2}{1+\alpha_2})$
Rohde (2009)	$2\beta[(\beta-1)\log(\frac{\beta-1}{\beta}) + 1] - 1$

Note: B is the Beta function. ${}_1F_1$ is the confluent hyper-geometric function.

The Pietra index P (Pietra, 1932) is also widely used to measure inequality, it can be defined as

$$P = \frac{E(|X - E(X)|)}{2E(X)} \tag{II.7}$$

The Pietra index can also be defined in terms of the Lorenz curve, it corresponds to the maximum deviation between $L(p)$ and p as

$$P = \max_{p \in [0,1]} \{p - L(p)\} \tag{II.8}$$

In addition, it corresponds to two times the area of the largest triangle that can be inscribed between $L(p)$ and the equalitarian line (Arnold, 2008).

In his book, Theil (1967) also suggests inequality indices, Theil's indexes T_L (low) and T_H (high), based on the generalized entropy (see Cowell (2011) for a discussion on generalized entropy measures).

$$T_L = GE_0 = -E(\log(\frac{X}{\mu})) \tag{II.9}$$

$$T_H = GE_1 = E(\frac{X}{\mu}\log(\frac{X}{\mu})) \tag{II.10}$$

Rohde (2008) relates the concept of Generalized Entropy (GE) to the Lorenz curve and provides

mathematical expressions to define the GE measures in terms of the Lorenz curve.

$$T_L = -\int_0^1 \log(L'(p))dp \tag{II.11}$$

$$T_H = \int_0^1 L'(p)\log(L'(p))dp \tag{II.12}$$

where $L'(p)$ is the first derivative of the Lorenz Curve.

3 Applications

In a first part, we will simulate samples to evaluate the performance of the methodology and compare it to other methods, then, we will use it to get local inequality indices in Paris.

3.1 Simulated data

Simulations are performed both to evaluate the performance of the method and to compare it with the midpoint method. The different methods are applied on income distribution simulations. Several known distributions are selected to perform the simulations such as GB2, Log-normal and Singh-Maddala. The parameters of the distributions are determined to ensure realistic income levels. For each simulation, the income of 2000 individuals is simulated according to a defined distribution. Two areas are considered, one with a low Gini index and one with a high index. Three distributions[2] are used: for the High Gini Index case, $LN(10.6, 1.01)$, $GB2(40000, 1.7, 0.98, 1.02)$ and $SM(30000, 1.9, 0.7)$, respectively with Gini index 0.52, 0.58 and 0.69, and for the Low Gini Index case, $LN(10.5, 0.7)$, $GB2(35000, 2.5, 0.95, 1.02)$ and $SM(50000, 2.2, 1.8)$, respectively with Gini index 0.37, 0.39 and 0.35.

For each of them,

(i) we generate a sample of size 2000 and calculate Gini index from those individual observations directly - called Ind.

(ii) we can get the shares and partial information (quantiles), and then compute Gini index - called Shares

[2]Lognormal distribution $LN(\mu, \sigma)$, GB2 distribution $GB2(\mu, \sigma, \nu, \tau)$ and Singh-Maddala distribution $SM(\mu, \sigma, \nu)$.

we can use only quantiles - and estimate the shares

(iii) the first step of estimating a GB2 distribution allows to compute a Gini index - called First Step

(iv) with the conditional expectation method - called Cond. Expectation

(v) with the midpoint method - called Midpoint.

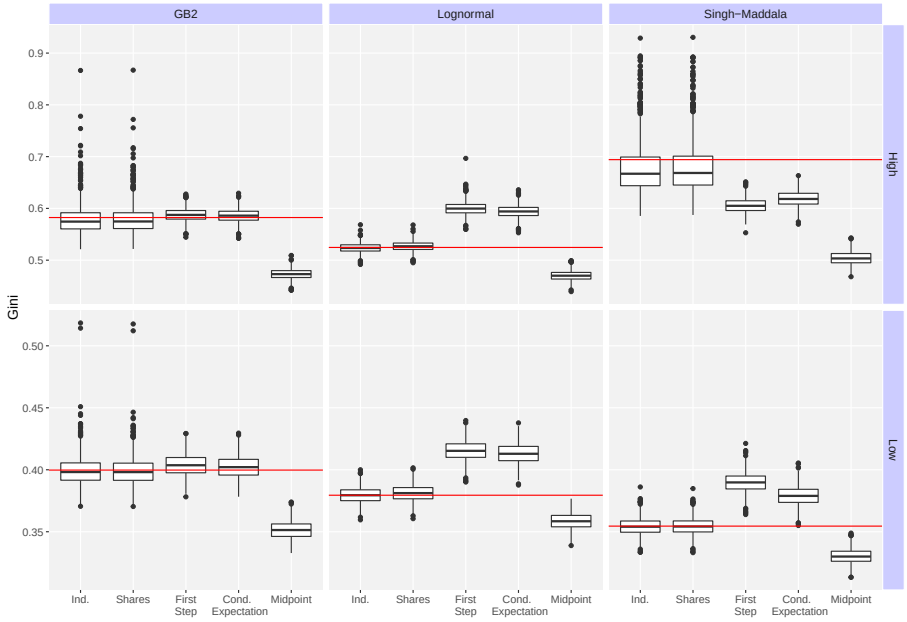


Figure II.1. Comparison of the different methods in terms of estimated Gini

Note: The results are based on 1000 simulations for each cases. The cases are: *Ind.* the Gini on the individual data, *Shares* the Gini estimated on shares, *First Step* the Gini estimated with the GB2 fitting, *Cond. Expectation* the Gini estimated on quantiles with the Conditional Expectation method and *Midpoint* the Gini estimated on quantiles with the Midpoint method. The red line corresponds to the true value of the Gini index of the simulated distribution (obtained by $1,000,000$ simulations). The parameters of the simulated distributions are: *High* $LN(10.6, 1.01)$, $GB2(40000, 1.7, 0.98, 1.02)$ and $SM(30000, 1.9, 0.7)$ and *Low* $LN(10.5, 0.7)$, $GB2(35000, 2.5, 0.95, 1.02)$ and $SM(50000, 2.2, 1.8)$.

Figure II.1 provides a comparison of the different methods in terms of estimated Gini index.

Method (i) is the one with detailed data, with the other four are based on aggregated quantities. The distribution of the Gini estimate based on shares data (ii) is very close to that observed on the individual data (i). Then, when data are in the form of shares, having aggregate or individual data gives a fairly close result. However, if we make as if the shares are not available (as in the dataset we have) and only quantiles are available, the lack of information leads to estimating a Gini index less close to the true value.

Admittedly, the first step of the Conditional Expectation method (iii) also allows to compute a Gini index. This consists of fitting a GB2 distribution on the quantiles to determine condi-

tional expectations, however, indices and the Lorenz curve can be computed using the GB2 distribution. Nevertheless, the underlying income distribution is not necessarily GB2 and, in addition, the GB2 distribution is a four-parameter distribution estimated on only a few points, and can therefore be misspecified, especially on the distribution tail. The second step (iv) should adjust the method by improving the results. The results suggest that the Conditional Expectation method (iv) computes a Gini index close than the GB2 one (iii). When the underlying distribution is GB2, both methods provide the same value close to the true value. When the distribution is not GB2, the results indicate an overestimation of Gini coefficient (or underestimation in the case of Singh-Maddala *High Gini Index*), however, the Gini index of the second step is more accurate. As a result, the Conditional Expectation method adjusts the estimate of Gini index.

Conditional expectation (iv) and Midpoint (v) methods can also be compared. The Midpoint method underestimates the true values in all cases, while the conditional expectation method only underestimates the true values for Singh-Maddala *High Gini Index* and overestimates for Singh-Maddala *Low Gini Index* and Log-normal. If the distribution is GB2 or Singh-Maddala, the Conditional Expectation method significantly outperforms the Midpoint method. When the distribution is Log-normal, the result is less clear. The Conditional Expectation method overestimates the true value while the midpoint method underestimates it. The performance of both methods is then related to the underlying distribution.

3.2 Parisian income data

This paper focuses on the measure of inequality within Paris. The Institut National de la Statistique et des Études Économiques (INSEE) provides income data at a very narrow scale named iris (see Figure II.2 for a visualisation of median income in the iris). The deciles and quartiles of income[3] in these areas are available. Paris includes 966 iris which are either living (861), activity (88) or miscellaneous (17). For some iris with insufficient population (especially activity or miscellaneous iris), income information is not available. Thus, 865 iris are reported with income data. Using these data, a Lorenz curve can then be modelled, the Gini index can be determined and compared with that provided by INSEE, and the Pietra and Theil indices not included can be calculated.

[3]Quantiles of declared household income per consumption unit for the year 2014. The used scale (OECD scale) has the following weighting: 1 UC for the first adult in the household, 0.5 UC for other persons aged 14 or over and 0.3 UC for children under 14 years.

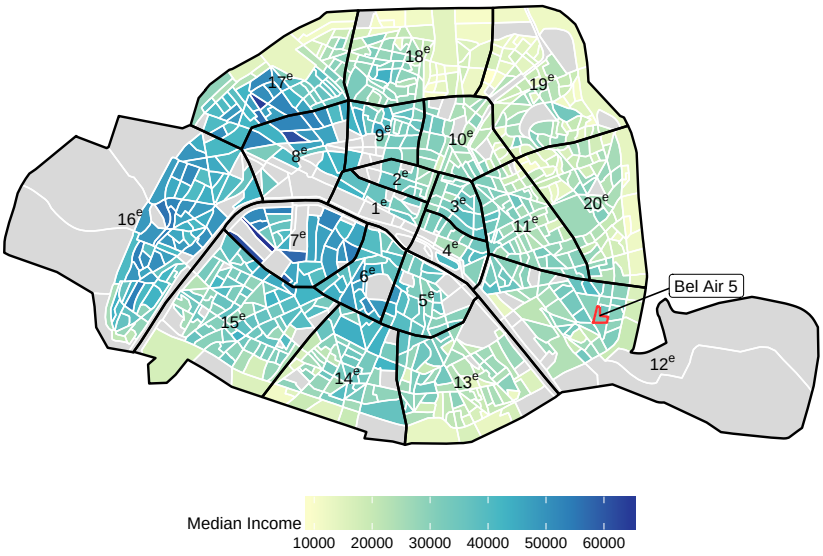


Figure II.2. Median income of the Parisian iris

Application of the methodology: an example with the Bel Air 5 iris

The quantiles provided by INSEE can be used to determine the different income ranges. Table II.4 displays an example[4] of the data obtained for an iris ("Bel Air 5" in the 12th arrondissement). In this area, the first decile is at €10,570, implying that 10% of the population have an income below €10,570. In order to derive income shares, income means within these ranges are required. The method is based on the fitting of a GB2 distribution to compute these conditional means. Therefore, the first step is to determine the parameters of the GB2 distribution by MLE (see dotted line in Figure II.3). The parameters are used to identify the density and the distribution of income in the area. Conditional means of income bins can then be calculated with this distribution (see blue lines in Figure II.3 and column 5 of Table II.4). Those means are used to compute the cumulative shares of income and the cumulative shares of population are the shares between two quantiles (see columns 6 and 7 of Table II.4). In the first range of Bel Air 5, the conditional mean is found to be at €7,403. The cumulative share of income can be derived from the total income (weighted sum of the conditional means). In this area, the poorest 10% of the population receive 2% of the total income. By plotting the cumulative share of income in relation to the cumulative share of the population, the points of the empirical Lorenz curve are observed (Figure II.4).

All alternative functional forms can be applied to the cumulative shares of income and

[4]Further examples are provided in the appendix 5.1.

Table II.4. Tabulated data (€) obtained using the conditional expectation method for Bel Air 5 iris

	Quantile	Income Range	Proportion	Mean Income	Income share	Population share
D_1	10,570	Below 10,570	0.10	7,403.64	0.02	0.10
D_2	18,696	10,570 - 18,696	0.10	14,947.50	0.06	0.20
Q_1	21,558	18,696 - 21,558	0.05	20,135.63	0.08	0.25
D_3	24,344	21,558 - 24,344	0.05	22,948.68	0.12	0.30
D_4	28,298	24,344 - 28,298	0.10	26,296.94	0.18	0.40
Q_2	32,626	28,298 - 32,626	0.10	30,412.82	0.26	0.50
D_6	37,782	32,626 - 37,782	0.10	35,113.38	0.36	0.60
D_7	43,444	37,782 - 43,444	0.10	40,488.39	0.46	0.70
Q_3	46,088	43,444 - 46,088	0.05	44,737.59	0.52	0.75
D_8	50,926	46,088 - 50,926	0.05	48,410.77	0.58	0.80
D_9	61,920	50,926 - 61,920	0.10	55,936.68	0.73	0.90
		61,920 and over	0.10	102,703.30	1	1

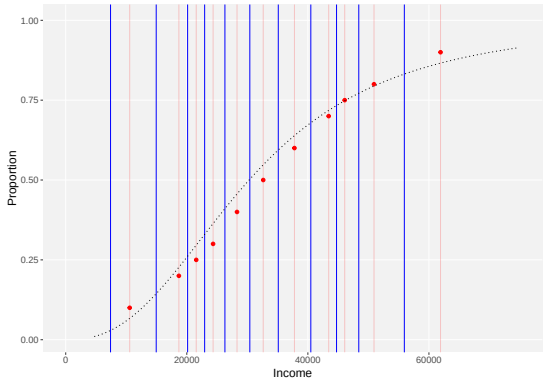


Figure II.3. Estimated means of income per bins of Bel Air 5 iris

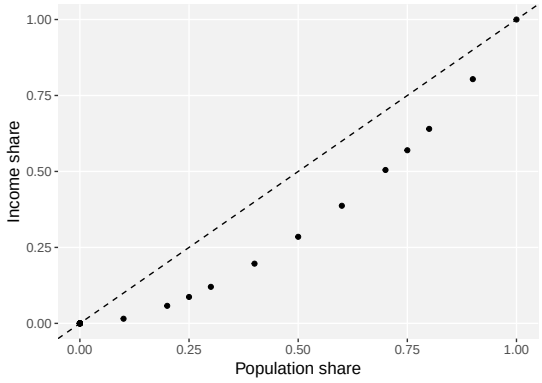


Figure II.4. Empirical Lorenz curve of Bel Air 5 iris

Note: The red lines and dots represent the observed quantiles. The blue lines represent the estimated means between two quantiles. The dotted line is the estimate of the GB2 cumulative distribution function.

population in order to approximate the empirical Lorenz curve. The parameters are optimized by NLS estimator. For each functional form, the most optimal parameters and Lorenz curve are obtained. The different forms could also be compared and ranked with each other using goodness-of-fit measures (see column 2 of the Table II.5). For Bel Air 5, the functional form of Sarabia (1997) ranks first with a chi-square statistic of 0.00036, then Ortega et al. (1991) functional form with a chi-square statistic of 0.00112 and Rasche et al. (1980) functional form with a chi-square statistic of 0.00154.

Table II.5. Inequality measures for the different functional forms of the Bel Air 5 iris

Functional form	χ^2	Rank	Gini	Pietra	T_L	T_H
Kakwani and Podder (1973)	0.01028	4	0.345	0.261	0.205	0.185
Rasche et al. (1980)	0.00154	3	0.351	0.248	0.221	0.214
Arnold (1986)	0.01493	6	0.342	0.259	0.189	0.184
Chotikapanich (1993)	0.01102	5	0.344	0.261	0.199	0.184
Sarabia (1997)	0.00036	1	0.356	0.244	0.24	0.287
Ortega et al. (1991)	0.00112	2	0.352	0.247	0.225	0.221
Rohde (2009)	0.02149	7	0.340	0.259	0.183	0.183

Best choice of functional forms

The ranking of functional forms according to the chi-square statistic can be obtained for each Parisian iris. The figure II.6 depicts the shares of each functional form per rank. The Ortega et al. (1991), Sarabia (1997) and Rasche et al. (1980) shapes predominate among the top rankings. 58% of the iris tend to be in favour of the Sarabia (1997) form, 38% for Ortega et al. (1991) form and 3% for Rasche et al. (1980) form. The figure II.5 spatially represents the best functional form of each iris in Paris.

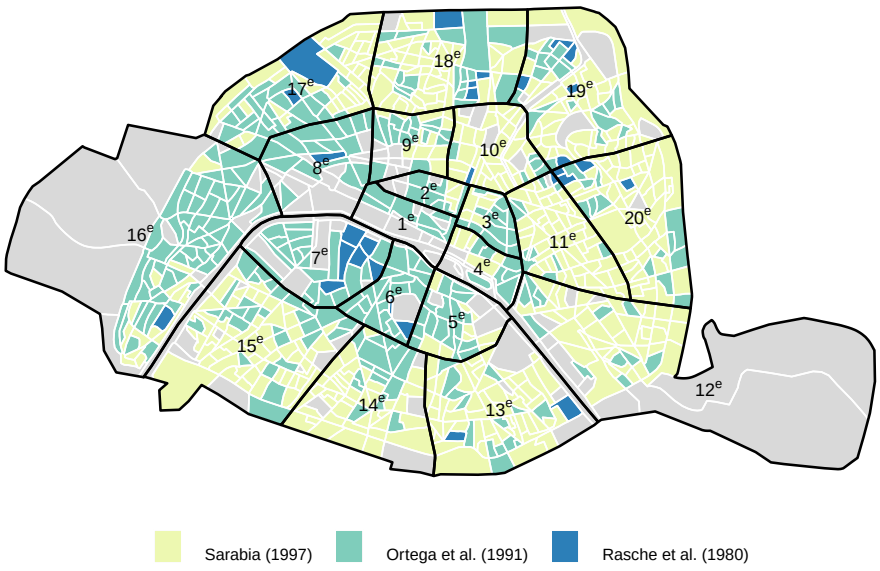


Figure II.5. Best choice of functional forms obtained for each Parisian iris

A rather notable pattern appears, the iris of the center and 16th arrondissements are generally in favour of an Ortega et al. (1991) form, while the iris of the outlying arrondissements are in favour of Sarabia (1997) form. This spatial pattern is similar to the one of inequality indices, and hence the functional form choice appears to be related to the income inequalities of the area. The means of the income quantiles can be examined according to the optimal functional

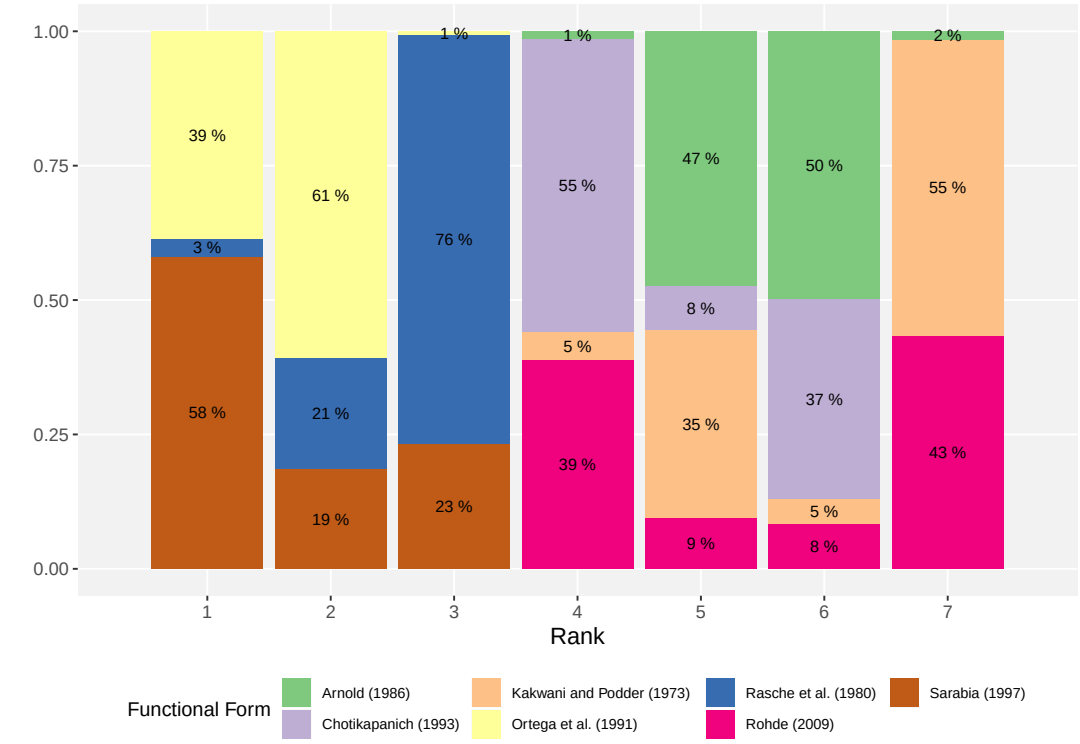


Figure II.6. Proportions of the different functional forms per rank

form (see figure II.7). The iris where the Ortega et al. (1991) form is preferred have, on average, higher income quantiles than the Sarabia (1997) form iris[5].

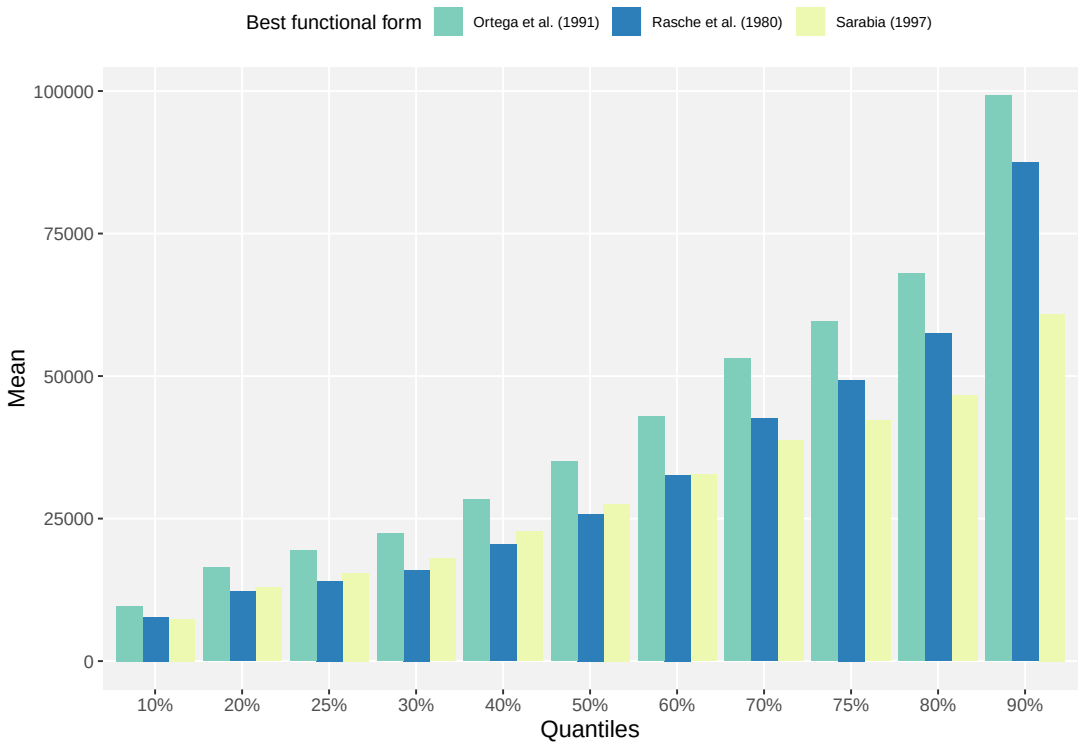


Figure II.7. Mean values of quantiles by best functional forms

[5]There are very few iris of the Rasche et al. (1980) form and the interpretation can be misleading.

Inequality measure

The Lorenz curve is used to compute inequality indices (see Table II.5 for Bel Air 5 iris). The various indexes can be defined as a function of each functional form parameters. The iris database provides the Gini index calculated from individual data. Then, a comparison between the one found with the functional form and the individual one should be considered. Figure II.8 depicts the Gini derived from functional forms as a function of the one given by INSEE and Figure II.9 the difference between them for each functional form. The method tends to slightly overestimate the Gini index compared to the individual one. The difference remains quite small, in most cases ranging from 0 to 0.02.

Similarly, an analysis with the *Midpoint method* is performed (See Appendix 5.2). The pattern of best choice of functional forms differs from the one obtained with the *Conditional Expectation method*. The Rasche et al. (1980) functional form tends to be preferred (56%), after Ortega et al. (1991) (38%) and, to a lesser extent, Arnold (1986) (3%), Kakwani and Podder (1973) (2%) and Sarabia (1997) (1%). Additionally, the *Midpoint method* seems to underestimate the true value of the Gini index, with a difference ranging from −0.06 to −0.03 (see Figures II.23 and II.24).

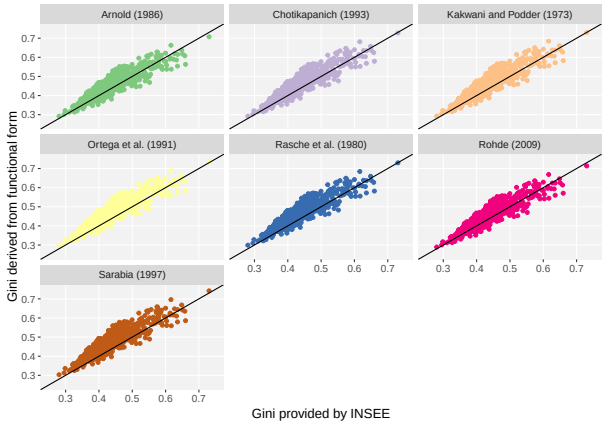


Figure II.8. Comparison between estimated and observed Gini index by functional form

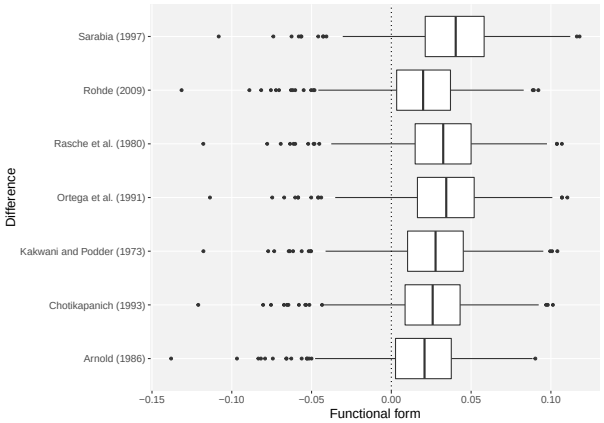


Figure II.9. Difference between estimated and observed Gini index

In addition, the method enables the calculation of inequality indices not provided by INSEE, such as the Pietra index or Theil indexes (see Figures II.10 for Gini index, II.12 for Pietra index, II.11 for Theil L index and II.13 for Theil H index). The pattern of income inequality in Paris is similar in terms of the inequality index used, with higher inequalities in the city center and the 16th arrondissement and lower inequalities in the outer arrondissements.

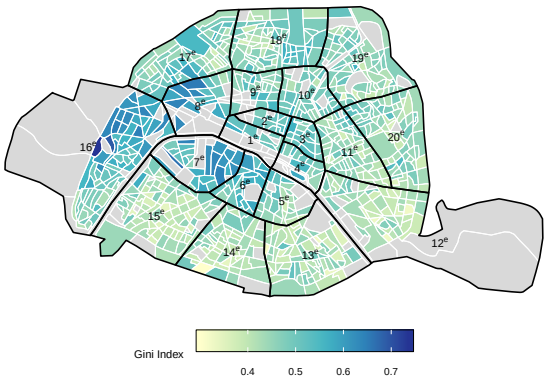


Figure II.10. Gini Index

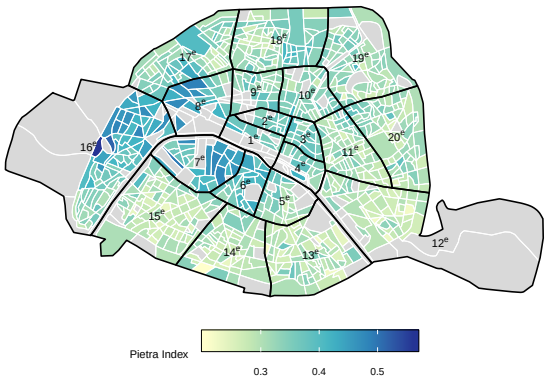


Figure II.12. Pietra Index

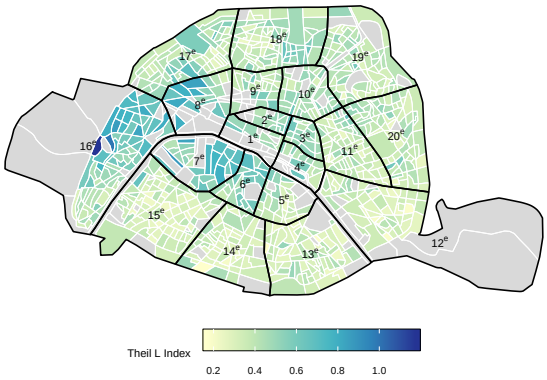


Figure II.11. Theil L Index

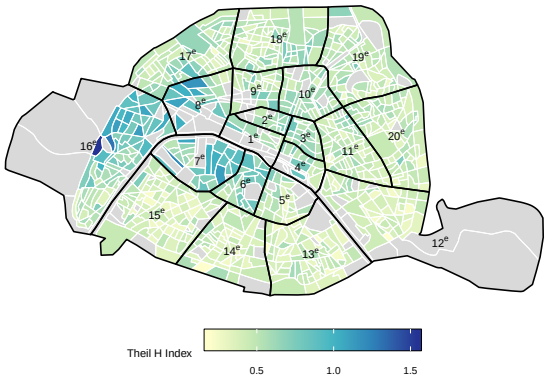


Figure II.13. Theil H Index

4 Conclusion

The objective of this paper is to propose an innovative method to model Lorenz curves and estimate inequality indices on small populations, when only quantiles are available. The method is based on conditional expectation in order to find the different income shares and thus model a Lorenz curve with the functional forms already proposed in the literature. Real and simulated data are used to evaluate the proposed method and compare it to other methods used. We note from simulated data that it is more difficult to estimate a Gini index when shares are not available and only quantiles are available. However, the proposed Conditional Expectation method outperforms the traditional Midpoint method. Similarly, the method applied to the Parisian iris data provides a Gini index very similar to the true value. Finally, the proposed methodology enables to model a Lorenz curve and hence to estimate inequality indices with quantile data.

Therefore, this method is useful for measuring inequalities when data are limited. This approach can be applied on income data in quantile form. However, it can also be used for

data in class form with an underlying distribution to find conditional expectations. Hence, this method overcomes the problems of imperfect data in the context of income inequality and enables the depiction of inequality measures when they appeared to be limited.

5 Appendix

5.1 Methodology examples on Parisian iris

This appendix provides examples of the methodology applied to other Parisian iris (Plaisance 1, Jardin des Plantes 1 and Invalides 1). Each example is composed of the same figures and tables as those displayed for the Bel Air 5 example, i.e. a table with quantiles, conditional means and income shares (Tables II.6, II.8, and II.10), a figure with the adjustment of the GB2 distribution (Figures II.14, II.16, and II.18), a figure with the empirical Lorenz curve (Figures II.15, II.17, and II.19) and finally a table with the inequality measures for the different functional forms and the goodness of fit (Tables II.7, II.9, and II.11).

Plaisance 1	
Arrondissement	14th
Population	2,169
Gini Index (INSEE)	0.2803

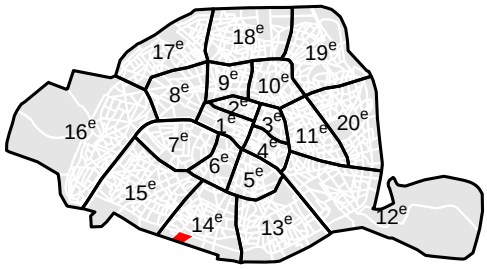


Table II.6. Tabulated data (€) obtained using the conditional expectation method for Plaisance 1 iris

	Quantile	Income Range	Proportion	Mean Income	Income share	Population share
D_1	7,630	Below 7,630	0.10	5,609.10	0.03	0.10
D_2	11,872	7,630 - 11,872	0.10	9,932.46	0.07	0.20
Q_1	13,486	11,872 - 13,486	0.05	12,688.12	0.10	0.25
D_3	14,782	13,486 - 14,782	0.05	14,135.49	0.14	0.30
D_4	17,132	14,782 - 17,132	0.10	15,947.17	0.21	0.40
Q_2	19,778	17,132 - 19,778	0.10	18,423.46	0.30	0.50
D_6	22,118	19,778 - 22,118	0.10	20,913.41	0.39	0.60
D_7	24,784	22,118 - 24,784	0.10	23,398.36	0.50	0.70
Q_3	26,130	24,784 - 26,130	0.05	25,442.64	0.56	0.75
D_8	27,920	26,130 - 27,920	0.05	26,998.92	0.62	0.80
D_9	33,202	27,920 - 33,202	0.10	30,331.71	0.77	0.90
		33,202 and over	0.10	50,102.78	1	1

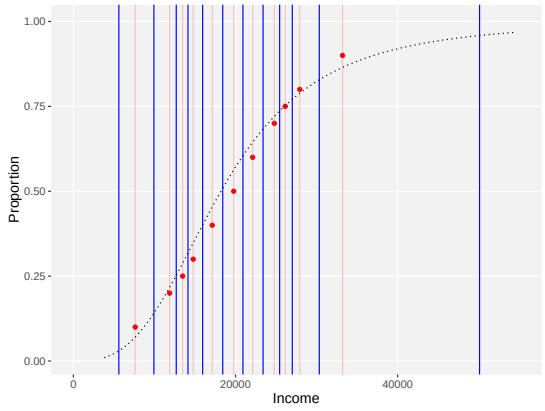


Figure II.14. Estimated means of income per bins of Plaisance 1 iris

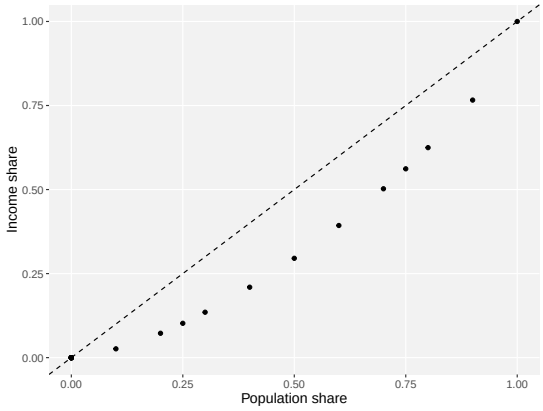


Figure II.15. Empirical Lorenz curve of Plaisance 1 iris

Table II.7. Inequality measures for the different functional forms of the Plaisance 1 iris

Functional form	χ^2	Rank	Gini	Pietra	T_L	T_H
Kakwani and Podder (1973)	0.00449	4	0.279	0.209	0.131	0.12
Rasche et al. (1980)	0.00025	3	0.284	0.2	0.149	0.133
Arnold (1986)	0.00623	5	0.278	0.209	0.127	0.119
Chotikapanich (1993)	0.00948	6	0.277	0.209	0.124	0.118
Sarabia (1997)	0.00006	1	0.286	0.199	0.155	0.156
Ortega et al. (1991)	0.00018	2	0.284	0.2	0.151	0.135
Rohde (2009)	0.0202	7	0.274	0.207	0.116	0.116

Jardins des Plantes 1	
Arrondissement	5th
Population	1,879
Gini Index (INSEE)	0.3497

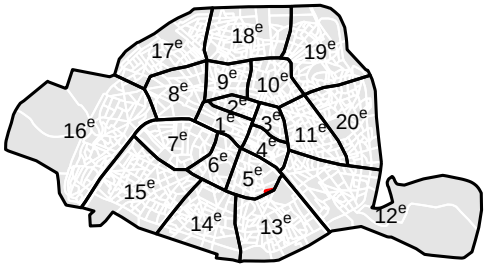


Table II.8. Tabulated data (€) obtained using the conditional expectation method for Jardin des Plantes 1 iris

	Quantile	Income Range	Proportion	Mean Income	Income share	Population share
D_1	12,762	Below 12,762	0.10	8,902.00	0.02	0.10
D_2	19,864	12,762 - 19,864	0.10	16,521.60	0.06	0.20
Q_1	23,802	19,864 - 23,802	0.05	21,851.10	0.08	0.25
D_3	26,770	23,802 - 26,770	0.05	25,284.70	0.11	0.30
D_4	32,078	26,770 - 32,078	0.10	29,387.13	0.17	0.40
Q_2	36,778	32,078 - 36,778	0.10	34,377.55	0.25	0.50
D_6	41,700	36,778 - 41,700	0.10	39,169.42	0.34	0.60
D_7	48,638	41,700 - 48,638	0.10	45,011.65	0.44	0.70
Q_3	53,198	48,638 - 53,198	0.05	50,846.34	0.50	0.75
D_8	59,716	53,198 - 59,716	0.05	56,309.24	0.56	0.80
D_9	76,948	59,716 - 76,948	0.10	67,345.96	0.71	0.90
		76,948 and over	0.10	130,729.00	1	1

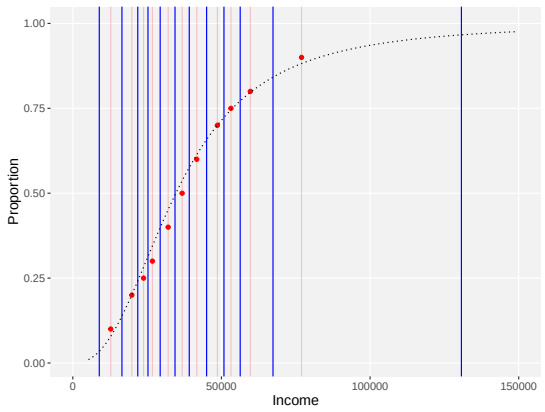


Figure II.16. Estimated means of income per bins of Jardin des Plantes 1 iris

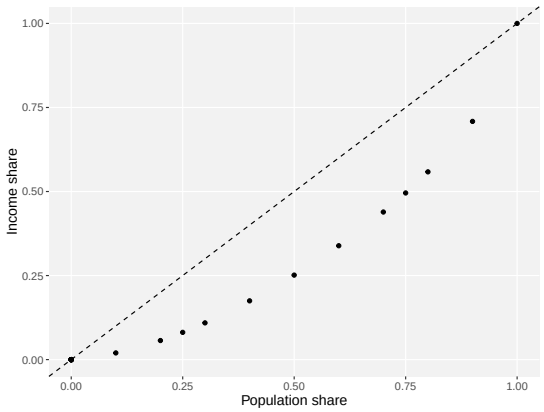


Figure II.17. Empirical Lorenz curve of Jardin des Plantes 1 iris

Table II.9. Inequality measures for the different functional forms of the Jardin des Plantes 1 iris

Functional form	χ^2	Rank	Gini	Pietra	T_L	T_H
Kakwani and Podder (1973)	0.00764	4	0.344	0.261	0.205	0.185
Rasche et al. (1980)	0.00039	2	0.35	0.248	0.224	0.211
Arnold (1986)	0.0109	6	0.342	0.259	0.191	0.183
Chotikapanich (1993)	0.00838	5	0.343	0.261	0.198	0.184
Sarabia (1997)	0.00065	3	0.353	0.247	0.225	0.248
Ortega et al. (1991)	0.00023	1	0.351	0.247	0.228	0.217
Rohde (2009)	0.0194	7	0.339	0.258	0.182	0.182

Invalides 1	
Arrondissement	7th
Population	2,057
Gini Index (INSEE)	0.6509

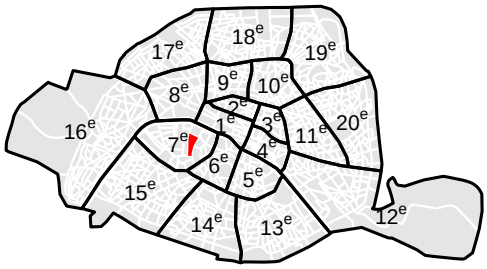


Table II.10. Tabulated data (€) obtained using the conditional expectation method for Invalides 1 iris

	Quantile	Income Range	Proportion	Mean Income	Income share	Population share
D_1	13,848	Below 13,848	0.10	8,356.21	0.01	0.10
D_2	21,006	13,848 - 21,006	0.10	17,459.94	0.02	0.20
Q_1	25,462	21,006 - 25,462	0.05	23,226.59	0.03	0.25
D_3	29,680	25,462 - 29,680	0.05	27,556.18	0.05	0.30
D_4	37,692	29,680 - 37,692	0.10	33,606.38	0.08	0.40
Q_2	46,840	37,692 - 46,840	0.10	42,136.14	0.12	0.50
D_6	57,324	46,840 - 57,324	0.10	51,895.25	0.17	0.60
D_7	80,502	57,324 - 80,502	0.10	67,993.51	0.23	0.70
Q_3	96,546	80,502 - 96,546	0.05	88,109.88	0.27	0.75
D_8	119,556	96,546 - 119,556	0.05	107,273.40	0.32	0.80
D_9	210,860	119,556 - 210,860	0.10	155,904.60	0.46	0.90
		210,860 and over	0.10	575,902.50	1	1

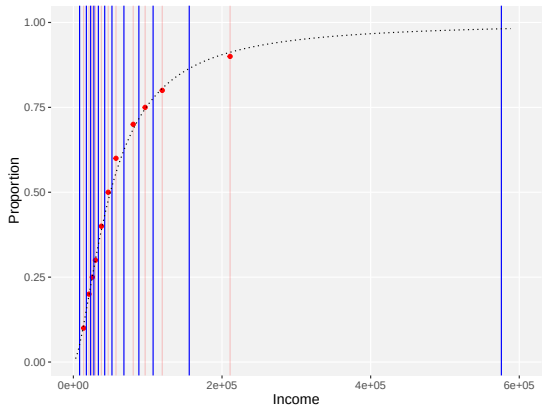


Figure II.18. Estimated means of income per bins of Invalides 1 iris

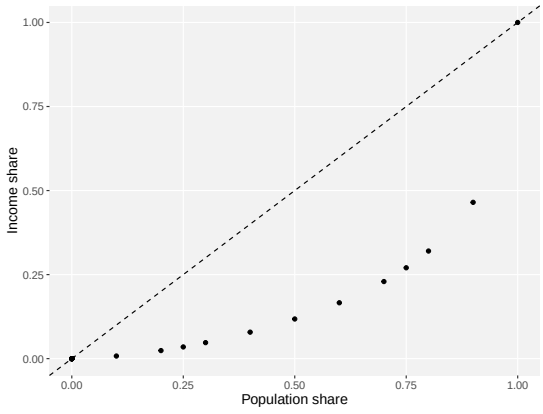


Figure II.19. Empirical Lorenz curve of Invalides 1 iris

Table II.11. Inequality measures for the different functional forms of the Invalides 1 iris

Functional form	χ^2	Rank	Gini	Pietra	T_L	T_H
Kakwani and Podder (1973)	0.108	7	0.556	0.432	0.692	0.518
Rasche et al. (1980)	0.00256	1	0.558	0.408	0.605	0.641
Arnold (1986)	0.0114	5	0.542	0.422	0.51	0.515
Chotikapanich (1993)	0.0670	6	0.553	0.431	0.646	0.512
Sarabia (1997)	0.00622	3	0.556	0.412	0.551	0.693
Ortega et al. (1991)	0.00588	2	0.561	0.405	0.633	0.697
Rohde (2009)	0.00983	4	0.543	0.422	0.514	0.514

5.2 Midpoint method on Parisian income data

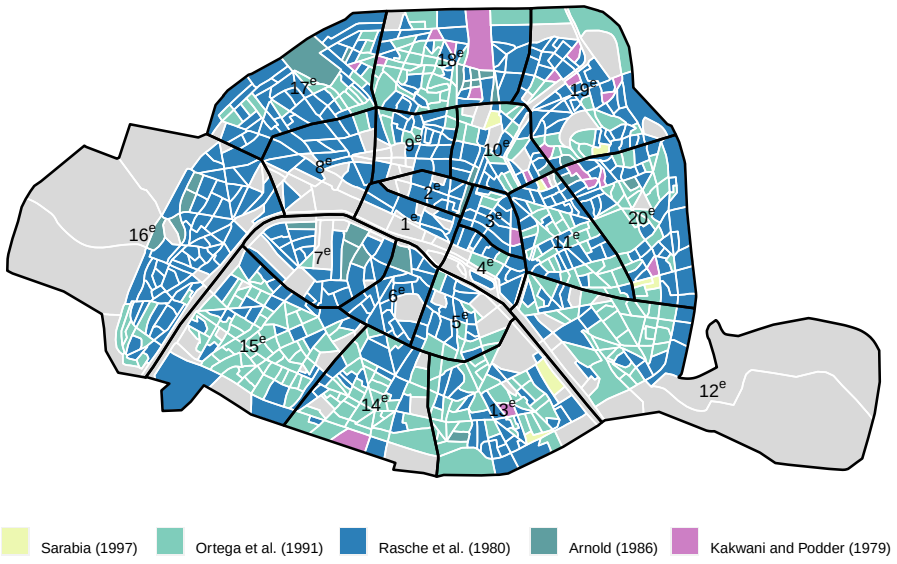


Figure II.20. Best choice of functional forms obtained for each Parisian iris (*Midpoint method*)

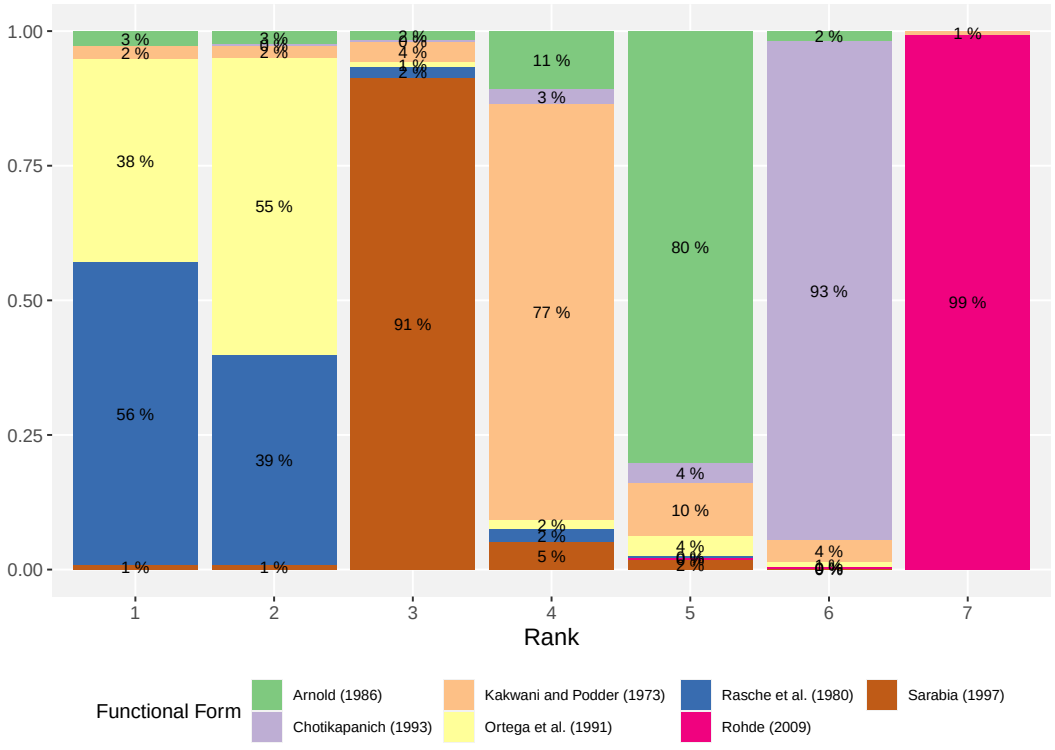


Figure II.21. Proportions of the different functional forms per rank (*Midpoint method*)

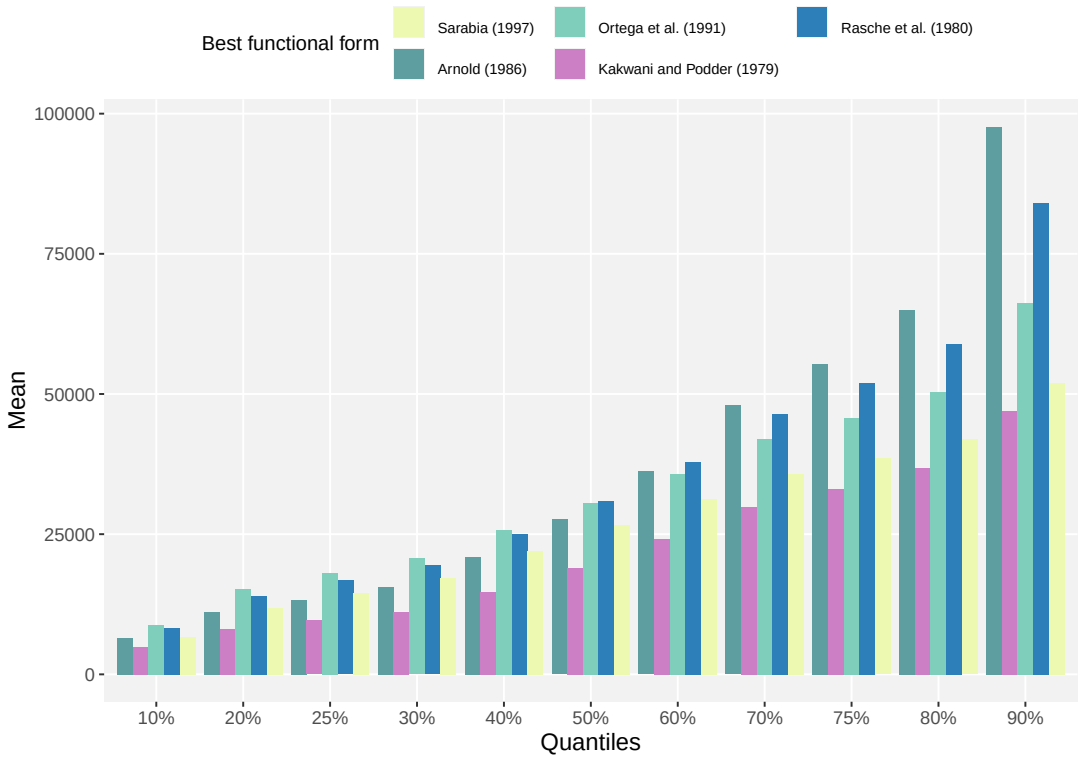


Figure II.22. Mean values of quantiles by best functional forms (*Midpoint method*)

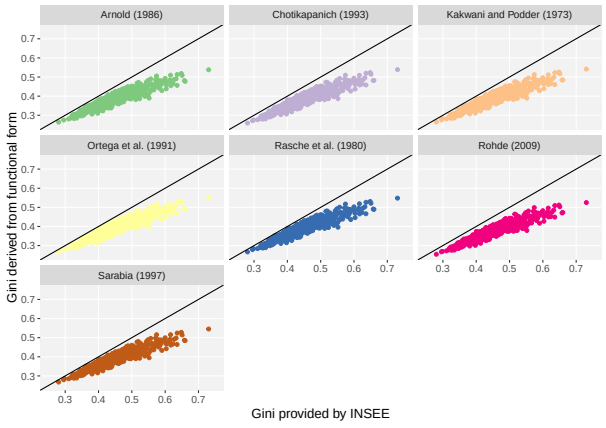


Figure II.23. Comparison between estimated and observed Gini index by functional form (*Midpoint method*)

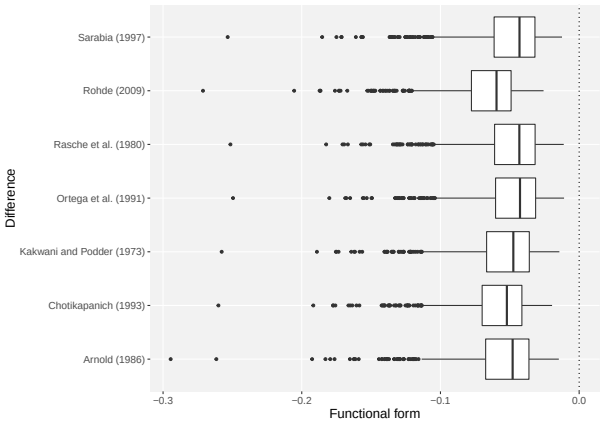


Figure II.24. Difference between estimated and observed Gini index (*Midpoint method*)

Bibliography

B. C. Arnold. A class of hyperbolic Lorenz curves. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 427–436, 1986.

B. C. Arnold. Pareto and generalized pareto distributions. In *Modeling income distributions and Lorenz curves*, pages 119–145. Springer, 2008.

R. L. Basmann, K. J. Hayes, D. J. Slottje, and J. Johnson. A general functional form for approximating the Lorenz curve. *Journal of Econometrics*, 43(1-2):77–90, 1990.

T. Blanchet, J. Fournier, and T. Piketty. Generalized pareto curves: theory and applications. 2017.

E. Castillo, A. S. Hadi, and J. M. Sarabia. A method for estimating Lorenz curves. *Communications in statistics-theory and methods*, 27(8):2037–2063, 1998.

D. G. Champernowne and F. A. Cowell. *Economic inequality and income distribution*. Cambridge University Press, 1998.

D. Chotikapanich. A comparison of alternative functional forms for the Lorenz curve. *Economics Letters*, 41(2):129–138, 1993.

D. Chotikapanich. *Modeling Income Distributions and Lorenz Curves*, volume 5. Springer Science & Business Media, 2008.

D. Chotikapanich and W. E. Griffiths. Estimating Lorenz curves using a Dirichlet distribution. *Journal of Business & Economic Statistics*, 20(2):290–295, 2002.

D. Chotikapanich and W. E. Griffiths. Averaging lorenz curves. *The Journal of Economic Inequality*, 3(1):1–19, 2005.

F. Cowell. *Measuring inequality*. Oxford University Press, 2011.

J. Fellman et al. Modelling lorenz curve. *Journal of Statistical and Econometric Methods*, 1(3): 53–62, 2012.

J. L. Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

C. Gini. Sulla Misura della Concentrazione e della Variabilità dei Caratteri. *Atti del Reale Istituto veneto di scienze, lettere ed arti*, 73:1203–1248, 1914.

W. Griffiths and G. Hajargasht. On gmm estimation of distributions from grouped data. *Economics Letters*, 126:122–126, 2015.

M. R. Gupta. Functional form for estimating the Lorenz curve. *Econometrica*, 52(5):1313–1314, 1984.

G. Hajargasht and W. E. Griffiths. Minimum distance estimation of parametric lorenz curves based on grouped data. *Econometric Reviews*, 39(4):344–361, 2020.

G. Hajargasht, W. E. Griffiths, J. Brice, D. P. Rao, and D. Chotikapanich. Inference for income distributions using grouped data. *Journal of Business & Economic Statistics*, 30(4):563–575, 2012.

N. C. Kakwani and N. Podder. On the estimation of Lorenz curves from grouped observations. *International Economic Review*, pages 278–292, 1973.

L. Kaplow. Why measure inequality? *The Journal of Economic Inequality*, 3(1):65–79, 2005.

C. Kleiber and S. Kotz. *Statistical size distributions in Economics and Actuarial Sciences.*, volume 470. John Wiley & Sons, 2003.

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219, 1905.

J. B. McDonald. Some Generalized Functions for the Size Distribution of Income. *Econometrica: journal of the Econometric Society*, pages 647–663, 1984.

P. Ortega, G. Martin, A. Fernandez, M. Ladoux, and A. Garcia. A new functional form for estimating Lorenz curves. *Review of Income and Wealth*, 37(4):447–452, 1991.

A. G. Pakes. *On Income Distributions and their Lorenz Curves.* Department of Mathematics, University of Western Australia, 1981.

G. Pietra. *Nuovi contributi alla metodologia degli indici di variabilita e di concentrazione.* Ferrari, 1932.

R. Rasche, J. Gaffney, A. Koo, and N. Obst. Functional forms for estimating the Lorenz curve: comment. *Econometrica*, 48(4):1061–1062, 1980.

N. Rohde. Lorenz curves and generalised entropy inequality measures. In *Modeling Income Distributions and Lorenz Curves*, pages 271–283. Springer, 2008.

N. Rohde. An alternative functional form for estimating the Lorenz curve. *Economics Letters*, 105(1):61–63, 2009.

J.-M. Sarabia. A hierarchy of Lorenz curves based on the generalized Tukey's lambda distribution. *Econometric Reviews*, 16(3):305–320, 1997.

J.-M. Sarabia, E. Castillo, and D. J. Slottje. An ordered family of Lorenz curves. *Journal of Econometrics*, 91(1):43–60, 1999.

H. Theil. Economics and information theory. Technical report, 1967.

J. Villaseñor and B. C. Arnold. Elliptical Lorenz curves. *Journal of Econometrics*, 40(2): 327–338, 1989.

CHAPTER III

Bodily Injury Claims in France : Negotiation or Court ?[1]

[1] Chapter co-written with Arthur Charpentier and Pierre-Yves Geoffard

1 Introduction

A huge literature in economics and law outlines compensation systems for personal injury accidents. Economic models of the litigation process aim to describe and explain the decision to sue and the outcomes of trials. Many authors initially focused on tort regimes (Bebchuk, 1984, Nalebuff, 1987, Priest and Klein, 1984). Such models are bargaining models with asymmetric information (see Farmer and Pecorino (1996) for a survey of asymmetric information in litigation process). Players learn during the game, either about the amount of the victim's damages or the degree of fault of the defendant. Osborne (1999a) and Osborne (1999b) analyse the impact of information asymmetry on the decision to litigate.

As discussed in Sugarman (2010), many countries have based their compensations on the "no-fault" law. Sugarman (2000) reviews the historic changes in American personal injury law. One third of American states have adopted a no-fault or partially no-fault regime, but also other countries such as most European countries (Germany, France, Italy, Austria, Belgium, Spain or the Netherlands), some states in Canada (Quebec, 1978) and New Zealand in 1974. These systems are based on the compensation of the victim without fault and an insurer then stands in for the defendant. Devlin (2002) investigates the factors that influence the switch between a tort and no-fault regime and the types of no-fault regime.

In France, the law of 5 July 1985 is intended to improve the compensation process of traffic accident victims. Victims are entitled to receive compensation for all their damages. Compensation must then reflect both financial losses, such as medical expenses or loss of income, and non-financial losses, such as pain and suffering. The French system is a three-step system. The insurance company proposes an out-of-court compensation based on medical expertise. Following this first proposal, the victim is faced with a simple choice: either accept this compensation or go to court. If the victim decides to go to court, a second amount will be offered. A final step can occur in case of appeal.

The compensation procedure is a decision-making problem. The choice between the amount negotiated by the insurance company and the decision to go to trial is based on the actual amount proposed after negotiation and a prediction of the amount that might be obtained at trial. Risk aversion and time preferences of the victim also influence the decision to pursue or negotiate. From an econometric approach, the problem is that the only information available is the amount finally obtained, the total duration of the procedure and the procedure selected.

The availability of information thus depends on the decision. Based on limited-dependent model of Maddala (1986), a five-equation model is developed: a decision equation to model the choice between negotiated settlement versus trial, two amount equations to model the compensation received in out-of-court and court proceedings, and two time equations to model the duration of out-of-court and court proceedings.

The paper proceeds as follows: Section 2 describes the compensation process and in particular the French system, Section 3 presents the theoretical decision model, Section 4 explains our limited-dependent econometric model, Section 5 details our dataset, Section 6 outlines the overall regression considered and the results, and Section 7 concludes.

2 Process

The various liability systems for bodily injury claims are described in section 2.1. The french one is based on no-fault law, voted in 1985, the so-called "loi Badinter" (see section 2.2). The amount of compensation must reflect the victim's pain and suffering and also personal aspect, defined by guidelines and scales, listed in section 2.3. Finally, section 2.4 outlines the timing of the french compensation system.

2.1 Liability system

Basically, two types of automobile insurance systems exist : "no-fault" and "tort" law. In "tort" law, the degree of responsability defines the amount of the compensation. Victims may bring an action against any person who caused the harm negligently or intentionally. Bebchuk (1984) assumed that there is, on average, a positive gain when going to court due to asymmetric information. Asymmetric information includes the level of damage (on the plaintiff's side) or the level of negligence (on the defendant's side). Bebchuk (1984) and Nalebuff (1987) propose models where asymmetric information refers to the defendant's level of fault.

In "no-fault" law, the insured are compensated by their own insurance companies, regardless of their liability. In such laws, pain and suffering are also considered. Victims collect their financial and non-financial losses directly from their own insurer regardless the fault (Cummins and Weiss, 1991). New-Zealand introduced this legislation in 1974 and many countries in Europe and States in United States of America and Canada have adopted similar legal systems. Devlin

(2002) investigates the factors that influence the switch between a tort and no-fault regime and the types of no-fault regime. According to Devlin, the insurance industry and the state's regulatory environment are both involved in the existence of a no-fault regime.

No-fault liability systems encourage risky behaviour by weakening the tort deterrent. These legislations lead to an increase in motoring and in the number and probability of accidents (Brown, 1985). In Cummins et al. (2001), the link between fatal accidents rate and no-fault systems is investigated both empirically and theoritically. It seems ambiguous from a theoritical perspective but empirically positive. Browne and Schmit (2008) finds that the rate of increase in bodily injury claims slowed after tort reforms based on U.S. automobile accidents in 1977, 1987 and 1997. Liao and White (2002) also compares efficiency under tort and no-fault systems. According to them, neither system dominates the others on efficiency and this efficiency relies on the cost of care.

2.2 Loi Badinter

In France, the law of 5 July 1985 —also known as the *Loi Badinter*— sets out the rights and liabilities for the compensation of bodily injuries. This law is applicable if the accident involves at least one motor vehicle in a traffic accident, the plaintiff has suffered an injury and a causal link is established between the injury and the accident. It is a "no-fault" liability regime. Indeed, the driver of the vehicle is always liable and "non-driver" victims (pedestrians, cyclists, passengers) are the most privileged. Compensation is due, whatever the responsibility of the driver. In France, car insurance is compulsory and the insurance company is then financially responsible. Note that if the guardian is not insured, a special Guarantee Fund exists.

Insurers are required to send a written questionnaire and make a compensation offer within eight months from the date of the collision (and a definitive offer within five months after the insurer is aware of the consolidation of the victim's injuries). If the victim accepts the offer, the case will be settled, see Chauchard (1989). The limitation period is ten years from the date of the injury. Table (III.1) reports average and median lengths by procedure. An out-of-court procedure takes about 356 days on average, compared with 832 days in the first instance and 1417 days on appeal. On Figure (III.1), we can vizualise the difference of procedure lengths by court location.

Table III.1. Number of accidents, lenght and indemnity per type of procedure and court type *(source: AGIRA)*.

	Number of accidents	Lenght total (mean)	Lenght total (median)	Indemnity (mean, €)	Indemnity (median, €)
Compromise	243465	356 days	241 days	20448	7062
Judicial (first)	13004	832 days	688 days	72022	21484
Judicial (appeal)	1626	1417 days	1363 days	233688	66179
Judicial (civil)	10987	914 days	755 days	82900	22478
Judicial (criminal)	3609	847 days	664 days	110067	28570

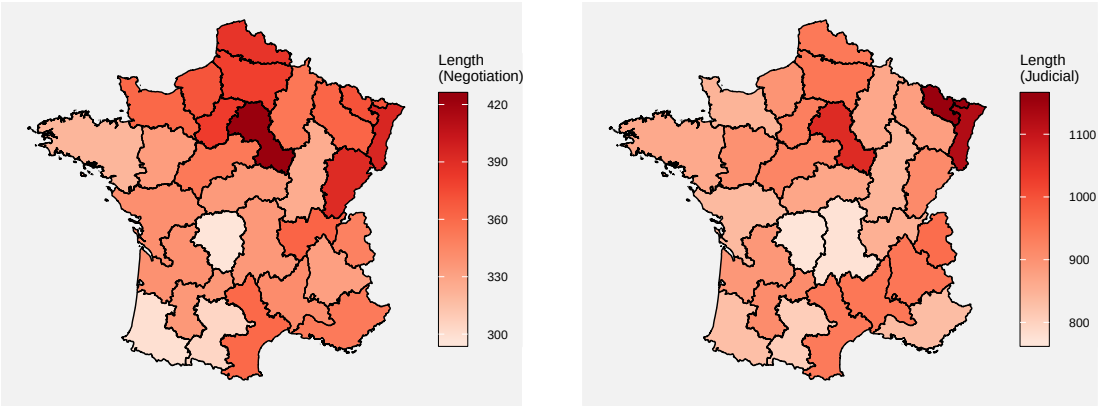


Figure III.1. Average procedure length settled by negotiation with the insurance company on the left, and on the right settled by courts.

2.3 Compensations

Compensation must reflect the victim's pain and suffering caused by the accident, but also consider personal aspects. A monetary amount must be defined to cover all damages. The so-called Dintilhac nomenclature lists the components considered in the amount of compensation and specifies guidelines and scales (Dintilhac, 2005). This report provides details on the compensatory elements for reference. It includes both patrimonial and extrapatrimonial damages. Figure (III.2) depicts average indemnities settled in negotiation and by the courts in the different jurisdictions. The average amount under negotiation is 20448€ and 72022€ at trial (Table III.1).

Patrimonial damages are the plaintiff's financial losses, which cover both pre- and post-consolidation costs. This refers to current and future health expenditures, current and future losses of professional earnings, professional incidence or third party assistance. Health expenses cover medical, pharmaceutical and hospitalisation expenses related to the accident. The losses of professional earnings is due to the victim's incapacity to work or apply for a job following the accident (*pertes de gains professionnels actuels*) and income loss or decrease due to permanent

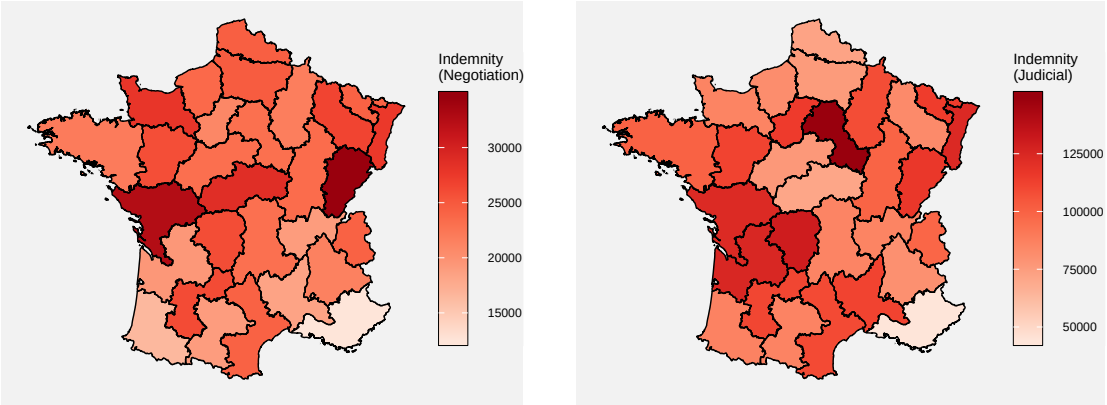


Figure III.2. Average indemnity settled by negotiation with the insurance company on the left, and on the right settled by courts.

disability (*pertes de gains professionnels futures*). Finally, the third party assistance (*assistance par une tierce personne*) covers the costs of a permanent assistance due to his or her disability.

Extrapatrimonial damages are non-financial losses. Extrapatrimonial damages also include pre- and post-consolidation injuries. These related to the reduction of physical and psychological potential and the pain endured by the victim. Several aspects shall be considered, such as: suffering endured, temporary or permanent functional deficit, damage to physiological and psychological integrity, aesthetic damage or agreement damage. Several guidelines are available for each of its components. The suffering endured (previously called *pretium doloris*) constitutes a contribution to the physical and moral suffering endured by the victim, measured on a scale of 0-7. The permanent functional deficit (*déficit fonctionnel permanent*) or partial incapacity (*incapacité permanente partielle*) is a percentage reflecting the degree of disability and the reduction of physical and psychological potential. The age of the victim is also a main criterion used to determine compensation for functional deficit. The aesthetic damage repairs any damage altering appearance or expression such as scars, mutilations or lameness, measured on a scale of 0-7. Finally, the agreement damage compensates the victim's inability to practice sports or leisure activities.

In case of death, they also include funeral expenses, loss of income expenses of family members and moral pain.

2.4 French compensation system

The system is a three-step process. Consider a car accident. The victim can obtained a financial compensation for all damages. After the victim's state has been consolidated, a medical expertise is carried out to value the damage suffered by the victim. On the basis of this expertise, the insurance company must offer compensation to the victim for all expenses and damages. After this bargaining step, the victim has two alternatives: accept the insurer's proposal out of court or go to court. Next, if she decides to pursue litigation, the second step is the court. The trial can be held in the criminal courts for a serious offence or in the civil courts. The court decides on an amount that the insurer will have to pay to the victim. A third step is possible if the victim or insurance company decided to appeal. This appeal decision is well described in several articles, see Santolino (2010).

In cases where compensation could not be settled out-of-court, the only possible recourse is the court. However, the competent court, whether civil or criminal, must be referred to. The civil court enables compensation without the possibility of criminal sanctions. In criminal matters, the victim seeks an expert opinion and compensation for his or her losses. Moreover, when the offence is serious or the victim is in a serious physical or psychological condition, it is legitimate to seek the offender's criminal conviction (suspension or cancellation of driving, suspended prison sentence or hard sentence, criminal fine, etc.).

On Figure (III.3) , we can visualize geographic specificies. For instance, in Aix-en-Provence, 87% of cases are settled by civil juridisctions, against 81% for the rest of France.

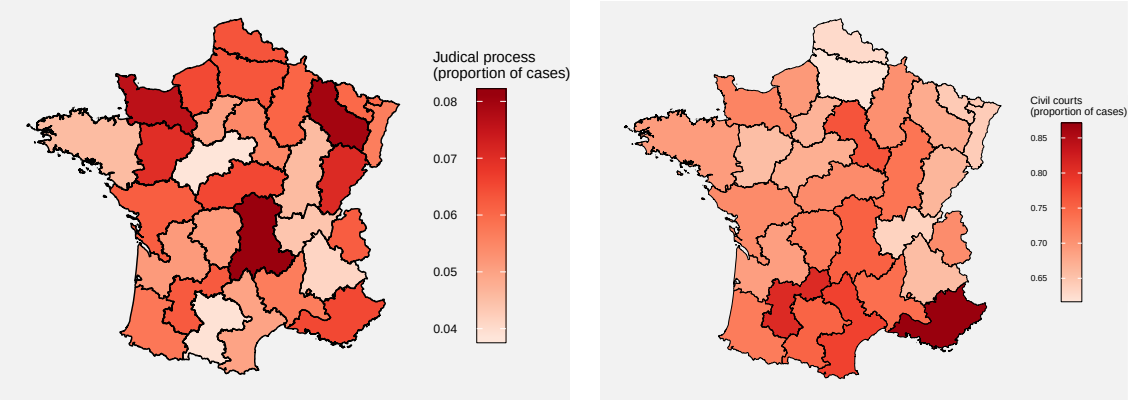


Figure III.3. Proportion of cases settled by courts on the left, and proportion of judicial cases settled by civil courts on the right.

3 Model

The bodily injury compensation procedure can be seen as a decision-making problem. Viscusi (1988) suggests that risk aversion has a statistically significant effect on the decision to go to trial, however, it is not the only determinant of claimant's behaviour. At time t, the insurance company is required to propose a compensation y_S to the victim. The victim is faced with a choice, accept this offer or decide to go to court. If she refers to the court, the court will decide the compensation y_T at time $t+h$. At this stage as well, the victim can accept the compensation or appeal and go to a higher degree court. Notice that the insurance company may also decide to appeal the judgment. The court of appeal will decide a compensation y_A at time T.

In this version of the model, we exclude the possibility of appeal. The victim faces thus a simple decision : accept y_S now ($d = 0$), or go to court ($d = 1$) and receive an amount y_T. This compensation y_T is unknown at time t, its distribution is the conditional distribution on the information available at time t, in particular about the extent of the prejudice. An individual victim may have quite information about this distribution, but lawyers may have some and, based on their knowledge, may advise the victim.

We assume preferences are represented by a CARA exponential function, $v(y) = -\exp(-\rho y)$. In that case, $\mathbb{E}v(y_T) = v[\mathbb{E}(y_T) - \frac{\rho}{2}\mathbb{V}(y_T)]$, where $\mathbb{E}(y_T)$ and $\mathbb{V}(y_T)$ denote, respectively, the expectancy and variance of y_T. That is, the victim is interested in maximizing $\mathbb{E}(y_T) - \frac{\rho}{2}\mathbb{V}(y_T)$. This utility is to be compared with the amount received at time t, compounded between t and $t+h$ at the risk free rate r, and therefore equal at time $t+h$ to $\exp(rh)y_S$. With intertemporally separables preferences and a rate δ of time preference, the choice is between y_S and the discount expected utility of y_T, which leads the victim to decide to go to court (d =1) if :

$$\exp(-\delta h)\left[\mathbb{E}(y_T) - \frac{\rho}{2}\mathbb{V}(y_T)\right] > y_S \tag{III.1}$$

Note here: the larger risk preference parameter ρ is, the more risk averse the agent is and the larger time preference parameter δ is, the more short-sighted the agent is. Based on the model, we find that, ceteris paribus, a higher rate of time preference δ, or a higher risk aversion parameter ρ, decreases the relative benefit of going to court. A lower expected payoff, and/or a higher variance of the future payoff, go in the same direction.

In the dataset, no information about preferences parameters is given, but the empirical

literature has identified that time preference δ and risk aversion ρ is related to individual characteristics. Arrondel and Masson (2013) notes that men and younger people are more risk tolerant than women and older ones. Women and older people are more likely far-sighted than men and younger people. The temporal preference difference between men and women is debatable and appears not significant in some cases (Arrondel and Masson, 2005). Consequently, the model predicts that, other things being equal, men should go more frenquently to court.

However, the prediction of the model with regard to age is not certain. The decision to go to trial is a trade-off between time preference and risk preference. Younger people are more risk tolerant but less far-sighted, therefore they have a lower risk aversion parameter ρ and a higher temporal preference parameter δ. In contrast, older people have a higher risk aversion parameter ρ and a lower temporal preference parameter δ. In such cases, the effect directions of the preference parameters in the model are opposite.

Arrondel and Masson (2005) defined four individuals types : "Hotheads" (low ρ, high δ), "Short-sighted prudent" (high ρ, high δ), "Entreprising" (low ρ, low δ) and "Armchairs investitors" (high ρ, low δ). "Short-sighed prudent" individuals are less likely to go to trial compared to "Entreprising" individuals. For "Hotheads" and "Armchairs investitors", a trade-off between time and risk preferences again occurs.

The decision to go to court also differs according to the type and severity of the accident. The severity and complexity of the injury suffered by the claimant affects the decision to sue and its outcome (Browne and Schmit, 2008). The most serious and complex damage can lead to higher compensation. The quantification of the settlement amount is more complex and could be more controversial. Hence, victims of more serious and complex accidents are more likely to sue in court, other things being equal.

Finally, the decision to bargain or litigate depends on the geographical area. The outcome of the court is strongly related to the conditions of the area. The decision of an individual whether or not to sue depends on the observed distribution of amounts within his/her jurisdiction. The model predicts that, ceteris paribus, in areas where there is a lower expectancy or a higher variance in judicial amounts, victims should go less frequently to court. A low expectancy reflects a low expected outcome at trial and a high variance reflects a more uncertain outcome at trial.

However, we only observe the actual compensation and thereby the truncated distributions

of y_S and y_T. In terms of data, we observe y_S if the victim accepts the negotiation compensation and y_T if the victim goes to court.

4 Econometric model

In this section, we will derive an econometric model to better understand the decision to go to trial. This model is necessary since the decision is based on both amounts and length of procedures. Nevertheless, it is a limited-dependent model because we only know the actual amount and the actual length of the procedure.

This section is divided as follows: Section 4.1 presents the amount model, Section 4.2 presents the time model, Section 4.3 presents the decision model and finally Section 4.4 presents the final model and Section 4.5 presents the inference.

4.1 Amount model

Let assume the quantities of interest y_S and y_T the compensations obtained respectively by negotiation and litigation for a given claim. These amounts are function of the characteristics of individual x_i and the geographical characteristics of the court. The amounts y_S and y_T are assumed to be not time-dependents. Furthermore, y_S and y_T are independents. Indeed, the amount decided at trial must be legally independent of the amount proposed by the insurer. The judge, when deciding on compensation from the court, is not informed of the amount previously chosen. Assumes that

$$\log(y_S) = \lambda_{S_0} + \boldsymbol{\lambda}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_S} + \varepsilon_S \tag{III.2}$$

for the negotiated settlement and

$$\log(y_T) = \lambda_{T_0} + \boldsymbol{\lambda}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_T} + \varepsilon_T \tag{III.3}$$

for the trial, where λ_S and λ_T are sets of parameters, ω_S and ω_T are the court specific heterogeneity and ε_S and ε_T are the error terms and supposed to be homoscedastic noises. In addition, either y_S or y_T is observed. We only observe y_S for out-of-court settlements ($d = 0$) and y_T for

trials ($d = 1$). More precisely,

$$Y = \begin{cases} y_S & \text{if } d = 0 \\ y_T & \text{if } d = 1 \end{cases}$$

Thus y_S and y_T is limited-dependant variables, as defined in Maddala (1986). If the (true) distributions of y_S and y_T is supposed the be Gaussian, distribution of observed quantities are truncated (from the left). Similarly, error terms are also truncated, and least square estimators are no longer convergent (see Heckman (1976) and Lee (1978)). More specifically, from the selection procedure, we have a selection bias in our observations. Observe for instance that

$$\mathbb{E}(\log(y_S)|d = 0) = \lambda_{S_0} + \boldsymbol{\lambda}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_S} + \mathbb{E}(\varepsilon_S|d = 0)$$

and

$$\mathbb{E}(\log(y_T)|d = 1) = \lambda_{T_0} + \boldsymbol{\lambda}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_T} + \mathbb{E}(\varepsilon_T|d = 1)$$

Here, both $\mathbb{E}(\varepsilon_S|d = 0)$ and $\mathbb{E}(\varepsilon_T|d = 1)$ are non-null terms, and the natural idea is to derive expressions for those two terms, to improve inference.

4.2 Time model

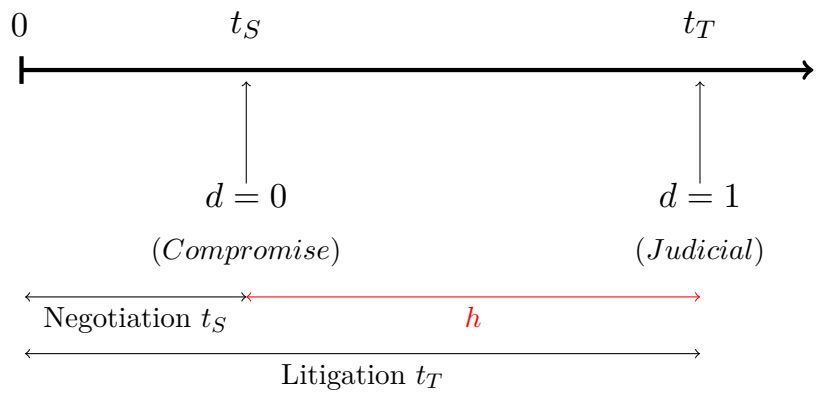


Figure III.4. Out-of-court and judicial settlement procedures

The purpose is to model the length between the insurance company's offer and the decision at trial. Note t_S the settlement date, t_T the trial date and the duration between them is h (see Figure (III.4)). For out-of-court settlements ($d = 0$), the duration t_S is known and for court settlements ($d = 1$), the duration t_T is known. For each observation, we have a single date t_{obs}, which is the duration of the procedure and therefore depends on the choice of the individual. As such, the data are censored for both procedures.

For the duration of the out-of-court process t_S, we observe t_{obs} for negotiated settlements and we know that t_S is less than t_{obs} for litigation cases. For the duration of the court process t_T, we observe t_{obs} for litigation cases and we know that t_T is greater than t_{obs} for negotiated settlements. Similarly to the amounts, either t_S or t_T is observed. Procedure times are therefore limited-dependant variables. More precisely,

$$T = \begin{cases} t_S & \text{if } d = 0 \\ t_T & \text{if } d = 1 \end{cases}$$

The aim is to model the duration between the two procedures. Assume that the procedure lengths are likely to depend on the regional characteristics and the victim characteristics x_i. Assumes that,

$$\log(t_S) = \beta_{S_0} + \boldsymbol{\beta}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_S} + \nu_S \tag{III.4}$$

and

$$h = \beta_{T_0} + \boldsymbol{\beta}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_T} + \nu_T \tag{III.5}$$

with $h = \log(t_T) - \log(t_S)$ and where β_S and β_T are sets of parameters, δ_S and δ_T are the regional specific heterogeneity and ν_S and ν_T are the error terms and supposed to be homoscedastic noises. Similarly to amounts, t_S and h are limited-dependent variables and error terms are also truncated. Assumes that,

$$\mathbb{E}(\log(t_S)|d = 0) = \beta_{S_0} + \boldsymbol{\beta}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_S} + \mathbb{E}(\nu_S|d = 0)$$

and

$$\mathbb{E}(h|d = 1) = \beta_{T_0} + \boldsymbol{\beta}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_T} + \mathbb{E}(\nu_T|d = 1)$$

Here, both $\mathbb{E}(\nu_S|d = 0)$ and $\mathbb{E}(\nu_T|d = 1)$ are non-null terms.

4.3 Decision model

As explained in the theoretical model, assuming the two amounts and the two procedural times are known, one should go to trial if

$$\exp(-\delta h)\left[\mathbb{E}(y_T) - \frac{\rho}{2}\mathbb{V}(y_T)\right] - y_S > 0$$

The decision equation depends both on a time factor and on the difference of the amounts, but also on the individual and the accident characteristics, and thus is defined as

$$d = \mathbb{1}(T^* > 0) \text{ with } T^* = \theta_0 + \boldsymbol{\theta_1 h} + \boldsymbol{\theta_2 \log(t_S)} + \boldsymbol{\theta_2(\log(y_T) - \log(y_S))} + \boldsymbol{\theta_3^\top x} + \boldsymbol{\alpha_c} + u \quad \text{(III.6)}$$

with $h = \log(t_T) - \log(t_S)$ and where θ is set of parameters, α_c is the court specific heterogeneity and u is supposed to be an homoscedastic noise, with variance σ_u^2.

Both the couple y_S and y_T and the couple t_s and h are limited-dependent variables. Thus, the error terms in the equations (III.2), (III.3), (III.4) and (III.5) are not null. The technique used here, described in Lee (1978) and Maddala (1986) is called the 'structural probit method' in Lee (1979). It is a two stage method, as the tobit model discussed in Tobin (1958), the main different is that the second step is not sufficient here to estimate all parameters. Some estimators for variance estimators should still be derived.

4.4 Final model

By inserting the amount and time equations, we can rewrite the Decision Equation (III.6) as

$$T^* = \frac{\gamma_0}{\sigma_{u^*}} + \frac{\gamma_1}{\sigma_{u^*}}\boldsymbol{x} + \frac{\gamma_c}{\sigma_{u^*}} + u^* \quad \text{(III.7)}$$

where $\gamma_0 = \theta_0 + \theta_1\beta_{T_0} + \theta_2\beta_{S_0} + \theta_3(\lambda_{T_0} - \lambda_{S_0})$, $\gamma_1 = \theta_3 + \theta_1\beta_{T1} + \theta_2\beta_{S1} + \theta_3(\lambda_{T1} - \lambda_{S1})$, $\gamma_c = \theta_1\delta_T + \theta_2\delta_S + \theta_3(\omega_T - \omega_S)$, $u^* = u + \theta_1\nu_T + \theta_2\nu_S + \theta_3(\varepsilon_T - \varepsilon_S)$ and $\sigma_{u^*} = \mathbb{V}(u + \theta_1\nu_T + \theta_2\nu_S + \theta_2(\varepsilon_T - \varepsilon_S))$. Assume that $\sigma_{u^*} = 1$ to have an identifiable model, then the decision equation is defined as

$$\Psi = \gamma_0 + \gamma_1\boldsymbol{x} + \gamma_c \quad \text{(III.8)}$$

So that, $T^* = \Psi + u^*$ and $\mathbb{P}(Y_T \text{ is observed}) = \mathbb{P}(\Psi > u^*) = \Phi(\Psi)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Observe further that the conditional distribution of $(\varepsilon_S, \varepsilon_T, \nu_S, \nu_T)$ given u^* is a joint Gaussian distribution (detailed in the Appendix 8.1). It is possible to write $\mathbb{E}[\varepsilon_S|d = 0]$, $\mathbb{E}[\varepsilon_T|d = 1]$, $\mathbb{E}[\nu_S|d = 0]$ and $\mathbb{E}[\nu_T|d = 1]$ as follows

$$\mathbb{E}[\varepsilon_S|d = 0] = \mathbb{E}[\varepsilon_S|\Psi > u^*] = \mathbb{E}[\sigma_{\varepsilon_S u^*}u^*|\Psi > u^*] = \sigma_{\varepsilon_S u^*}\left(\frac{-\phi(\Psi)}{\Phi(\Psi)}\right) = \sigma_{\varepsilon_S u^*}\pi_S,$$

$$\mathbb{E}[\varepsilon_T|d=1] = \mathbb{E}[\varepsilon_T|\Psi \leq u^*] = \mathbb{E}[\sigma_{\varepsilon_T u^*} u^*|\Psi \leq u^*] = \sigma_{\varepsilon_T u^*}\left(\frac{\phi(\Psi)}{1-\Phi(\Psi)}\right) = \sigma_{\varepsilon_T u^*}\pi_T,$$

$$\mathbb{E}[\nu_S|d=0] = \mathbb{E}[\nu_S|\Psi > u^*] = \mathbb{E}[\sigma_{\nu_S u^*} u^*|\Psi > u^*] = \sigma_{\nu_S u^*}\left(\frac{-\phi(\Psi)}{\Phi(\Psi)}\right) = \sigma_{\nu_S u^*}\pi_S$$

and

$$\mathbb{E}[\nu_T|d=1] = \mathbb{E}[\nu_T|\Psi \leq u^*] = \mathbb{E}[\sigma_{\nu_T u^*} u^*|\Psi \leq u^*] = \sigma_{\nu_T u^*}\left(\frac{\phi(\Psi)}{1-\Phi(\Psi)}\right) = \sigma_{\nu_T u^*}\pi_T$$

where $\sigma_{\varepsilon_S u^*}$ (respectively $\sigma_{\varepsilon_T u^*}$, $\sigma_{\nu_S u^*}$ and $\sigma_{\nu_T u^*}$) is the covariance between ε_S (resp. ε_T, ν_S and ν_T) and u^* and π_S and π_T are inverse-Mills ratio, defined as

$$\pi_S = \frac{-\phi(\Psi)}{\Phi(\Psi)} \text{ and } \pi_T = \frac{\phi(\Psi)}{1-\Phi(\Psi)}. \tag{III.9}$$

By including them in equations (III.2), (III.3), (III.4) and (III.5), observe that

$$\mathbb{E}(\log(y_S)|d=0) = \lambda_{S_0} + \boldsymbol{\lambda}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_S} + \sigma_{\varepsilon_S u^*}\pi_S, \mathbb{E}(\log(y_T)|d=1) = \lambda_{T_0} + \boldsymbol{\lambda}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_T} + \sigma_{\varepsilon_T u^*}\pi_T,$$

$\mathbb{E}(\log(t_S)|d=0) = \beta_{S_0} + \boldsymbol{\beta}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_S} + \sigma_{\nu_S u^*}\pi_S$ and $\mathbb{E}(h|d=1) = \beta_{T_0} + \boldsymbol{\beta}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_T} + \sigma_{\nu_S u^*}\pi_T$ thus we obtain the following joint linear model,

$$\log(y_S) = \lambda_{S_0} + \boldsymbol{\lambda}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_S} + \sigma_{\varepsilon_S u^*}\pi_S + \eta_S \quad \text{if } d=0 \tag{III.10}$$

$$\log(y_T) = \lambda_{T_0} + \boldsymbol{\lambda}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_T} + \sigma_{\varepsilon_T u^*}\pi_T + \eta_T \quad \text{if } d=1 \tag{III.11}$$

$$\log(t_S) = \beta_{S_0} + \boldsymbol{\beta}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_S} + \sigma_{\nu_T u^*}\pi_S + \upsilon_S \quad \text{if } d=0 \tag{III.12}$$

$$h = \beta_{T_0} + \boldsymbol{\beta}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_T} + \sigma_{\nu_S u^*}\pi_T + \upsilon_T \quad \text{if } d=1 \tag{III.13}$$

with $\mathbb{E}(\eta_S|d=0) = \mathbb{E}(\eta_T|d=1) = \mathbb{E}(\upsilon_S|d=0) = \mathbb{E}(\upsilon_T|d=1) = 0$. If we have been able to avoid the biased properties of the residual terms, we now face an heteroscedasticity problem since the variance of the residuals is no longer constant. Nevertheless, weighted least squares can be used to solve this problem (see Appendix 8.2).

4.5 Inference

The first step is to estimate γ's coefficients of the equation (III.7), assuming that $\sigma_{u^*}^2 = 1$. Then, Ψ's can be predicted using the $\widehat{\gamma}$'s coefficients. Given Ψ's, estimators of the inverse Mills ratio can then be derived,

$$\widehat{\pi}_{S,i} = \frac{-\phi(\widehat{\Psi}_i)}{\Phi(\widehat{\Psi}_i)} \text{ and } \widehat{\pi}_{T,i} = \frac{\phi(\widehat{\Psi}_i)}{1-\Phi(\widehat{\Psi}_i)} \text{ where } \widehat{\Psi}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 \boldsymbol{x_i} + \widehat{\gamma}_{c,i}$$

Then, the estimators of λ and β's can be obtained running regressions (III.10-III.13) using the selected sample :

$$\log(y_S) = \lambda_{S_0} + \boldsymbol{\lambda}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_S} + \sigma_{\varepsilon_S u^*}\pi_S + \eta_S \quad \text{if } d = 0$$

$$\log(y_T) = \lambda_{T_0} + \boldsymbol{\lambda}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\omega_T} + \sigma_{\varepsilon_T u^*}\pi_T + \eta_T \quad \text{if } d = 1$$

$$\log(t_S) = \beta_{S_0} + \boldsymbol{\beta}_{\boldsymbol{S_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_S} + \sigma_{\nu_S u^*}\pi_S + \upsilon_S \quad \text{if } d = 0$$

$$h = \beta_{T_0} + \boldsymbol{\beta}_{\boldsymbol{T_1}}^\top \boldsymbol{x} + \boldsymbol{\delta_T} + \sigma_{\nu_S u^*}\pi_T + \upsilon_T \quad \text{if } d = 1$$

Note that h is unknown for both types of procedures, therefore $\log(\widehat{t}_S)$ is used to compute the time difference for trial procedures. Further, we need estimators for $\sigma_{\varepsilon_S u^*}$, $\sigma_{\varepsilon_T u^*}$, $\sigma_{\nu_S u^*}$ and $\sigma_{\nu_T u^*}$. Using $\widehat{\lambda}$ and $\widehat{\beta}$, it is possible to estimate residuals. For instance, for settlement cases,

$$\widehat{\zeta}_{y_S,i} = \log(\widehat{y}_{S,i}) - \log(y_{S,i})$$

Using equations for the variance $\sigma^2_{y_S}$, we can consider the following estimator

$$\widehat{\sigma}^2_{\varepsilon_S} = \frac{1}{n_S}\sum_{i\in I_S}(\widehat{\zeta}_{S,i} - \widehat{\sigma}^2_{\varepsilon_S u^*}\widehat{\Psi}_i\widehat{\pi}_{S,i})$$

with similar expressions for $\widehat{\sigma}^2_{y_T}$, $\widehat{\sigma}^2_{t_S}$ and $\widehat{\sigma}^2_h$, where I_S is the subset of out-of-court settlement. The major drawback is that this estimator might be negative. A better estimator can be derived from

$$\mathbb{E}[\eta_S^2 | d_i = 0] = \sigma^2_{\varepsilon_S u^*}\Psi_i\pi_{S,i} + \sigma^2_{y_S} - \sigma^2_{\varepsilon_S u^*}\pi^2_{S,i}$$

i.e.

$$\widetilde{\sigma}^2_{\varepsilon_S} = \frac{1}{n_S}\sum_{i\in I_S}(\widehat{\eta}_S^2 - \widehat{\sigma}^2_{\varepsilon_S u^*}\widehat{\Psi}_i\widehat{\pi}_{S,i} + \widehat{\sigma}^2_{\varepsilon_S}\widehat{\pi}^2_{S,i})$$

with similar expressions for $\widetilde{\sigma}^2_{\varepsilon_T}$, $\widetilde{\sigma}^2_{\nu_S}$ and $\widetilde{\sigma}^2_{\nu_T}$. Then those estimates are plugged-in in Equations (III.10-III.13) and parameters of these equations then be estimated. The last step is to estimate the decision equation (III.6) with $\log(\widehat{y}_S)$, $\log(\widehat{y}_T)$, $\log(\widehat{t}_S)$ and $\widehat{h}$.

$$T_i^* = \theta_0 + \boldsymbol{\theta_1}\widehat{\boldsymbol{h}}_{\boldsymbol{i}} + \boldsymbol{\theta_2}\log(\widehat{\boldsymbol{t}}_{\boldsymbol{S;i}}) + \boldsymbol{\theta_2}(\log(\widehat{\boldsymbol{y}}_{\boldsymbol{T;i}}) - \log(\widehat{\boldsymbol{y}}_{\boldsymbol{S;i}})) + \boldsymbol{\theta_3}^\top\boldsymbol{x_i} + \boldsymbol{\alpha_{c;i}} + u$$

5 Dataset

The dataset is provided by Association pour la Gestion des Informations sur le Risque Automobile (AGIRA). This file has been set up to comply with legal requirements. This enables victims to be aware of the amounts they are entitled to. Statistics by year are presented in the table (III.2) and by procedure in the table (III.3).

The first information we have on accidents is the way the settlement was obtained: Negotiation or Judicial. In addition, we also know whether the amount was settled in the first instance or on appeal. We record 258095 road accident victims between 1997 and 2014, including 243465 (94%) out-of-court settlements, 13004 (5%) in first instance and 1626 (1%) in appeal. Our dataset will be too unbalanced to provide proper estimates if we consider the decision to go to appeal court. Then, claims on appeal are not considered in the analysis. Furthermore, we do not have information to understand which procedure is chosen, civil (75%) or criminal (25%). The proportion of cases going to trial decreases over time. Trials, particularly appeals, require longer procedural time and, at the reporting date, proceedings have not yet been completed for more recent accidents. Compensations at trial are also higher on average, 20448€ for out-of-court settlements, 72022€ for trials and 233688€ for appeals.

Information about the victim is also known: age and sex. Our database is composed of 42% of men, with a higher proportion of men going to trial (6% of men and 5% of women go to trial). The average age of the victims is 44 years old and the share of trials decreases with age (7% of 0-25 year olds, 6% of $26-65$ year olds and 4% of 65 year olds and over go to trial).

The dataset also provides information about the accident: type of victim (Driver, Passenger, Cyclist, Pedestrian), and related to the injury severity of the victim: degree of suffering, partial permanent disability and degree of aesthetic damage. A large number of unknown types are observed (46% of overall observations). This lack of information is mainly due to more recent observations where the data are less detailed (72% of "no details" in 2009-2014). Otherwise, the database is mainly constituted of drivers, then passengers and pedestrians (excluding non-detailed data, there are 42% drivers, 24% passengers, 7% cyclists and 25% pedestrians). Moreover, victims in court have, on average, a higher injury severity. Partial permanent disability is, on average, equal to 9.05 for judicial cases (compared to 4.71 for non-judicial cases), the degree of suffering is 3.17 (compared to 2.41) and the degree of permanent aesthetic damage is 1.25 (compared to 0.65).

We also have information about various lengths, including hospitalization, recovery, procedure and temporary work incapacity. Procedures in court naturally take longer because it is an additional step in the process. In addition, the length of hospitalization, recovery and temporary incapacity to work reflect a greater injury severity and are higher on average at the trial. Judicial proceedings have, on average, 13 days of hospitalization, 524 days of recovery and 147 days of incapacity for work (compared to 3 days of hospitalization, 326 days of recovery and 57 days of incapacity for work for non-judicial proceedings).

The department of the accident is also known. For judicial issues, the victim can choose either the court of the accident or the court of residence. There are 173 High Courts in France, which is more than the granularity we have in our dataset. We therefore consider the Courts of Appeal, of which there are 32, to carry out the analysis. We also remove victims without information on the location of the accident and fatal accidents (the compensation is very different for both procedures).

Table III.2. Summary of the dataset used, mean of the variables per year *(source: AGIRA)*.

	Global	1997-2000	2001-2008	2009-2014
Number of accidents	258095	34999	120091	103005
Amount of Compensation (€)	24390.26	36372.86	28727.88	15261.70
Age (years)	44.32	41.90	44.01	45.51
Male (%)	42.18	46.72	42.92	39.78
Driver (%)	22.79	41.79	31.35	6.35
Passenger (%)	13.05	25.40	15.95	5.48
Biker (%)	4.00	4.46	4.18	3.62
Pedestrian (%)	13.64	17.17	15.22	10.60
No details (%)	46.53	11.19	33.30	73.96
Compromise (%)	94.33	87.56	93.13	98.03
First degree (%)	5.04	10.74	6.09	1.88
Second degree (%)	0.63	1.70	0.78	0.09
Civil (%)	75.10	72.21	76.55	75.38
Length of Hospitalization (days)	3.58	2.39	3.96	3.53
Length of Recovery (days)	338.71	458.78	374.90	255.73
Length of Procedure (days)	386.30	502.00	420.68	306.90
Length Total (days)	725.01	960.78	795.58	562.63
Length of temporary work incapacity (days)	63.06	108.64	77.57	30.66
Partial permanent disability (IPP)	5.01	7.24	5.55	3.62
Degree of suffering	2.46	2.85	2.60	2.16
Degree of permanent aesthetic damage	0.70	0.81	0.70	0.61
Victim Salary (%)	18.91	18.96	17.66	20.34

Table III.3. Summary of the dataset used, mean of the variables per type of procedure *(source: AGIRA)*.

	Compromise	1st degree	2nd degree
Number of accidents	243465	13004	1626
Amount of Compensation (€)	20448.32	72022.43	233687.90
Age (years)	44.57	40.42	38.14
Male (%)	41.95	45.45	51.11
Driver (%)	22.59	26.36	22.94
Passenger (%)	12.98	14.39	12.24
Biker (%)	3.99	4.13	3.75
Pedestrian (%)	13.36	18.68	15.38
No details (%)	47.07	36.44	45.69
Civil (%)	-	81.25	25.89
Length of Hospitalization (days)	2.96	12.52	24.13
Length of Recovery (days)	325.75	526.38	778.48
Length of Procedure (days)	355.55	832.98	1417.53
Length Total (days)	681.30	1359.36	2196.00
Length of temporary work incapacity (days)	57.09	147.15	283.76
Partial permanent disability (IPP)	4.71	9.05	17.92
Degree of suffering	2.41	3.17	3.91
Degree of permanent aesthetic damage	0.65	1.25	1.98
Victim Salary (%)	18.65	22.80	25.65

6 Results

As previously explained, five models are estimated here: Equation (III.6) models the decision to go to court, Equation (III.2) and Equation (III.3) respectively the out of court and the court (log) amounts, Equation (III.4) the (log) length of the out of court procedure and finally Equation (III.5) the difference (in log) between the two procedure lengths. In order to have interpretable results for the court variable, we use the strategy of using contrasts. Geographical effects are estimated without the intercept to ensure identifiability. The equation can then be written with centered parameters (see Rosnow et al. (2000) and Appendix 8.3). The various lengths are transformed using a logarithmic function (if the length is positive, 0 otherwise). The age and the partial permanent disability are transformed into classes (age: "0-25", "26-65" and "65+" and IPP: "0-25", "26-70", "70+"). Tables III.4 and III.5 (for court coefficients) give the parameter estimators of the five models.

The results suggest that the gender of the victim has no impact on the decision to go to court. Men and women have, other things being equal, an equal probability of going to court or not. However, gender affects the determination of amounts. Men tend to have higher amounts than women, both at trial and out-of-court.

Age has a significant impact on the decision to go to trial. Older people, compared to younger people, have a lower probability of going to trial. They also have, ceteris paribus, higher amounts for both procedures.

The decision to go to trial also differs according to the type of victim: passengers tend to go to trial less than drivers, and pedestrians and cyclists tend to go to trial more. The amounts are also higher for these persons, with the exception of out-of-court pedestrians.

Finally, the complexity of the accident has a positive impact on the decision to go to court. The length of hospitalization, recovery or incapacity increases the probability of going to court. It also increases the amounts received in both cases. In addition, the suffering endured is positively related to the probability of going to court and the amounts proposed. Accidents resulting in permanent disability have a particular effect: the most disabled cases are less likely to be settled in court, compared to IPP between 0 and 25, but the amounts are higher in both procedures.

Table III.4. Parameter estimates for the five regression models

	Dependent variable:				
	Procedure	log(Amount)		log(Length Procedure)	Time difference
	Probit	Compromise	Judicial	Compromise	
	(1)	(2)	(3)	(4)	(5)
Intercept	282.07***	−260.90***	−64.27***	20.15***	-13.19***
Court Location			See Table (III.5)		
Time Difference	−0.31***				
	(0.06)				
log(Settlement date)	−0.08***				
	(0.02)				
log(Amount) diff.	0.26***				
	(0.04)				
log(Amount) diff. (+)	−0.36***				
	(0.06)				
Male	0.001	0.06***	0.05***	−0.01***	0.001
	(0.02)	(0.002)	(0.012)	(0.002)	(0.009)
Age 26-65 (ref: 0-25)	−0.01	0.08***	0.10***	−0.04***	0.07***
	(0.02)	(0.003)	(0.013)	(0.002)	(0.001)
Age 65+	−0.42***	0.40***	0.11***	−0.18***	0.11***
	(0.03)	(0.011)	(0.021)	(0.006)	(0.017)
No details (ref: Driver)	0.42***	−0.16***	0.10***	0.002	0.21***
	(0.03)	(0.009)	(0.015)	(0.005)	(0.014)
Passenger	−0.13***	0.18***	−0.01	0.01***	0.04**
	(0.03)	(0.005)	(0.019)	(0.003)	(0.014)
Biker	0.13**	0.03***	0.12***	0.08***	0.01
	(0.05)	(0.007)	(0.03)	(0.005)	(0.023)
Pedestrian	0.34***	−0.05***	0.09***	0.08***	0.01
	(0.03)	(0.009)	(0.02)	(0.006)	(0.015)
IPP 26-70 (ref: 0-25)	−0.30***	1.24***	0.93***	0.16***	−0.12***
	(0.05)	(0.014)	(0.027)	(0.009)	(0.018)
IPP 70+	−0.81***	2.51***	2.04***	0.30***	−0.21***
	(0.13)	(0.04)	(0.01)	(0.025)	(0.043)
Degree of Sufferring	0.41***	0.37***	0.62***	0.07***	−0.06***
	(0.01)	(0.01)	(0.01)	(0.004)	(0.008)
log(Length Hospitalization)	0.21***	0.06***	0.16***	0.06***	−0.06***
	(0.01)	(0.006)	(0.007)	(0.003)	(0.006)
log(Length Recovery)	0.23***	0.08***	0.25***	0.39***	−0.11***
	(0.02)	(0.005)	(0.009)	(0.002)	(0.002)
log(Length Incapacity)	0.02***	0.07***	0.09***	0.0004	−0.004*
	(0.005)	(0.001)	(0.003)	(0.0006)	(0.002)
Year of Accident	−0.14***	0.14***	0.04***	−0.01***	0.01***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)
Salary	0.18***				
	(0.02)				
IMR0		1.60***		0.07**	
		(0.054)		(0.028)	
IMR1			−0.32**		0.63***
			(0.135)		(0.096)
Observations	256 469	243 465	13 004	243 465	13 004
R^2		0.7317	0.7551	0.411	0.107
Adjusted R^2		0.7316	0.7542	0.4109	0.104
Log Likelihood	−45 234				
Akaike Inf. Crit.	90 564				
Residual Std. Error		0.54	0.63	0.45	0.48
F Statistic		15 434 ***	929.1***	3951***	36.09***

Note: *p<0.1; **p<0.05; ***p<0.01

6.1 Geographical Aspects

Table III.5. Estimation of fixed effects (court location variable).

Standards errors are computed using bootstraps of 1000 simulations.

	(1)	(2)	(3)	(4)	(5)
Agen	0.02	-0.01	-0.002	-0.03***	0.05
	(0.082)	(0.058)	(0.042)	(0.008)	(0.035)
Aix-en-Provence	0.52***	-0.39***	-0.06***	0.002	-0.10***
	(0.018)	(0.108)	(0.015)	(0.017)	(0.013)
Amiens	-0.05	0.10**	0.01	0.03***	-0.02
	(0.064)	(0.044)	(0.035)	(0.007)	(0.03)
Angers	-0.03	0.10**	0.09**	-0.04***	0.06**
	(0.077)	(0.053)	(0.046)	(0.008)	(0.031)
Besançon	-0.07	0.12**	0.01	0.04***	-0.04
	(0.073)	(0.051)	(0.043)	(0.008)	(0.033)
Bordeaux	-0.05	0.05	0.13***	-0.02***	0.03
	(0.048)	(0.035)	(0.028)	(0.005)	(0.023)
Bourges	-0.03	0.11	0.06	-0.07***	0.03
	(0.098)	(0.068)	(0.056)	(0.011)	(0.042)
Caen	0.12*	-0.03	0.10***	0.0002	-0.06**
	(0.06)	(0.048)	(0.035)	(0.008)	(0.029)
Chambéry	-0.07	0.05	-0.03	-0.04***	0.08**
	(0.074)	(0.054)	(0.047)	(0.008)	(0.033)
Colmar	-0.11	0.17***	0.05	0.03***	0.09**
	(0.081)	(0.062)	(0.043)	(0.009)	(0.038)
Dijon	-0.30***	0.21***	-0.02	-0.03**	0.03
	(0.067)	(0.074)	(0.045)	(0.012)	(0.032)
Douai	-0.06	0.08**	0.01	0.04***	0.03
	(0.048)	(0.035)	(0.027)	(0.006)	(0.022)
Grenoble	-0.40***	0.23***	-0.09**	-0.03**	0.08**
	(0.066)	(0.085)	(0.043)	(0.014)	(0.031)
Limoges	-0.23**	0.17**	0.04	-0.10***	0.08*
	(0.105)	(0.082)	(0.058)	(0.012)	(0.048)
Lyon	-0.29***	0.17***	-0.04	0.03***	-0.03
	(0.043)	(0.061)	(0.026)	(0.009)	(0.021)
Metz	0.05	0.08	-0.01	0.002	0.17***
	(0.075)	(0.052)	(0.044)	(0.009)	(0.044)
Montpellier	-0.07	0.12***	0.05*	0.004	0.04*
	(0.053)	(0.041)	(0.03)	(0.006)	(0.025)
Nancy	0.15**	-0.02	0.13***	-0.01	-0.07**
	(0.065)	(0.055)	(0.038)	(0.009)	(0.031)
Nîmes	0.02	-0.001	-0.04	-0.02***	0.02
	(0.054)	(0.037)	(0.032)	(0.005)	(0.026)
Orléans	-0.60***	0.33***	-0.07	0.004	0.01
	(0.085)	(0.126)	(0.053)	(0.02)	(0.043)
Paris	-0.12***	0.08**	0.05***	0.08***	0.05***
	(0.026)	(0.035)	(0.017)	(0.006)	(0.013)
Pau	0.09	-0.10**	0.01	-0.09***	0.05*
	(0.06)	(0.044)	(0.039)	(0.007)	(0.029)
Poitiers	-0.14*	0.15***	0.09**	-0.03***	0.01
	(0.07)	(0.053)	(0.044)	(0.008)	(0.029)
Reims	-0.10	0.11*	-0.04	-0.04***	0.01
	(0.082)	(0.059)	(0.055)	(0.009)	(0.039)
Rennes	-0.40***	0.25***	0.003	-0.05***	0.07***
	(0.042)	(0.085)	(0.028)	(0.013)	(0.021)
Riom	0.16**	-0.06	0.04	-0.04***	-0.04*
	(0.067)	(0.057)	(0.04)	(0.01)	(0.027)
Rouen	-0.08	0.08*	-0.02	0.03***	-0.003
	(0.06)	(0.045)	(0.029)	(0.007)	(0.026)
Toulouse	-0.35***	0.21***	-0.04	-0.06***	0.02
	(0.05)	(0.078)	(0.034)	(0.012)	(0.023)
Versailles	-0.14***	0.06	0.06*	0.03***	0.01
	(0.052)	(0.046)	(0.035)	(0.007)	(0.025)

Note: *p<0.1; **p<0.05; ***p<0.01

On Figures (III.5), (III.6) and (III.7), we can visualize the centered court location effects on French map (respectively for the Decision Equation, the Amount Equations and the Time Equations) and on Figure (III.8) with confidence interval.

The probability to go to court in Rennes, Grenoble or Orléans is significantly smaller than average situation in France. On the other hand, the probability to go to court in Aix-en-Provence, Riom or Nancy is significantly larger than the average situation in France, ceteris paribus.

Amounts settled out of court are also higher in Rennes, Grenoble or Orléans compared to the average situation in France, while the opposite occurs in the courts of Aix-en-Provence. In Riom and Nancy, the negotiated amounts do not significantly differ from the French average.

The situation is a slightly different for trial amounts. The court of Nancy has in fact proposed higher amounts compared to the rest of France, whereas in Aix-en-Provence the effect is significant negative. The impact of the courts of Rennes and Orléans is not significant on proposed judicial amounts compared to the rest of France and Grenoble has a significant negative effect.

Therefore, the amounts settled out of court are more likely to be higher in areas where the probability of going to court is lower than in the rest of the country. Conversely, no remarkable pattern appears on the amounts settled at trial.

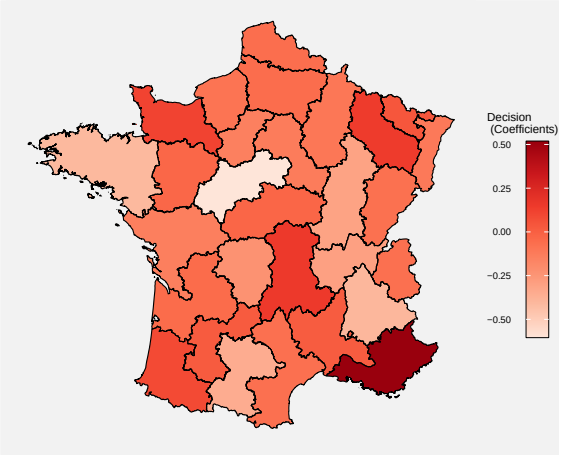


Figure III.5. Estimators of the regional categorical effects (centered on the average entire French population) for the decision Equation (III.6).

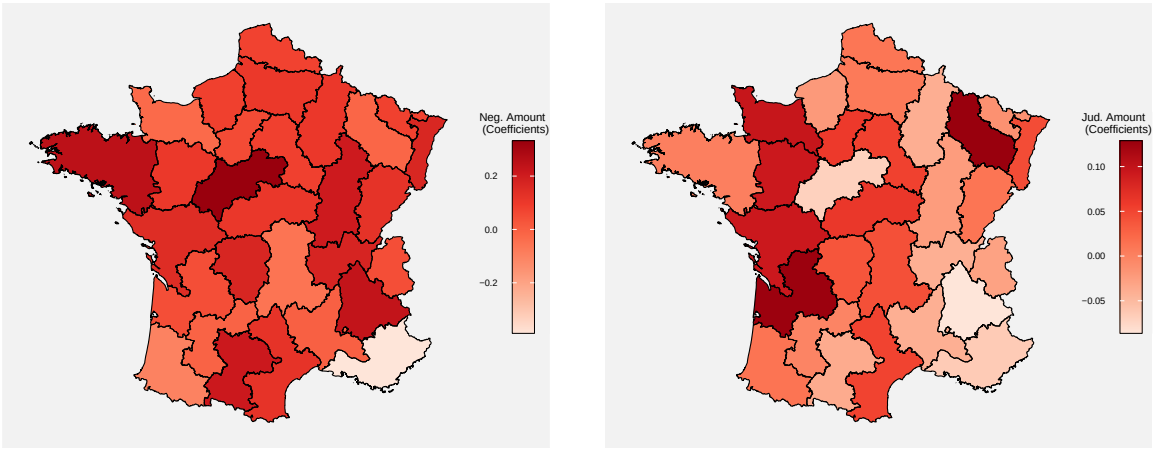


Figure III.6. Estimators of the regional categorical effects (centered on the average entire French population) for the negotiation amount Equation (III.2) in the left and for the judicial Equation (III.3) on the right.

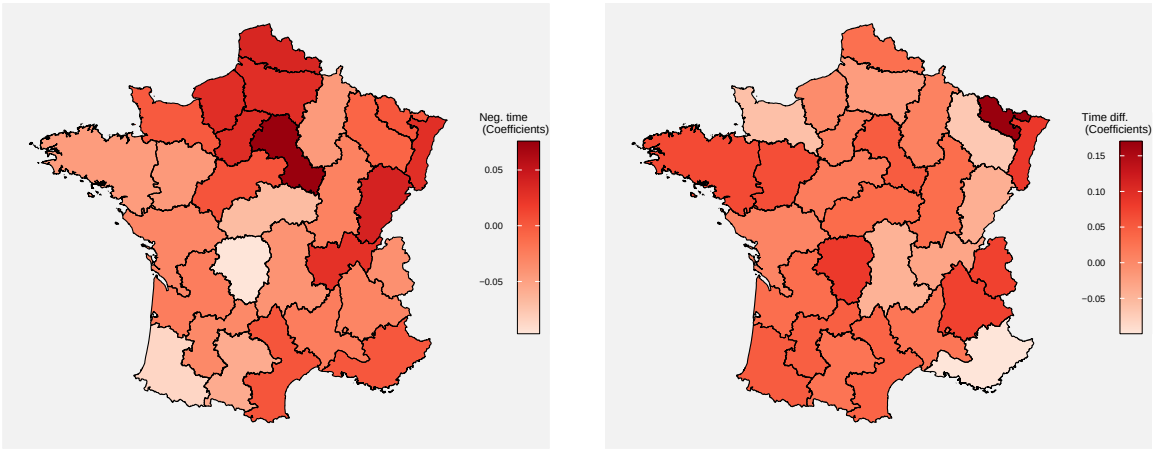


Figure III.7. Estimators of the regional categorical effects (centered on the average entire French population) for the negotiation length procedure Equation (III.4) in the left and for the time difference Equation (III.5) on the right.

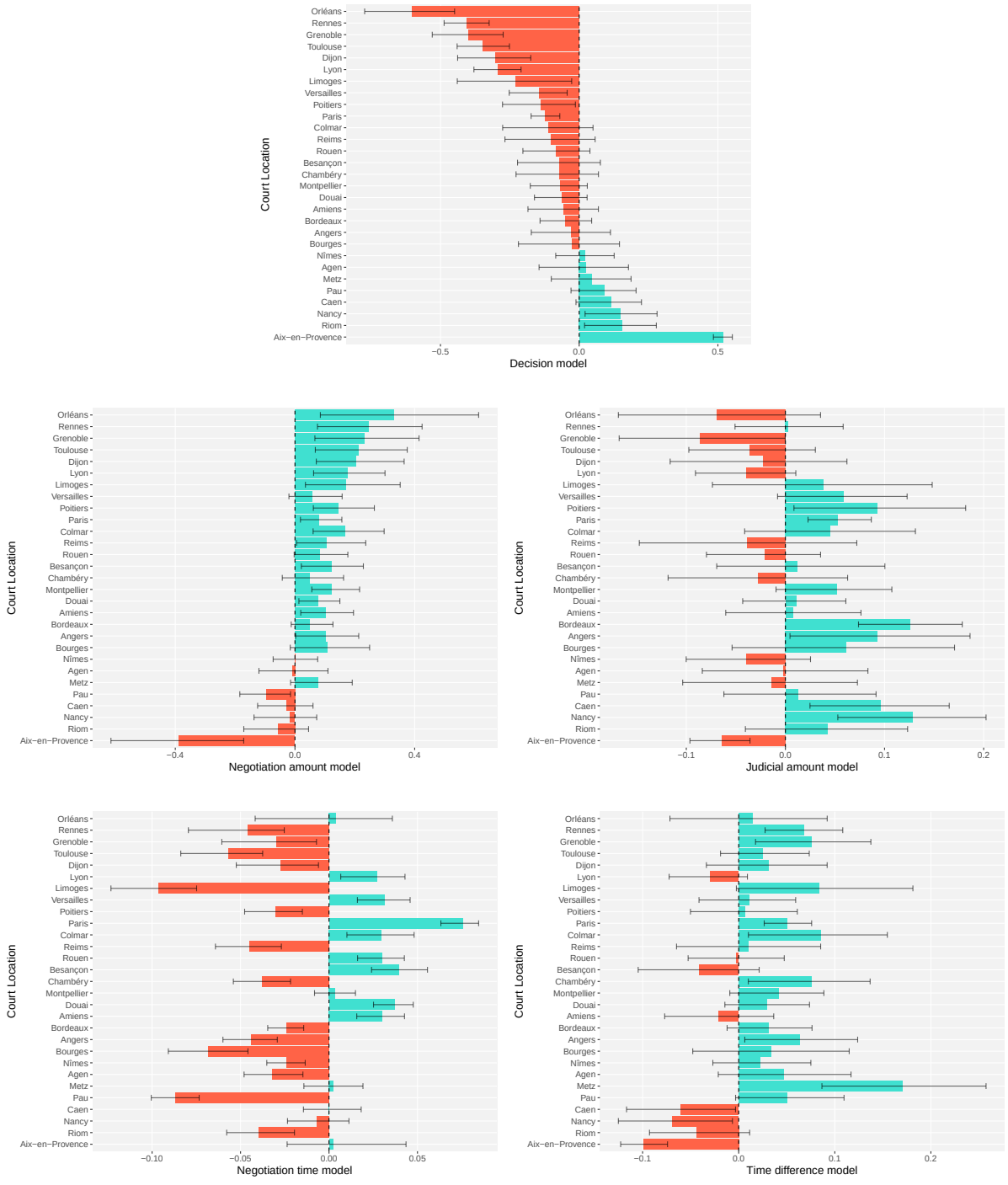


Figure III.8. Estimators of the regional categorical effects (centered on the average entire French population) for the five Equations. Court locations are ordered according to the decision equation coefficient. 95% confidence intervals are computed using bootstraps of 1000 simulations (2.5% and 97.5% quantiles).

7 Conclusion

In France, the law of 5 July 1985 established the concept of integral compensation for victims of road accidents. The compensation scheme for victims is described in detail in the law, and includes the extent of personal injury and property damage. This compensation must reflect both patrimonial damage, such as health expenses or loss of income, and extra-patrimonial damage, such as suffering or pain. The Act requires the insurer to initially make an out-of-court offer to the victim; nevertheless, the victim can sue and a judge will fix the amount of compensation (whether in civil or criminal court). In France, nearly 95% of victims are compensated through out-of-court settlements.

For this article, we obtained a data set of 258,095 victims of accidents that occurred between 1997 and 2014, with the final amount obtained by the victim as compensation. Our aim is to understand the decision to sue or not to sue. More specifically, we examined several individual and accident characteristics (age of the victim, severity of the accident, etc.) and geographical aspects on the amounts obtained, the duration and the decision to go to court. For this purpose, we proposed a limited dependent model suggested by Maddala (1986). The available data on amount or time are closely related to the individual's choice of whether or not to sue, thus the variables are limited dependent and a selection model is needed. A five-equation model is developed: a decision equation, two amount equations (judicial and extra-judicial) and two time equations (extra-judicial and inter-procedural).

Thus, the results suggest that individual characteristics impact both the decision and the amounts. Older individuals tend to go to court less but have higher amounts than younger individuals. However, no difference occurs between men and women on the decision to sue, but men have higher proposed amounts. We also find that the most severe and complex accidents have a higher probability of being settled in court and the compensations to the victims are also higher. In particular, the decision of litigation is significantly related positively to the degree of suffering and the length of hospitalisation and recovery, but negatively to the partial permanent disability. In addition, some systematic differences exist between the courts on the decision and the compensations. For instance, victims are more likely to settle a bodily injury claim in Aix-en-Provence than in Orléans. For extrajudicial amounts, the order is kept for these regions: in a region where out-of-court amounts are higher, victims will be reluctant to bring a lawsuit. The picture of judicial amounts is more contrasted and no consistent pattern seems to emerge.

Several limitations can be acknowledged in our study. Time preferences and risk aversion may not be elicited by any data in the database. In an attempt to capture these preferences, some related individual characteristics are considered. Nevertheless, no reliable measures of risk aversion and time preference can be included. Finally, although this study deals with a large number of road accidents, two aspects of compensation bodily injury claim are neglected. First, the analysis focuses only on the negotiation and the first instance trial. The appeal cases are excluded from both the theoretical and econometric models, as they represent an extremely low ratio in the database. Second, only non-fatal accidents are discussed here. For fatal accidents, the amounts are calculated very differently and are not based on the same scales and values. Fatal accidents can also be very controversial, since the aim is to "value" human life.

Despite these limitations, our work has implications for the literature on bodily injury claims. This paper highlights the determinants that influence the decision to prosecute and the amounts offered to victims at each stage. Several variables are pertinent in the determination of the amount and the decision of individuals to sue. Nevertheless, the scales and criteria used are not fixed and well-defined. The amounts, even between two cases that seem close, can be very different. In particular, we note that, even in a country where the law is the same everywhere, regional specificities emerge.

8 Appendix

8.1 Distribution of $(\varepsilon_S, \varepsilon_T, \nu_S, \nu_T)$ given u^*

The conditional distribution of $(\varepsilon_S, \varepsilon_T, \nu_S, \nu_T)$ given u^* is a joint Gaussian distribution

$$\begin{pmatrix} \varepsilon_S \\ \varepsilon_T \\ \nu_S \\ \nu_T \end{pmatrix} \Big| u^* \sim \mathcal{N}\left(\begin{pmatrix} \rho_{\varepsilon_S u^*} u^* \\ \rho_{\varepsilon_T u^*} u^* \\ \rho_{\nu_S u^*} u^* \\ \rho_{\nu_T u^*} u^* \end{pmatrix}, \begin{pmatrix} \sqrt{\sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*}} & \rho_{\varepsilon_S \varepsilon_T | u^*} & 0 & 0 \\ \rho_{\varepsilon_S \varepsilon_T | u^*} & \sqrt{\sigma^2_{\varepsilon_T} - \rho^2_{\varepsilon_T u^*}} & 0 & 0 \\ 0 & 0 & \sqrt{\sigma^2_{\nu_S} - \rho^2_{\nu_S u^*}} & \rho_{\nu_S \nu_T | u^*} \\ 0 & 0 & \rho_{\nu_S \nu_T | u^*} & \sqrt{\sigma^2_{\nu_T} - \rho^2_{\nu_T u^*}} \end{pmatrix} \right),$$

where $\rho_{\varepsilon_S u^*}$, $\rho_{\varepsilon_T u^*}$, $\rho_{\nu_S u^*}$ and $\rho_{\nu_T u^*}$ are correlations coefficients (respectively $\text{cor}[\varepsilon_S, u^*]$, $\text{cor}[\varepsilon_T, u^*]$, $\text{cor}[\nu_S, u^*]$ and $\text{cor}[\nu_T, u^*]$) while $\rho_{\varepsilon_S \varepsilon_T | u^*}$ and $\rho_{\nu_S \nu_T | u^*}$ denotes the partial correlations, defined as:

$$\rho_{\varepsilon_S \varepsilon_T | u^*} = \frac{\rho_{\varepsilon_S \varepsilon_T} - \rho_{\varepsilon_S u^*} \rho_{\varepsilon_T u^*}}{\sqrt{1 - \rho^2_{\varepsilon_S u^*}} \sqrt{1 - \rho^2_{\varepsilon_T u^*}}}$$

and

$$\rho_{\nu_S\nu_T|u^*} = \frac{\rho_{\nu_S\nu_T} - \rho_{\nu_S u^*}\rho_{\nu_T u^*}}{\sqrt{1-\rho^2_{\nu_S u^*}}\sqrt{1-\rho^2_{\nu_T u^*}}}$$

Thus, one can derive

$$\mathbb{E}[\varepsilon_S] = \mathbb{E}[\mathbb{E}[\varepsilon_S|u^*]] = \rho_{\varepsilon_S u^*}\mathbb{E}[u^*],$$

$$\mathbb{E}[\varepsilon_S^2] = \mathbb{E}[\mathbb{E}[\varepsilon_S^2|u^*] = \rho^2_{\varepsilon_S u^*}\mathbb{E}[u^{*2}] + \sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*}$$

and

$$\mathbb{E}[\varepsilon_S u^*] = \mathbb{E}[u^*\mathbb{E}[\varepsilon_S|u^*]] = \rho_{\varepsilon_S u^*}\mathbb{E}[u^{*2}],$$

with similar expressions when ε_T, ν_S and ν_T are substituted to ε_S. Thus, further

$$\sigma^2_{\varepsilon_S} = \mathbb{E}[\varepsilon_S^2] - \mathbb{E}[\varepsilon_S]^2 = \rho^2_{\varepsilon_S u^*}\mathbb{E}[u^{*2}] + \sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*} - \rho^2_{\varepsilon_S u^*}\mathbb{E}[u^*]^2 = \rho^2_{\varepsilon_S u^*}\sigma^2_{u^*} + \sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*},$$

and the covariance $\sigma_{\varepsilon_S u^*}$ is

$$\sigma_{\varepsilon_S u^*} = \mathbb{E}[\varepsilon_S u^*] - \mathbb{E}[\varepsilon_S]\mathbb{E}[u^*] = \rho_{\varepsilon_S u^*}\mathbb{E}[u^{*2}] - \rho_{\varepsilon_S u^*}\mathbb{E}[u^*]^2 = \rho_{\varepsilon_S u^*}\sigma^2_{u^*}$$

since $\sigma^2_{u^*}$ is standardized to 1, and $\sigma_{\varepsilon_S u^*} = \rho_{\varepsilon_S u^*}$ with similar expressions when ε_T, ν_S and ν_T are substituted to ε_S.

8.2 Variance of the residuals

From properties of the truncated normal distribution,

$$\mathbb{E}[u^{*2}|\Psi > u^*] = 1 + \Psi\pi_S \text{ and } \mathbb{E}[u^{*2}|\Psi \leqslant u^*] = 1 + \Psi\pi_T$$

so finally,

$$\mathbb{E}[\varepsilon_S^2|d=0] = \mathbb{E}[\varepsilon_S^2|\Psi > \epsilon^*] = \mathbb{E}[\rho^2_{\varepsilon_S u^*}u^{*2} + \sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*}|\Psi > u^*]$$

i.e.

$$\mathbb{E}[\varepsilon_S^2|d=0] = \rho^2_{\varepsilon_S u^*}(1+\Psi\pi_S) + \sigma^2_{\varepsilon_S} - \rho^2_{\varepsilon_S u^*} = \sigma^2_{\varepsilon_S u^*}\Psi\pi_S + \sigma^2_{\varepsilon_S}$$

while similarly $\mathbb{E}[\varepsilon_T^2|d=1] = \sigma^2_{\varepsilon_T u^*}\Psi\pi_T + \sigma^2_{\varepsilon_T}$, $\mathbb{E}[\nu_S^2|d=0] = \sigma^2_{\nu_S u^*}\Psi\pi_S + \sigma^2_{\nu_S}$ and $\mathbb{E}[\nu_T^2|d=1] = \sigma^2_{\nu_T u^*}\Psi\pi_T + \sigma^2_{\nu_T}$. Hence, we can now derive expressions for variances of the residuals,

$$\text{Var}[\eta_S|d=0] = \mathbb{E}[\eta_S^2|d=0] - \mathbb{E}[\eta_S|d=0]^2 = \mathbb{E}[(\varepsilon_S - \sigma_{\varepsilon_S u^*}\pi_S)^2|d=0]$$

i.e.

$$\text{Var}[\eta_S|d=0] = \sigma^2_{\varepsilon_S u^*}\Psi\pi_S + \sigma^2_{\varepsilon_S} - 2(\sigma_{\varepsilon_S u^*}\pi_S)(\sigma_{\varepsilon_S u^*}\pi_S) + (\sigma_{\varepsilon_S u^*}\pi_S)^2 = \sigma^2_{\varepsilon_S u^*}\Psi\pi_S + \sigma^2_{\varepsilon_S} - \sigma^2_{\varepsilon_S u^*}\pi_S^2$$

while similarly $\text{Var}[\eta_T|d=1] = \sigma^2_{\varepsilon_T u^*}\Psi\pi_T + \sigma^2_{\varepsilon_T} - \sigma^2_{\varepsilon_T u^*}\pi_T^2$, $\text{Var}[\upsilon_S|d=0] = \sigma^2_{\nu_S u^*}\Psi\pi_S + \sigma^2_{\nu_S} - \sigma^2_{\nu_S u^*}\pi_S^2$ and $\text{Var}[\upsilon_T|d=1] = \sigma^2_{\nu_T u^*}\Psi\pi_T + \sigma^2_{\nu_T} - \sigma^2_{\nu_T u^*}\pi_T^2$.

8.3 The Spatial Component

Consider a linear regression of Y on variables $\boldsymbol{X} = (X_1, \boldsymbol{X}_{(1)})$, where X_1 is some categorical variable (say the region), taking values in $R = \{R_1, \cdots, R_k\}$. Consider a linear regression without an intercept so that no reference is considered, and k parameters are estimated

$$Y = \sum_{j=1}^{k} \beta_{1,j}\mathbf{1}_{X_1=R_j} + \boldsymbol{\beta}_{(1)}^{\mathsf{T}}\boldsymbol{X}_{(1)} + \varepsilon. \tag{*}$$

Let $\widehat{\boldsymbol{\beta}}$ denote OLS-estimators (or equivalently MLE estimators assuming Gaussian residuals). Equivalently, one can consider the following equivalent notation

$$Y = \mu + \sum_{j=1}^{k} [\beta_{1,j} - \mu]\mathbf{1}_{X_1=R_j} + \boldsymbol{\beta}_{(1)}^{\mathsf{T}}\boldsymbol{X}_{(1)} + \varepsilon. \tag{*'}$$

These two models are strictly equivalent, whatever μ. Let $\widetilde{\beta}_{1,j} = \widehat{\beta}_{1,j} - \mu$. This is OLS-estimator of $\beta_{1,j} - \mu$ in Equation $(*')$

One classical case is the one where μ is equal to one specific parameter (say β_{1,j_0}), so that region j_0 is the reference : all regions will be compared to j_0. Testing significance with a Student t test is the test of $H_0 : \beta_{1,j} = \beta_{1,j_0}$ (against $H_0 : \beta_{1,j} \neq \beta_{1,j_0}$). It is the default technique used in many statistical softwares.

Another popular case is the one where μ is chosen so the average of all parameters is zero

(also called 'centering' strategy). In that case

$$\mu = \overline{\beta}_1 = \sum_{j=1}^{k} \omega_j \widehat{\beta}_{1,j} \text{ where } \omega_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_{1,i}=R_j}$$

The model was estimated without the intercept, to keep all the regions, as in $(*)$. In order to have comparable results, in Table III.5 coefficients have been centered, with $\widetilde{\theta}_{1,j}$'s. Thus, the reference is the the average situation in the entire population. Here, significance is mentioned with respect to that reference. It is based on Student t test on $\widetilde{\theta}_{1,j}$'s with the first order approximation $\text{Var}[\widetilde{\theta}_{1,j}] \sim \text{Var}[\widehat{\theta}_{1,j}]$. See Chapter 13 of Seltman (2015) for further discussion.

Bibliography

L. Arrondel and A. Masson. Risk and time preferences: Saver types. 2005.

L. Arrondel and A. Masson. Measuring savers' preferences how and why? 2013.

L. A. Bebchuk. Litigation and settlement under imperfect information. *The RAND Journal of Economics*, pages 404–415, 1984.

C. Brown. Deterrence in tort and no-fault: the new zealand experience. *Calif. L. Rev.*, 73:976, 1985.

M. J. Browne and J. T. Schmit. Litigation patterns in automobile bodily injury claims 1977–1997: Effects of time and tort reforms. *Journal of Risk and Insurance*, 75(1):83–100, 2008.

J.-P. Chauchard. La transaction dans l'indemnisation du préjudice corporel. pages 1–39. 1989.

J. D. Cummins and M. A. Weiss. The effects of no fault on automobile insurance loss costs. *Geneva Papers on Risk and Insurance. Issues and Practice*, pages 20–38, 1991.

J. D. Cummins, R. D. Phillips, and M. A. Weiss. The incentive effects of no-fault automobile insurance. *The Journal of Law and Economics*, 44(2):427–464, 2001.

R. A. Devlin. Determinants of no-fault insurance measures. *Journal of Risk and Insurance*, 69 (4):555–576, 2002.

J.-P. Dintilhac. Rapport du groupe de travail chargé d'élaborer une nomenclature des préjudices corporels, 2005. URL https://solidarites-sante.gouv.fr/IMG/pdf/Rapport_groupe_de_travail_nomenclature_des_prejudices_corporels_de_Jean-Pierre_Dintilhac.pdf.

A. Farmer and P. Pecorino. Issues of informational asymmetry in legal bargaining. *Dispute resolution: Bridging the settlement gap*, 2:79–105, 1996.

J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

L.-F. Lee. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International economic review*, pages 415–433, 1978.

L.-F. Lee. Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica: Journal of the Econometric Society*, pages 977–996, 1979.

Y.-P. Liao and M. J. White. No-fault for motor vehicles: An economic analysis. *American Law and Economics Review*, 4(2):258–294, 2002.

G. S. Maddala. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.

B. Nalebuff. Credible pretrial negotiation. *The RAND Journal of Economics*, pages 198–210, 1987.

E. Osborne. Courts as casinos? an empirical investigation of randomness and efficiency in civil litigation. *The Journal of Legal Studies*, 28(1):187–203, 1999a.

E. Osborne. Who should be worried about asymmetric information in litigation? *International Review of Law and Economics*, 19(3):399–409, 1999b.

G. L. Priest and B. Klein. The selection of disputes for litigation. *The Journal of Legal Studies*, 13(1):1–55, 1984.

R. L. Rosnow, R. Rosenthal, and D. B. Rubin. Contrasts and correlations in effect-size estimation. *Psychological science*, 11(6):446–453, 2000.

M. Santolino. Determinants of the decision to appeal against motor bodily injury judgements made by spanish trial courts. *International Review of Law and Economics*, 30(1):37–45, 2010.

H. J. Seltman. Experimental design and analysis. *Retrieved from*, 2015.

M. Solignac. "Ubi lex distinguit, distinguere debemus", une approche economique de l'indemnisation des dommages corporels. 2014.

S. D. Sugarman. A century of change in personal injury law. *Calif. L. Rev.*, 88:2403, 2000.

S. D. Sugarman. Compensation for accidental personal injury: What nations might learn from each other. *Pepp. L. Rev.*, 38:597, 2010.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

W. K. Viscusi. Product liability litigation with risk aversion. *The Journal of Legal Studies*, 17 (1):101–121, 1988.

General Conclusion

Unfortunately, the economic data are not perfect. Econometrics aims to reduce the gap between economic models and data. Imperfect data of any kind should be duly included in the models to avoid any misleading analysis or results. In particular, the thesis aims to provide solutions to the issue of aggregated and limited dependent data. The reforms of personal data protection, the inherent nature of data and the growth of Big Data are introducing greater constraints on the treatment and analysis of the data encountered.

The first chapter examines the topic of compositional data. Very common in economics, they are not often considered correctly. This chapter provides an overview of suitable methods with compositions. More precisely, the use of simplex geometry and linear and non-linear regressions on compositions is explained. Understanding the particular nature of the compositions enables a more coherent and intuitive interpretation of the results. In many cases, this seems more relevant even if the results and forecasts remain locally close.

The second chapter explores income inequality measures. It proposes a methodology for adapting traditional methods to the potential data encountered. Personal income is a private datum and, as such, is rarely accessible. Especially when dealing with restricted areas, the only information available should be the quantiles (deciles, quartiles, etc.) of income. The proposed approach is to reconstruct income shares using conditional expectations and to approximate a Lorenz curve with conventional methods. It allows to describe inequalities when data are constrained and outperforms the traditional midpoint method.

Finally, the third and last chapter analyses the process of the French traffic accident compensation system. This three-stage system is a decision problem; therefore, the data are limited dependent and the model is a decision model. An adapted approach, using a five-equation model, helps to identify the determinants involved in the decision to prosecute and in the determination of compensation amounts for victims.

A variety of alternative solutions are proposed in these three chapters to deal with imperfect data. The imperfect and incomplete aspect of the data raises several problems, from the conception to the interpretability of a statistical model. Hence, the goal is to adapt or create new methodologies to comply with the nature of the data. This adaptability can occur upstream of modelling by transforming the data to a compatible input of the model (Chapter I). Alternatively, it can be achieved on a part of the model and then apply conventional theory afterwards (Chapter II). Finally, it can be done by applying and developing a new model in order to include the specificities of the data in the model (Chapter III).

Such models ensure a suitable analysis of the data. The data encountered may be inherently imperfect, and the model is therefore the most appropriate. However, some data are subject to alterations and there are in fact more detailed and "perfect" data, although they are not available or costly. These models help to adapt and fit the reality, but they can not outperform or even equal the models on more realistic data. An adequate understanding of the limits of the model and an appropriate interpretation is then essential.

In addition, the difference between adapted and classical methodologies is sometimes negligible. The implementation of this type of model, which is more complex and more demanding, could be debatable. Identifying the purposes of the study is therefore necessary in order to perceive the potential problems of each modelling. The importance given to imperfect data in the analysis can be used to judge whether or not further consideration of these data is necessary.

This thesis leads to a better understanding and documentation of the problem of imperfect data. The important point is to recognise the issues and problems related to the data. Data, as perfect as they may seem, can never capture the whole real world. No one can survey an entire population, at best a sample, or have complete knowledge of a phenomenon, at best an overview. The real world is complex and involves many parameters to be measured. The idea is therefore to reflect reality as closely as possible and to define the scope and stakes of a study.

In the future, further perspectives and developments of this research can be envisaged. First,

this thesis considers only two types of imperfect data. In the economic world, other types of data are problematic and need to be overcome. The panorama of imperfect data leads to questions about missing data and imputation mechanisms, accuracy of data tainted by measurement errors or even outlier treatment. The consideration of data consistency in models is relevant to prevent statistical bias. In addition, the analysis of models on imperfect data can be extended to measure the information loss due to the substitution of such variables. These measures could provide a more complete picture of the sometimes necessary use of these models and would reinforce the importance of dealing with the characteristics of the data. Finally, the analysis of spatial aggregation could be extended. The imposed areas – departments, districts, cantons, cities – are intrinsically linked to the demographic characteristics of their population. Increasingly, new zones are designed to counteract the reciprocity between geographical boundaries and demographic aggregates. Hence, it could be relevant to explore the sensitivity of the relationships with regard to the choice of geographical area.

List of Figures

List of Tables

UNIVERSITÉ DE RENNES 1

Titre : Économétrie des données imparfaites : méthodes et applications

Mots clés : économétrie appliquée ; données compositionnelles ; données agrégées ; variables dépendantes limitées ; inégalités ; dommages corporels.

Résumé : L'Économétrie permet de formaliser et de styliser un phénomène économique. Au fil du temps, les économètres font face à de nouveaux problèmes liés à l'évolution des données. À titre d'exemple, la prédominance de données massives ou les nouvelles réformes sur la protection des données font évoluer la nature, la forme ou encore la disponibilité des données. Cette transformation soulève de nouveaux défis et problèmes pratiques dans la prise en compte de ces données par les économètres.
L'objectif de cette thèse est de proposer et d'étudier des méthodologies adaptées aux données imparfaites, plus particulièrement, les variables agrégées et les variables dépendantes limitées à travers deux phénomènes économiques : premièrement, les mesures d'inégalités, et deuxièmement, l'indemnisation des dommages corporels.
Dans chaque contexte, un ensemble de méthodes pertinentes est proposé afin d'appréhender les données disponibles et d'expliquer chaque phénomène. Dans le cadre des données agrégées, elle fournit une étude des variables compositionnelles liée à l'agrégation des données catégorielles (Chapitre I) et une méthode d'estimation des mesures d'inégalités à partir des données quantiles (Chapitre II). Dans le cadre des données censurées, elle propose une méthode d'analyse des montants d'indemnisation pour les préjudices corporels (Chapitre III).

Title : Econometrics of imperfect data: methods and applications

Keywords : applied econometrics; compositional data; aggregated data; limited dependent variable; inequalities; bodily injury claim.

Abstract : The Econometrics approach enables to formalize and stylize an economic phenomenon. Over time, econometricians face new problems related to changes in data. For example, the prevalence of massive data or new data protection reforms are changing the nature, the form or the availability of data. As a result, econometricians face new challenges and practical problems in dealing with data.
The aim of this thesis is to propose and study methodologies adapted to imperfect data, more specifically, aggregated variables and limited dependent variables through two economic phenomena: first, inequality measures, and second, bodily injury claims.
In each context, a set of relevant methods is proposed to deal with the available data and to explain each phenomenon. In the context of aggregated data, a study of compositional variables related to the aggregation of categorical data (Chapter I) and a method for estimating inequality measures from quantile data (Chapter II) are provided. In the context of censored data, a method for analyzing bodily injury claims amounts is proposed (Chapter III).