HAL
open science

Données Agrégées et Variables Compositionnelles : Note Méthodologique

Enora Belz, Arthur Charpentier

To cite this version:

Enora Belz, Arthur Charpentier. Données Agrégées et Variables Compositionnelles : Note Méthodologique. 2019. hal-02097031

HAL Id: hal-02097031

https://hal.science/hal-02097031

Preprint submitted on 12 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Données Agrégées et Variables Compositionnelles Note Méthodologique

Enora Belz[1,*] **and Arthur Charpentier**[1,2]

[1]Univ Rennes, CNRS, CREM - UMR 6211, F-35000 Rennes, France
[2]Université du Québec à Montréal, Canada
[*]enora.belz@univ-rennes1.fr

ABSTRACT

La réforme du droit sur les données personnelles en Europe rend difficile l'accès aux données individuelles (même souvent non-nominatives), surtout quand on cherche des données jugées sensibles (et souvent, le revenu entre dans cette catégorie). Une solution souvent envisagée est la mise à disposition de données agrégées spatialement. Ces données posent toutefois deux problèmes techniques. Le premier est que les données catégorielles deviennent des compositions. Le second est lié au paradoxe écologique qui dit qu'il est dangereux d'inférer des relations économétriques individuelles à partir de données agregées. Nous verrons ici comment travailler avec des données compositionnelles (pour éventuellement juste valider une approche classique de régression linéaire - plus simple à interpréter). Et nous évoquerons le second, mais qui reste malheureusement trop général pour pouvoir être traité de manière satisfaisante.

1 Introduction et Motivation

Quand on veut faire de la micro-économétrie, on peut utiliser deux types de données : les données à caractère personnel et 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 car 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. Au lieu de mettre à disposition des données individuelles pour faire des études (revenus fiscaux des contribuables, résultats scolaires d'élèves scolarisées, etc.), il est devenu fréquent de publier des donnéees agrégées (revenus par zone IRIS, résultats par établissement scolaire, etc.). Comme nous allons le voir, 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).

Nous allons présenter dans la section 2 les données agrégées, et plus particulièrement spatialement. 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 européen 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, x_i, z_i), avec $z \in \mathscr{Z} = \{\zeta_1, \zeta_2, \cdots, \zeta_m\}$ un ensemble fini (correspondant aux différents groupes au sein desquels on va agréger les données) et un modèle de régression

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

Une autre écriture usuelle est

$$y_i = \beta_0 + x_i^\top \beta + \varepsilon_i, \tag{2}$$

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 (1984a) pour une revue exhaustive ou Clark & Avery (1976) une analyse des quartiers de Los Angeles), un bureau de vote (Klima *et al.* (2017) 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 $(\overline{y}_j, \overline{x}_j, z_j)$ et on considère la régression

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

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

Sur la Figure 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 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 la littératie 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 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 micro et macro, 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 agrége 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 menage, des 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 $\overline{x}_j = (\overline{x}_{1,j}, \overline{x}_{2,j}, \cdots, \overline{x}_{d,j})$ quand la variable x prend k. On peut ainsi avoir $d = 2$, avec deux modalités, par exemple le genre (masculin et 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.

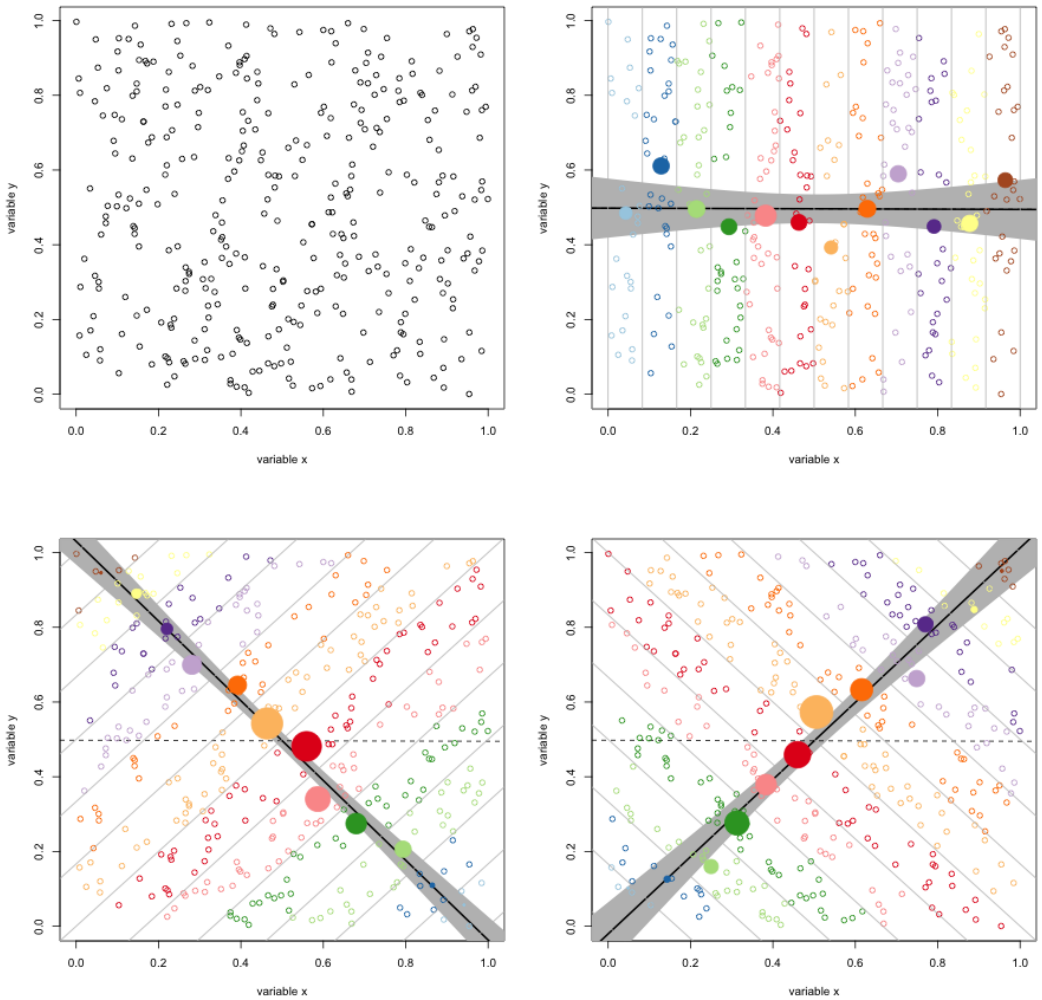

Figure 1. Paradoxe écologique : corrélation entre $\overline{y}_j$ et $\overline{x}_j$ en fonction du regroupement par classe, sur un même jeu de données (en haut à gauche). Les trois autres graphiques présentent (en ronds pleins) différentes agrégations. En haut à droite, les classes sont verticales (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 $\overline{x}_j$ (comme sur les données initiales, x_i et y_i étant ici indépendantes), alors qu'en bas, $\overline{y}_j$ et $\overline{x}_j$ sont respectivement corrélés négativement et positivement.

On pourrait aussi parler de données floues[1]. 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[2].

[1] 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 $(5\%, 85\%, 10\%)$).

[2] 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 $\overline{x}_{\text{vert},j}$ ou $\overline{x}_{\text{bleu},j}$ seront nuls).

Definition 1. *Une variable compositionnelle $\bar{x}$ est un vecteur du sous-simplexe*[3] $\tilde{\mathscr{S}}_d \subset \mathbb{R}^d$,

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

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

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

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

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

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

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

Formellement, $n = (n_1, \cdots, n_d)$ est vu comme la réalisation d'une loi multinomiale $\mathscr{M}(n, p)$ où $p \in \mathscr{S}_d$ est le vecteur de probabilité inconnu. $\bar{x} = \mathscr{C}(n)$ correspond aux fréquences empiriques dans chacunes des classes, et correspond à l'estimateur du maximum de vraisemblance du paramètre p. La contrainte forte $\bar{x}_1 + \cdots + \bar{x}_d = 1$ impose une corrélation négative entre les composantes : si $N \sim \mathscr{M}(n, p)$, et $\overline{X} = \mathscr{C}(N)$, 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 Utiliser des 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 note Loonis & 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 agrégée 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 & Harel (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 spatialement les individus. Openshaw (1984b) 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 (2015).

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[4]. 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 carré, 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).

[3] Le simplexe traditionnel est une portion d'hyperplan de $\mathbb{R}^d$, correspondant à l'enveloppe convexe de d points extremaux, 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.

[4] https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_grids

2.2 Le carroyage INSEE

L'INSEE est allé plus loin en proposant des carrés de 200m de côté[5]. 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'explique Loonis & 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 granualarité 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 mentionnées dans Loonis & Bellefon (2018) évoque trois 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 Deichman *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 de département (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 dernière a poussé l'INSEE à introduire l'IRIS[6], 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 2, on peut voir la distribution de la population par IRIS à gauche, et sur une carte à droite.

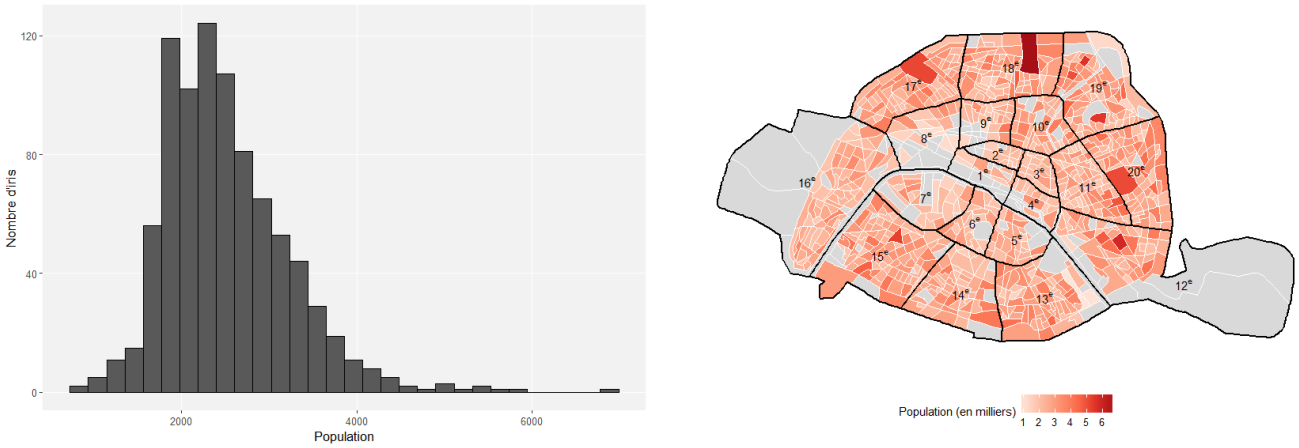

Figure 2. Distribution de la population par IRIS dans Paris.

2.4 Utiliser des Données Agrégées

Même si Gehlke & 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 & Taylor (1979), qui proposait d'évaluer systématiquement la variabilité des corrélations lorsque différents regroupement spatiaux étaient considérés. En particulier, Openshaw & Taylor (1979) a 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 & Rao (1985) a établi que la corrélation entre le chômage et l'absence de voiture pouvait

[5] https://www.insee.fr/fr/statistiques?debut=0&categorie=1&geo=ICQ-1
[6] https://www.insee.fr/fr/metadonnees/definition/c1523

varier entre -1,00 et +1,00, à partir de plus de 2500 ED (*enumeration districts*) dans le comté de Merseyside. Le rapport ESPON (2006) présente des agrégats fréquemment utilisées en Irlande, en Suède et en Allemagne.

Fotheringham & Wong (1991) revient 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 (2013), 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 comportement 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 & Mackenzie (2018) revient 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, on va étudier 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, x_i), où y est la variable continue que l'on souhaite décrire (par exemple le revenu) et x est une composition associée à une variable qualitative (par exemple le niveau d'étude). On peut écrire la régression linéaire

$$y_i = \beta_0 + x^\top \beta + \varepsilon_i$$

tout en gardant en mémoire que pour avoir l'identifiabilité, une modalité devra servir de référence. Mais comme le notait Chayes (1960), la contrainte $x^\top \mathbb{1} = x_1 + \cdots + x_d = 1$ (et donc la corrélation négative entre les composantes[7]) 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'Equation (1), $\mathbb{E}[Y|X=x] = \beta_0 + x^\top \beta$ le second terme correspond au produit scalaire $\langle x, \beta \rangle$ 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. Les transformations les plus populaires sont basées sur les log-ratio, dès lors qu'aucune composante n'est nulle. Soit $x \in \mathbb{R}^d_+$, la transformation additive ALR_j avec pour référence la $j$ème composante est la transformation $\widetilde{\mathscr{S}_d} \to \mathbb{R}^{d-1}$

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

On peut définir la transformation inverse en notant que $\mathrm{ALR}_j^{-1}(y) = \mathscr{C}(\exp[y_0^j])$, où $y_0^j = (y_1, \cdots, y_{j-1}, 0, y_j, \cdots, y_d)$.

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

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

[7] 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.

et la transformation inverse est tout simplement $\mathrm{CLR}^{-1}(y) = \mathscr{C}(\exp[y])$. Cette symétrie rend l'interprétation relativement simple, comme le montre Filzmoser & Hron (2009). Enfin pour la transformation dite isométrique ILR, on se donne une base orthonormée[8] $e = \{e_1, \cdots, e_{d-1}\}$ de $\mathscr{S}_d$, et on note ILR_e la transformation $\widetilde{\mathscr{S}_d} \to \mathbb{R}^{d-1}$

$$\mathrm{ILR}_e(x) = (\langle x, e_1\rangle, \cdots, \langle x, e_{d-1}\rangle).$$

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

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

qui redevient une régression linéaire usuelle[9] et l'estimateur naturel de β est alors $\widehat{\beta} = \mathrm{ILR}_e^{-1}\left(\widehat{b}^{\mathrm{OLS}}\right)$

$$\widehat{b}^{\mathrm{OLS}} = (\mathrm{ILR}_e(X)^\top \mathrm{ILR}_e(X))^{-1} \mathrm{ILR}_e(X)^\top y.$$

En dimension $d = 3$, la tranformation du log-ratio additif est

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

alors que la transformation inverse est

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

qui rappelle la transformation logistique pour les variables multinomiales, introduite par Theil (1969). La tranformation du log-ratio centré est

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

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

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

3.3 Visualisation d'une régression

Si la variable explicative 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 (2013). Ce graphique permet de visualiser des obervations x_i, comme sur la Figure 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 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), $x_i = (x_{\mathsf{BEP},i}, x_{\mathsf{BAC},i}, x_{\mathsf{SUP},i})$. Le premier modèle considéré est la régression classique

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

On peut aussi considérer une régression sur une transformation de x,

$$y_i = b_0 + b_1 \mathrm{ILR}_e(x_i)_1 + b_2 \mathrm{ILR}_e(x_i)_2 + \varepsilon_i \tag{ILR-1}$$

L'interprétation de la partie gauche du Tableau 1 - pour la régression linéaire (classique) - est assez naturelle ici.

Pour la seconde régression, avec la transformation des variables, dont l'estimation est à droite dans le Tableau 1, on peut transformer le paramètre estimé,

$$\widehat{\beta} = \mathrm{ILR}_e^{-1}(\widehat{b}) = (0.0565; 0.4892; 0.4542)$$

[8] Pour le produit scalaire d'Aitchison, comme définies dans Egozcue & Pawlowsky-Glahn (2015). Notons que si $x, y \in \tilde{\mathscr{S}}_d$, on peut définir le produit scalaire d'Aitchison en posant

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

.

[9] Techniquement, on peut doter le simplexe d'opérateurs de sommes et de produits, en posant $x \oplus y = \mathscr{C}(xy)$ et $b \otimes x = C(x^b)$ pour $b \in \mathbb{R}$. Dans ce cas, le modèle de régression multivarié peut formelle 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.

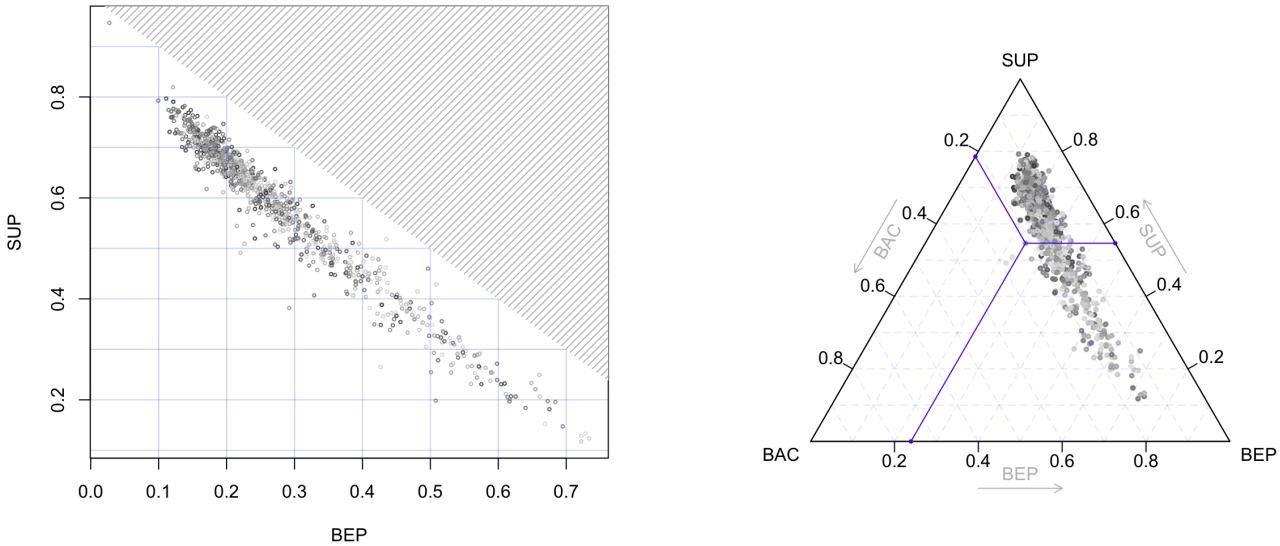

Figure 3. Nuage des points (x_2, x_3) dans $\mathbb{R}^2$ à gauche, et nuage des points (x_1, x_2, x_3) dans $\widetilde{S}_3$ à droite. à gauche, la partie supérieure droite est inatteignable, puisqu'alors $x_2 + x_3 \geq 1$. à droite est également représenté le point $x = (21\%, 24\%, 55\%)$ dans l'espace $(\mathsf{BAC}, \mathsf{BEP}, \mathsf{SUP})$.

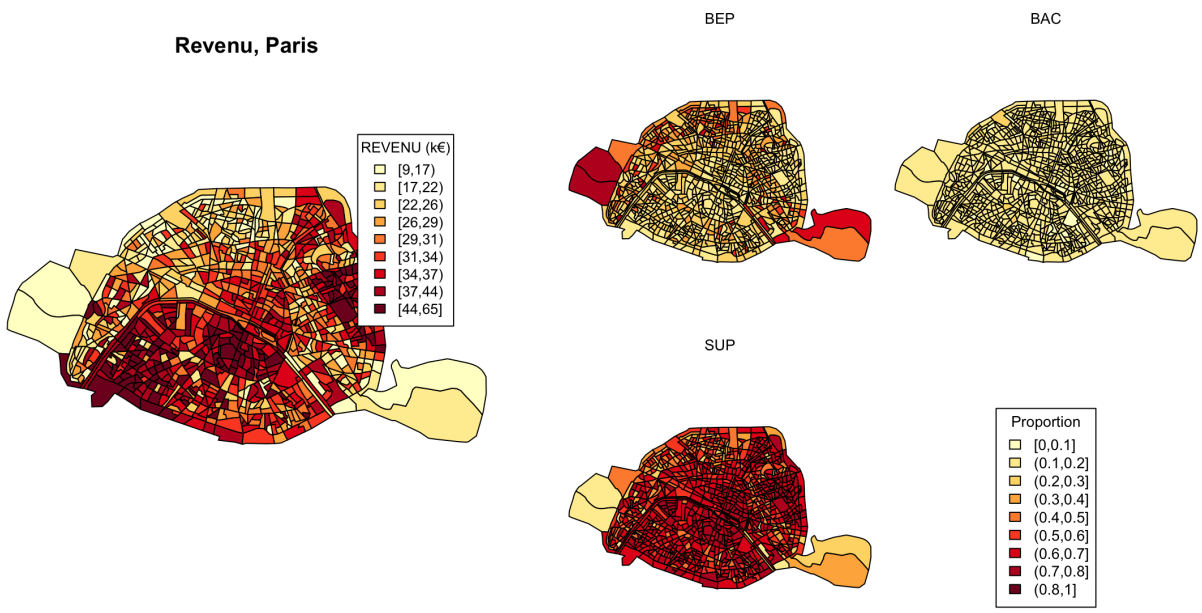

Figure 4. Distribution du revenu médian à gauche, par quartier, à Paris, et distribution du niveau d'étude à droite, avec la proportion de BEP, BAC et SUP.

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

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

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

	(OLS-1) revenu		(ILR-1) revenu
x_{BEP}	2.777***	$\text{ILR}_e(x)_1$	1.028***
	(0.671)		(0.122)
x_{SUP}	7.559***	$\text{ILR}_e(x)_2$	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
F Statistic	873.7***		1009***

le maximum étant obtenu quand $u = \beta$.

Mais probablement plus intéressant, on peut visualiser les prévisions, soit dans $\mathbb{R}^2$, soit dans le simplexe $\mathscr{S}_d$, comme sur la Figure 5.

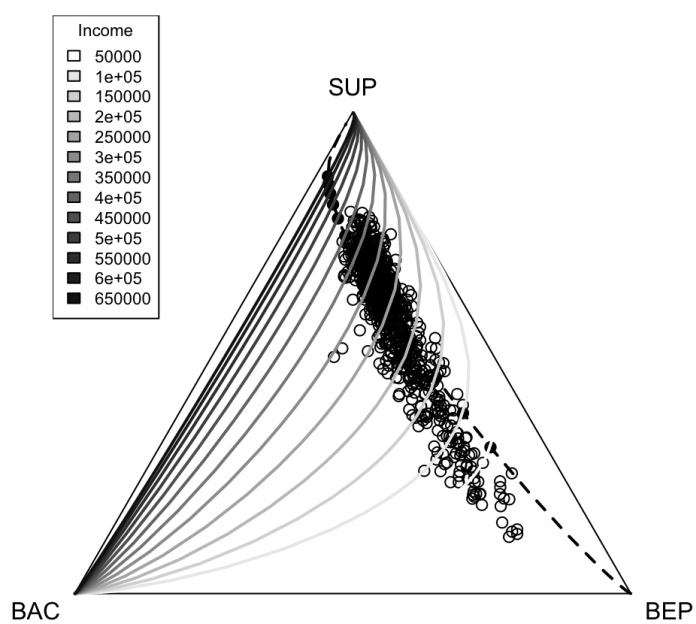

Figure 5. Courbes d'iso-niveau de y, avec le modèle (ILR-1) dans le sous-simplexe $\widetilde{S}_3$.

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 $(x_{\mathsf{BEP}}, x_{\mathsf{SUP}})$, comme sur la figure 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 iso-niveau de la fonction

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

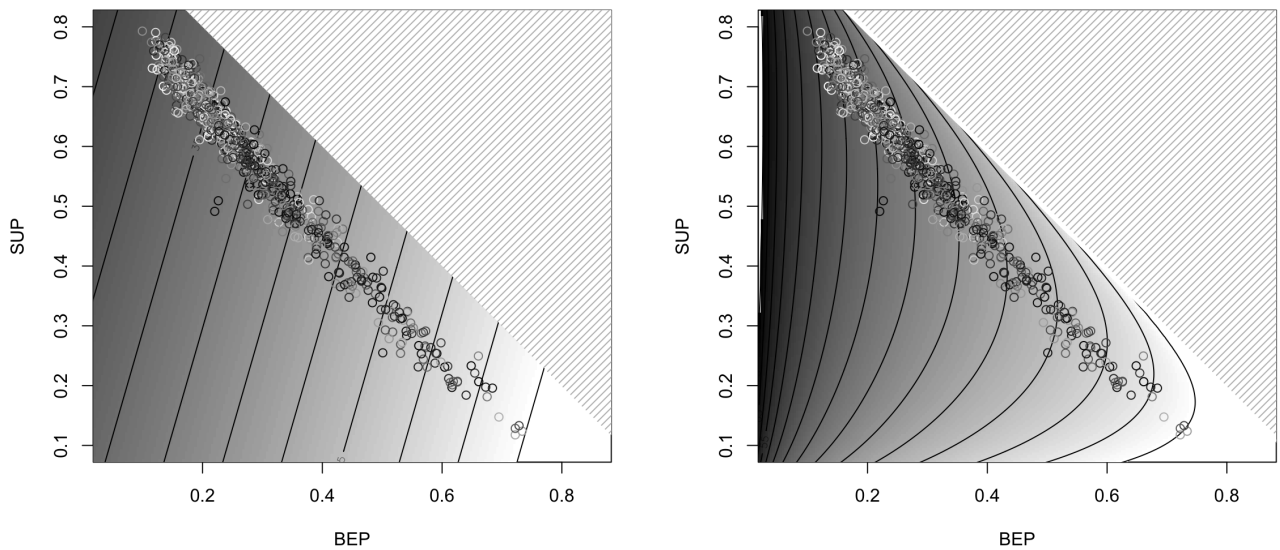

Figure 6. Courbes d'iso-niveau de y, avec le modèle (OLS-1) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-1) à droite.

A droite de la Figure 7, on peut voir l'évolution de la prévision en fonction de $x = (u, 15\%, 85\% - u)$, 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 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.

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 7, le revenu ne semble pas augmenter linéairement avec la proportion de diplômés de l'enseignement supérieur, ou décroitre linéairement avec la proportions de titulaire d'un brevet des collèges. On peut alors naturellement tenter des transformations de type Box-Cox (décrites dans Box & 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 9. 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 moyen.

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

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

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

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

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

$$\log y_i = b_0 + \mathrm{ILR}_e(x_i)^\top b + \varepsilon_i \tag{ILR-2}$$

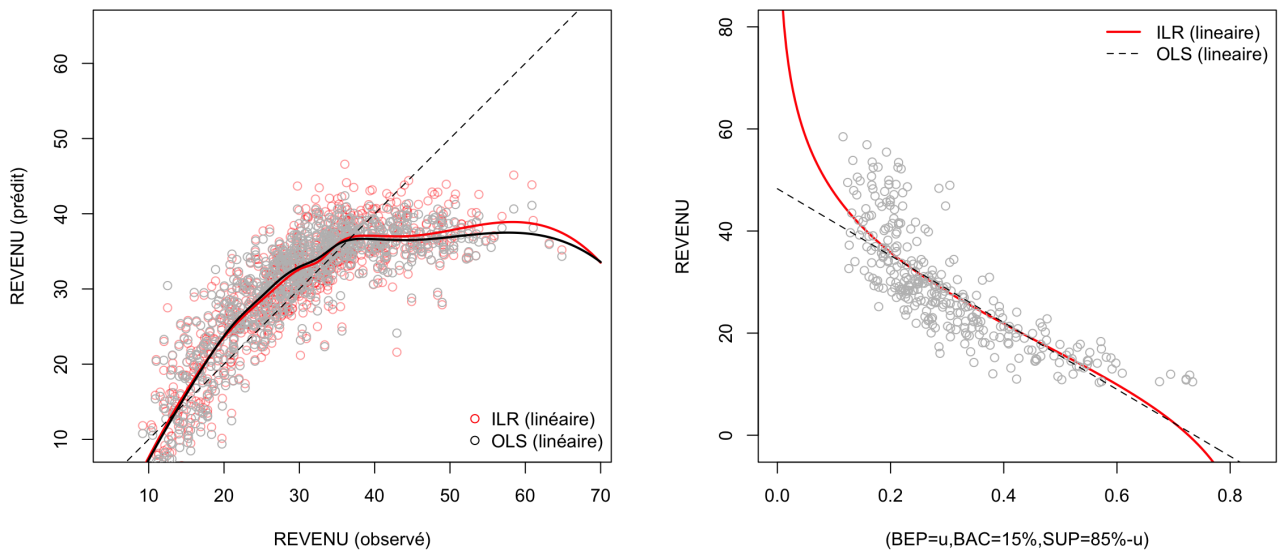

Figure 7. à gauche, comparaison des prévisions de niveau moyen de revenu, par quartier, avec les modèles (OLS-1) et (ILR-1). à droite, prévision en fonction de $x = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%.

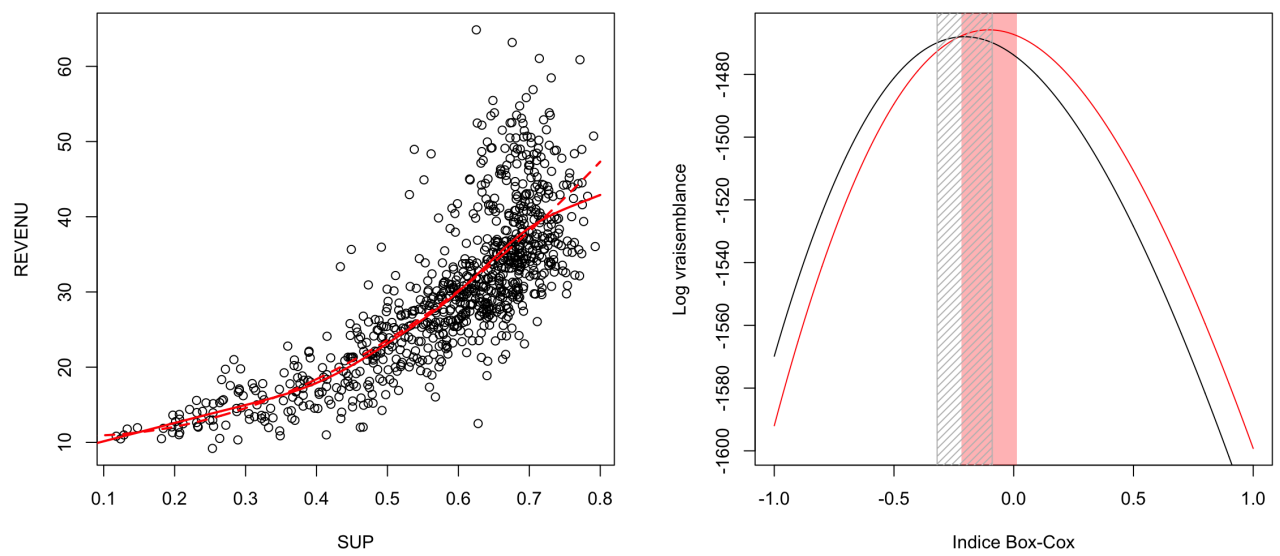

Figure 8. Évolution du revenu moyen en fonction de la proportion de diplômés de l'enseignement supérieur, par quartier, à gauche, et log-vraisemblance profilée du modèle de Box-Cox à droite.

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

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

On peut également transformer (ILR-1), pour introduire une composante quadratique, de la forme $y_i = \beta_0 + \langle x_i, \beta \rangle + \langle x_i, Bx_i \rangle +$

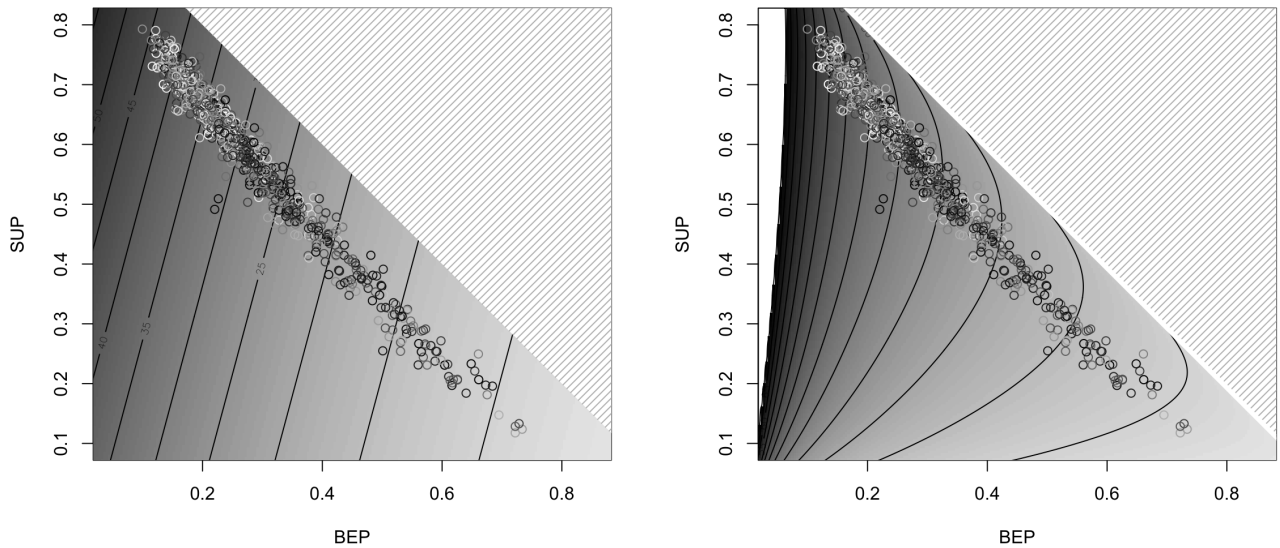

Figure 9. Courbes d'iso-niveau de y, avec le modèle (OLS-2) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-2) à droite.

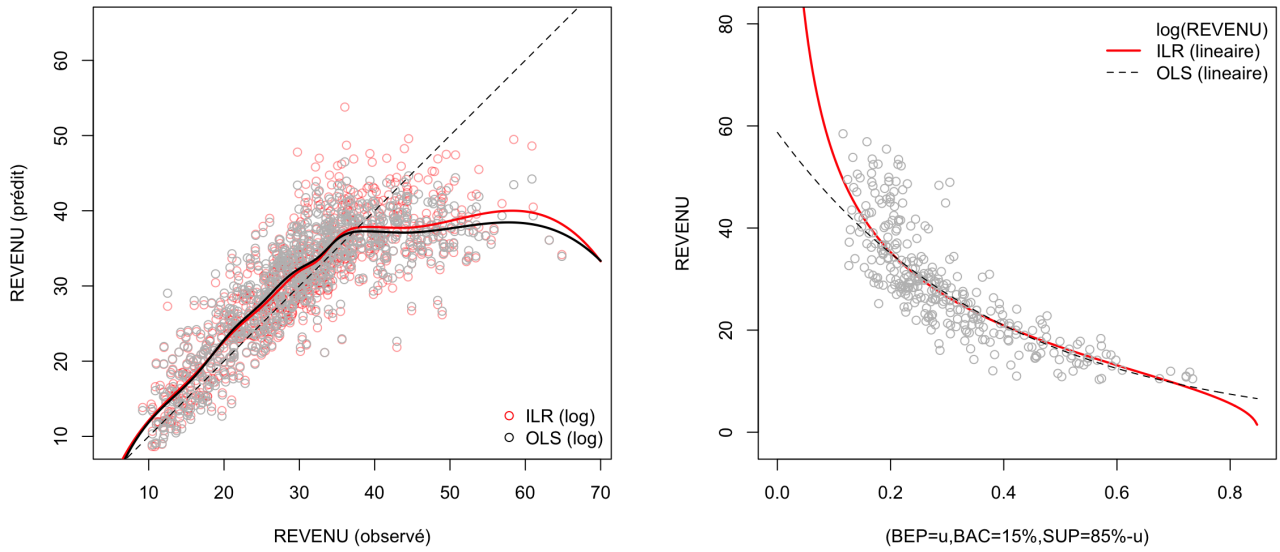

Figure 10. à droite, prévision en fonction de $x = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%, et à gauche, comparaison des prévisions avec les deux modèles (OLS-2) et (ILR-2) sur le logarithme du revenu.

ε_i, avec une matrice carrée symétrique B, soit

$$y_i = b_0 + \mathrm{ILR}_e(x_i)^\top b + \mathrm{ILR}_e(x_i)^\top B\, \mathrm{ILR}_e(x) + \varepsilon_i \qquad \text{(ILR-3)}$$

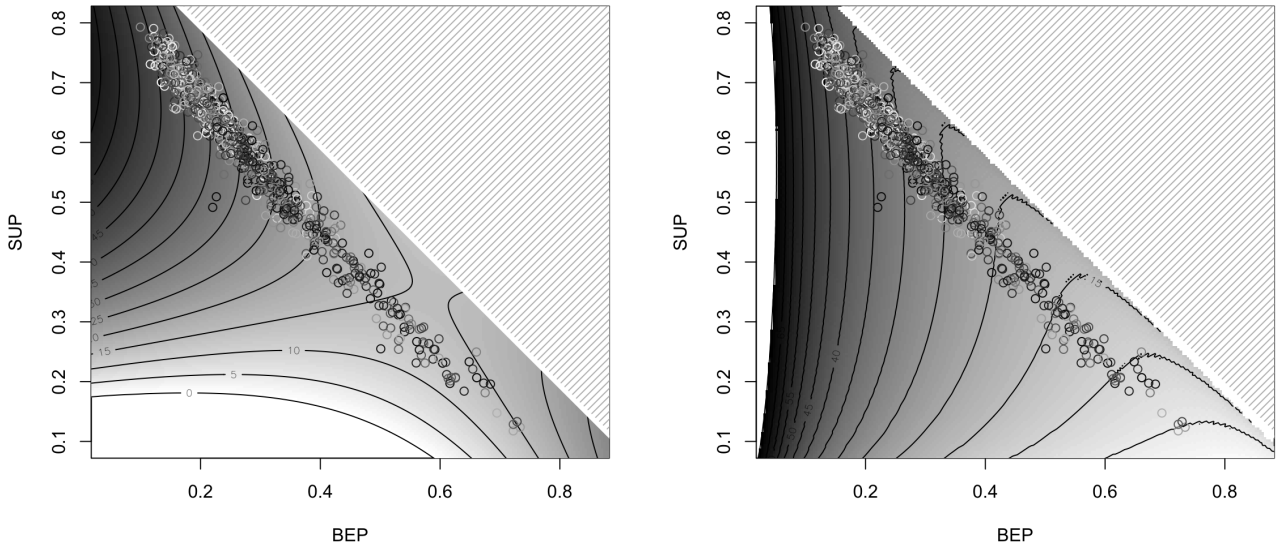

Figure 11. Courbes d'iso-niveau de y, avec le modèle (OLS-3) dans le plan (BEP,SUP) de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR-3) à droite.

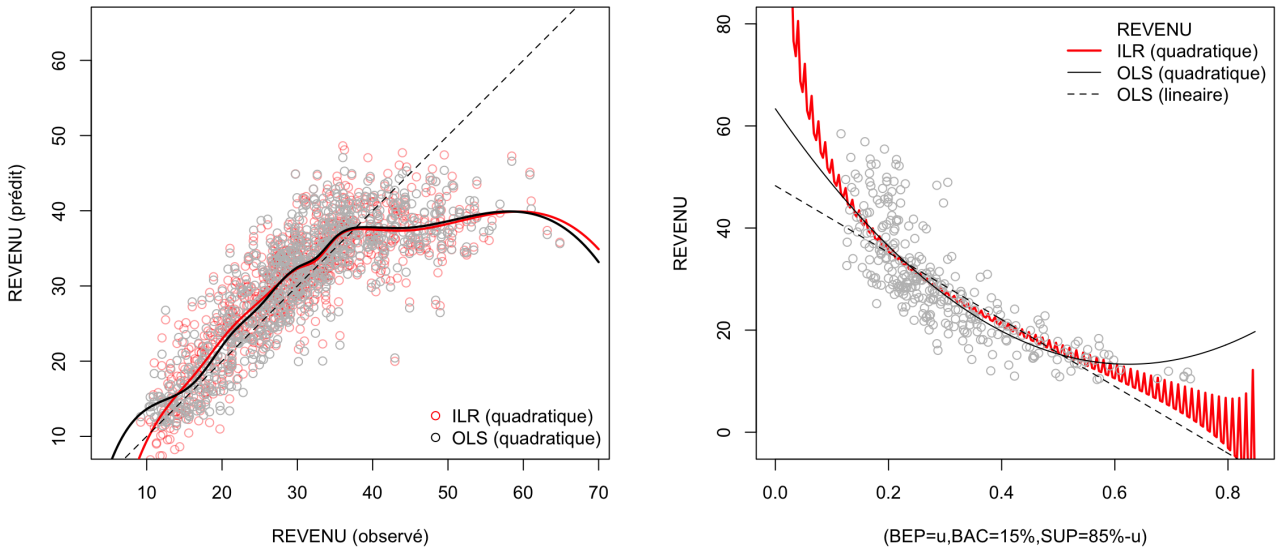

Figure 12. à droite, prévision en fonction de $x = (u, 15\%, 85\% - u)$, pour u variant de 0% à 85%, et à gauche comparaison des prévisions avec les deux modèles (OLS-3) et (ILR-3) avec une transformation quadratique de x.

3.6 Propriétés des estimateurs

La matrice $\text{ILR}(X)$ 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 Rousseeuw & Kaufman (1990) ou Bondell & Reich (2009). En fait, si les groupes sont constitués aléatoirement, indépendemment de y et de x, alors le rang[10] de $\text{ILR}(X)$ correspond au nombre de modalités "*réel*". Autrement dit, si x est une variable à 8 modalités mais que le rang de

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

ILR(X) est 5, c'est que 3 modalités peuvent être fusionnées entre elles, pour constituer au final 5 "*vraies*" catégorie (une étant la référence).

Table 2. Régression quadratique: revenu en fonction du niveau d'étude.

	(OLS-3) revenu		(ILR-3) revenu
x_{BEP}	19.441***	$\mathrm{ILR}_e(x)_1$	0.162
	(3.742)		(0.163)
x_{SUP}	10.800***	$\mathrm{ILR}_e(x)_2$	1.005***
	(2.724)		(0.069)
x^2_{BEP}	-7.621	$\mathrm{ILR}_e(x)_1^2$	0.241***
	(5.231)		(0.058)
x^2_{SUP}	6.182**	$\mathrm{ILR}_e(x)_2^2$	0.292***
	(1.908)		(0.051)
x^2_{BAC}	15.890 ***	$\mathrm{ILR}_e(x)_1(x)_2$	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***

3.7 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.

La Figure 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 15.

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

$$y_i = b_0 + \mathrm{ILR}_e(x_{1,i})^\top b_1 + \cdots + \mathrm{ILR}_e(x_{k_i})^\top b_k + \varepsilon_i$$

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".

Table 3. Régression linéaire: revenu en fonction de la superficie du logement.

	(OLS) revenu		(OLR) revenu
$x_{40\text{-}}$	11.056***	$\text{ILR}_e(x)_1$	4.790***
	(1.665)		(0.477)
x_{100+}	82.309***	$\text{ILR}_e(x)_2$	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***

Table 4. Régression sur deux variables explicatives : revenu en fonction du niveau d'étude et de la taille du logement.

	(OLS) revenu		(ILR) revenu
$x_{1,\text{BEP}}$	-9.978·	$\text{ILR}_e(x_1)_1$	7.013***
	(5.754)		(0.967)
$x_{1,\text{SUP}}$	34.956***	$\text{ILR}_e(x_1)_2$	9.240***
	(5.146)		(0.580)
$x_{2,40\text{-}}$	-6.001***	$\text{ILR}_e(x_2)_1$	0.203
	(1.096)		(0.411)
$x_{2,100+}$	51.115**	$\text{ILR}_e(x_2)_2$	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***

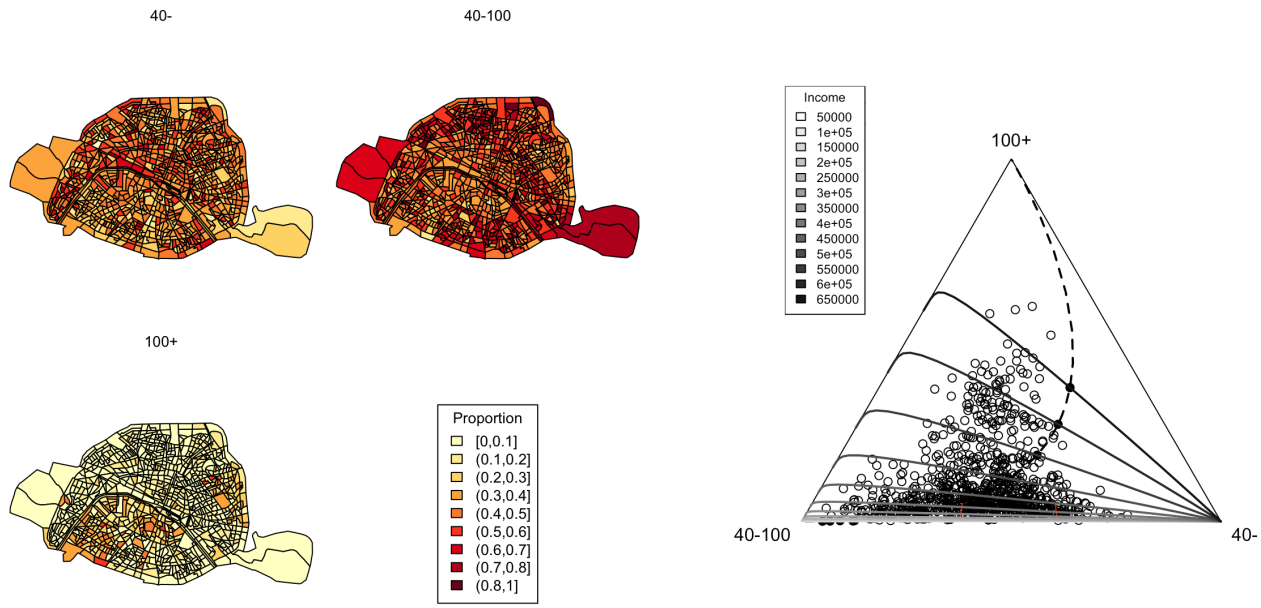

Figure 13. Variable x_2 correspondant à la proportion de personnes dans le quartier habitant un logement (résidence principale) de moins de 40m^2, entre 40m^2 et 100m^2, et plus de 100m^2. La Figure de droite est la régression linéaire ILR (Table 3).

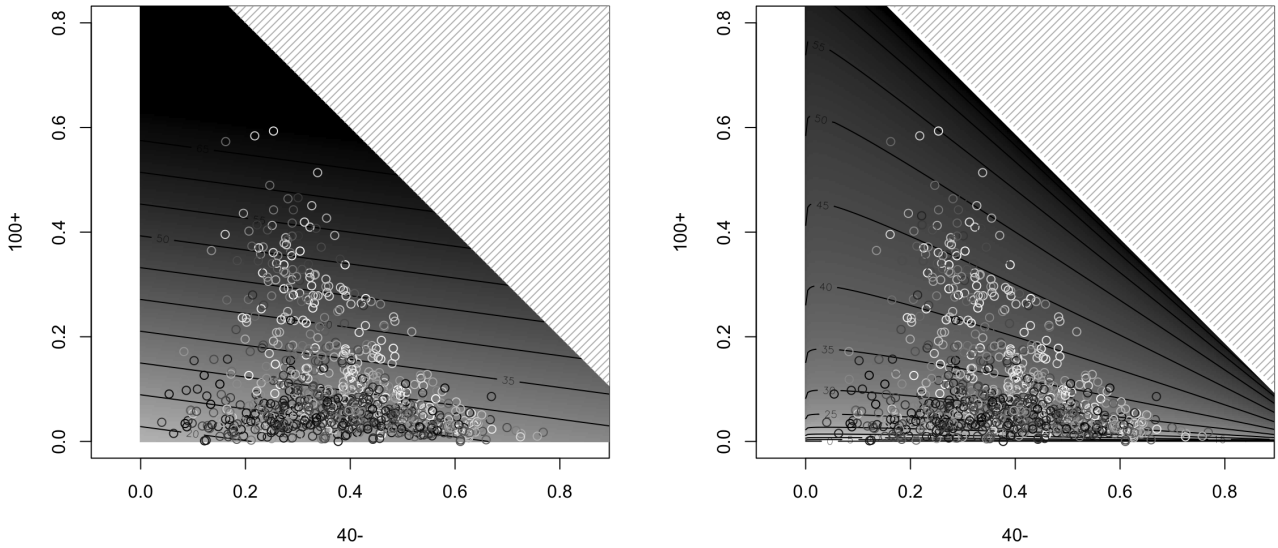

Figure 14. Courbes d'iso-niveau du revenu y, avec le modèle (OLS) dans le plan (moins de 40m^2, plus de 100m^2), de $\mathbb{R}^2$ à gauche, et avec le modèle (ILR) à droite.

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

$$p_{i:j} = g(x_{i:j}, \alpha)$$

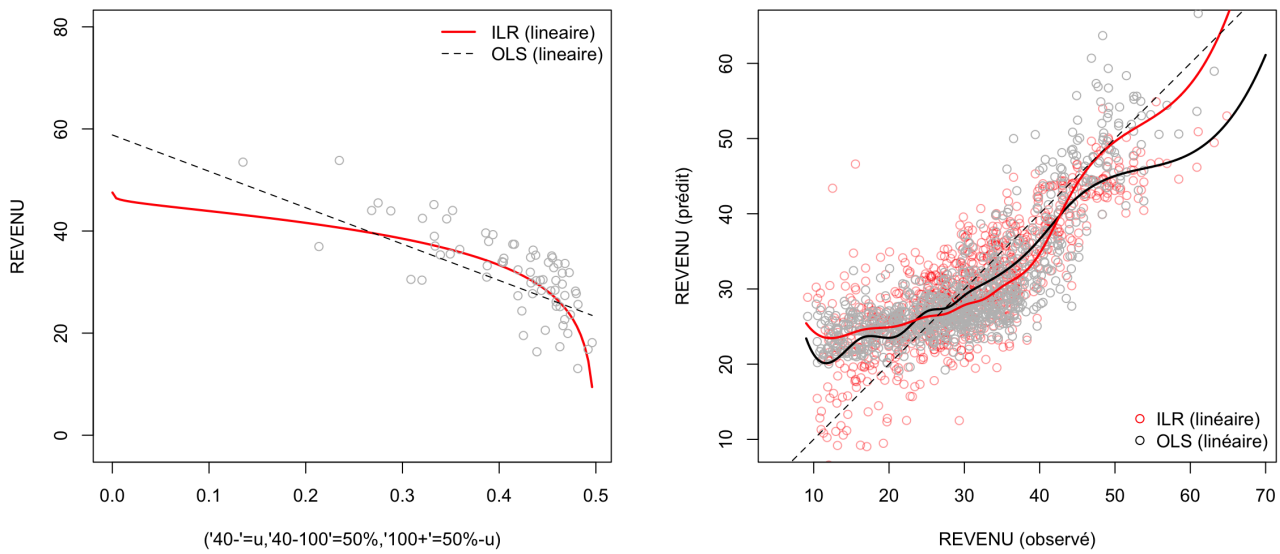

Figure 15. À gauche, prévision du revenu en fonction de $x = (u, 50\%, 50\% - u)$, pour u variant de 0% à 50%, et à droite comparaison des prévisions avec les deux modèles (OLS) et (ILR).

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 suit une binomiale de paramètres n_j et $\overline{p}_j$. La probabilité d'être $\overline{p}_j$ 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

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

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 & 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, l'absence d'autocorrélation spatiale et l'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 Annexes : Aspects computationnels

Plusieurs librairies sous R proposent des fonctions pour faire de l'analyse de données compositionelles, y compris des régressions, comme en atteste l'ouvrage de Van den Boogaart & 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)
```

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)
```

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

References

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

Berlin, J.A., Santanna, J., Schmid, C.H., Szczech, L.A. & Feldman, H.I. (2002). Individual patient versus group level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medecine*, 22, 371–387 doi:10.1002/sim.1023

Best, N., Cockings, S., Bennett, J., Wakefield, J. & Elliott, P. (2001). 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*. **164**, 155–174.

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

Box, G. E. P., & D. R. Cox. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26, 2, 211–252.

Buccianti, A. & Pawlowsky-Glahn, V. (2011). Compositional Data Analysis: Theory and Applications. John Wiley & Sons.

Chayes, F. (1960). On Correlation between Variables of Constant Sum. *Journal of Geophysical Research*, 65:12, 4185–4193

Clark, W.A.V. & Avery, K.L. (1976). The Effects of Data Aggregation in Statistical Analysis. *Geographical Analysis*, 8, 428-438, doi:10.1111/j.1538-4632.1976.tb00549.x

Crawford, C.A.G. & Young, L.J. (2004). A spatial view of the ecological inference problem. *Chapter 10 in* Ecological Inference: New Methodoligcal Strategies (King, Rosen & Tanner Eds.). Cambridge University Press.

Deichman, U., Balk, D. & Yetman, G. (2001). Transforming population data for interdisciplinary usages : from census to grid . Washington (DC) : Center for International Earth Science Information Network. http://sedac.ciesin.org/downloads/docs/gpw-v3/gpwdocumentation.pdf

Duncan, O.D. & Davis, B. (1953). An Alternative to Ecological Correlation. *American Sociological Review*, 18:6, 665–666.

Egozcue, J.J. & Pawlowsky-Glahn, V. (2015) Simplicial geometry for compositional data. *in* Compositional Data Analysis in the Geosciences: From Theory to Practice. Buccianti, Mateu-Figueras & Pawlowsky-Glah Eds, Geological Society, London, Special Publications, 264, 145-159

El Emam K., Brown, A. & AbdelMalik, P. (2008). Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk. *Journal of the American Medical Informatics Association*, **16**:2, 256–266.

Elff, M. (2009) Social divisions, party positions, and electoral behaviour. *Electoral Studies*, 28:2, 297–308. doi:10.1016/j.electstud.2009.02.002

ESPON (2006) The modifiable areas unit problem. *Scientific Support Project 3.4.3*.

Filzmoser, P. & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40, 233-248.

Fotheringham, A.S. & Wong, D.W.S. (1991). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A*, **23**, 1025–1044.

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

Fry, T. (2011). Applications in Economics. *in* Compositional Data Analysis. Pawlowsky-Glahn & Buccianti Eds., Wiley Interscience, 318-328.

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

Gelman, A. (2009). Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do. Princeton University Press.

Greenland, S. (2001). Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30, 1343–1350 doi:10.1093/ije/30.6.1343

Goodman, L.A. (1953) Ecological regressions and the behavior of individuals. *American Sociological Review*, 18, 663-664.

Hannan, M.T. & Burstein, L. (1974). Estimation from Grouped Observations. *American Sociological Review*, 39, 374-392.

Holt, D., Steel, D.G., Tranmer, M. & Wrigley, N. (2001). Aggregation and Ecological Effects in Geographically Based Data. *Geographical Analysis* 28, 244-261, doi:10.1111/j.1538-4632.1996.tb00933.x

King, G. (1997). A Solution to the Ecological Inference Problem. Princeton University Press.

Klima, A., Schlesinger, T., Thurner, P.W. & Küchenhoff, H. (2017) Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions. *Sociological Methods & Research*, 244-261.

Lin, W., Shi, P., Feng, R., & Li, H. (2014). Variable selection in regression with compositional covariates. Biometrika, 101(4), 785–797. doi:10.1093/biomet/asu031

Loonis, V. & Bellefon, M.-P. (dir.) (2018). Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R, *Insee Méthodes*, 131, Insee, Eurostat.

Loth, A. (2015). Risques de ré-identification dans les bases de données de santé, moyens de s'en prémunir : un projet de loi conciliant ouverture et protection. *DREES Santé*, **64**, 7–18.

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

Morais, J. (2017). Impact of media investments on brands' market shares : a compositional data analysis approach. Thèse de Doctorat, Université de Toulouse 1.

Openshaw, S. & Taylor, P.J. (1979). A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem. *in* Statistical Applications in the Spatial Sciences, N. Wrigley ed., 127–144. London: Pion.

Openshaw, S. (1981). Le problème de l'agrégation spatiale en géographie. L'Espace géographique. **10**, 15–24.

Openshaw, S. (1984a). The Modifiable Areal Unit Problem. CATMOG No. 38, Geo Books, Norwich, U.K.

Openshaw, S. (1984b). Ecological fallacies and the analysis of areal census data. *Environment and Planning A: Economy and Space*,**16**, 17–31.

Openshaw, S. & Rao, L. (1985). Algorithms for Reengineering 1991 Census Geography. *Environment and Planning A: Economy and Space*,**17**, 425–44.

Pearl, J. & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books.

Pearson, K. (1897) 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

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-435.

Ronning, G. (1992). Share equations in econometrics: A story of repression, frustration and dead ends. *Statistical Papers*, 33, 307–334.

Rousseeuw, P. & Kaufman, L. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Schwartz, S. (1994) The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *American Journal of Public Health*

Shively, W.P. (1969) "Ecological" Inference: The Use of Aggregate Data to Study Individuals. *The American Political Science Review*, 63, 1183–1196.

Smith, R.J. (1985) A Selective Review of Confidentiality Research Published Since 1975. United Census, https://bit.ly/2MwU8Mq.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, **10**:5, 571–588.

Theil, H. (1954) Linear Aggregation of Economics Relations. Amsterdam: North-Holland.

Theil, H. (1969) A multinomial extension of the linear logit model. *International Economic Review*, 10, 251–259.

Van den Boogaart, K.G. & Tolosana-Delgado, R. (2013). Analyzing compositional data with R. Springer.

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

Wakefield, J.C. & Lyons, H. (2010). Spatial Aggregation and the Ecological Fallacy *in* Handbook of Spatial Statistics, Alan E. Gelfand, Peter Diggle, Peter Guttorp, Montserrat Fuentes Eds., Chapman & Hall CRC.

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

West, D. (2013). Ternary Equilibrium Diagrams. 2nd ed. New York: Springer Verlag.

Wong, D.W.S. (2013). The Modifiable Areal Unit Problem (MAUP) *in* WorldMinds: Geographical Perspectives on 100 Problems, D.G. Janelle, B. Warf & K. Hansen eds., 571-575, Springer Verlag.

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