Aurait-on pu construire des régions selon des critères statistiques ?

Marc CHRISTINE[1] (INSEE, DMCSI),
Michel ISNARD[2] (Insee, Secrétariat général)

Version préliminaire

MOTS-CLÉS : Agrégation, CAH, contiguïté, inertie

0 RÉSUMÉ

La (re)construction de régions a constitué, depuis plusieurs mois, un objet de débats intenses au sein de la classe politique et de controverses dans lesquelles les préoccupations des différents acteurs semblaient assez éloignées de considérations statistiques. On a vu fleurir au fil des jours différentes cartes : projets, souhaits, simulations… avec apparition de 12, 13, 14 .. macro-régions, sans parler de la difficulté sémantique de désignation des entités ainsi construites.

A l'issue de ces débats et au prix d'arbitrages qui n'ont pas nécessairement été acceptés unanimement, le Parlement a voté au final ce qui devient la **loi n° 2015-29 du 16 janvier 2015**, **qui établit un découpage de la France continentale en 12 régions**.

Tous les exercices de partitionnement géographique évoqués ci-dessus ont été réalisés dans le cadre extrêmement simple consistant à partir des régions existantes et à agréger certaines d'entre elles pour arriver à un total de 12 à 15 macro-régions (si l'on se limite aux régions continentales). Même si le nombre de combinaisons possibles demeure élevé, la contrainte de *contiguïté* des régions à agréger réduit sensiblement le champ des possibles, sans parler de la restriction imposée par les considérations politiques, historiques ou sociales qui « interdiraient » certains regroupements.

Reprenant une problématique initialisée aux JMS 2000[3], puis développée aux JMS 2012[4], le présent article se propose d'explorer des solutions statistiques de partitions du territoire en « régions », en procédant par des agrégations d'unités géographiques élémentaires plus fines que les actuelles régions : départements, arrondissements, zones d'emploi, cantons, voire communes.

Il s'agit de la mise en application des techniques, des algorithmes et des outils informatiques développés notamment lors de la dernière présentation citée, pour constituer, au sein d'une population de référence (les communes du territoire métropolitain - Corse mise à part) des sous-ensembles ou *classes* présentant des conditions d'homogénéité ou d'hétérogénéité maximales vis-à-vis de certaines caractéristiques quantitatives.

Mais les classes ainsi constituées doivent aussi respecter la contrainte forte de *connexité*, relativement au positionnement géographique des unités élémentaires. Celui-ci est appréhendé par le concept de *contiguïté*, qui constitue une relation binaire entre les unités géographiques élémentaires. Les tailles de ces classes (exprimées selon une unité ad hoc, qui peut être la population, la

[1] marc.christine@insee.fr
[2] michel.isnard@insee.fr
[3] « Un algorithme de regroupement d'unités statistiques selon certains critères de similitude », Marc CHRISTINE et Michel ISNARD, VII[èmes] Journées de Méthodologie statistique, 4-5 décembre 2000.
[4] « Agrégation optimale sous contrainte de contiguïté : aspects théoriques et mise en œuvre avec applications à des cas pratiques », Marc CHRISTINE et Michel ISNARD, XI[èmes] Journées de Méthodologie statistique, 24-26 janvier 2012.
.

superficie…) sont limitées pour éviter des distorsions trop importantes. Enfin, le nombre total de classes à constituer est fixé a priori (par exemple par le pouvoir politique).

On simulera donc différents partitionnements, selon les critères d'optimisation et selon les jeux de variables retenus : le cas standard est celui où l'on cherche à minimiser ou maximiser la variance intra-classes pour une variable quantitative standard : ceci illustre le cas de la distance euclidienne et d'une dispersion appréhendée par le concept d'*inertie*.

D'autres critères seront analysés : avec plusieurs variables d'intérêt, quantitatives ou qualitatives ou avec des statistiques non linéaires.

Au final, il sera intéressant de comparer les cartes obtenues au découpage des régions actuelles ou aux projets de regroupement votés récemment au Parlement.

Table des matières

1 Histoire de la région[5]

Sans revenir à la situation d'avant la Révolution française où existaient 39 circonscriptions territoriales, dénommées provinces, généralités, comtés, marches, duchés.., il faut rappeler qu'entre les deux guerres mondiales, différentes mesures avaient préfiguré les régions (régions économiques regroupant des chambres de commerce, régions touristiques portées par les fédérations de syndicats d'initiative…). C'est le régime de Vichy qui a, par le décret du 30 juin 1941, créé une division du territoire par un découpage regroupant des départements, proche des actuelles régions, en attribuant à certains préfets les pouvoirs de préfets régionaux.

La mesure fut abrogée à la Libération mais le Gouvernement provisoire, par l'ordonnance du 10 janvier 1944, instaura des régions administratives placées sous l'autorité d'un commissaire de la République. Celles-ci sont dissoutes en janvier 1946 mais une loi du 21 mars 1948 met en place des Inspecteurs généraux de l'administration en mission extraordinaire (IGAME) chargés de coordonner au sein de 13 circonscriptions (les « *igamies* ») l'action des régions de défense et des préfets de départements.

En 1954, les comités régionaux d'expansion, d'initiative privée, sont officiellement agréés. Puis un décret du 30 juin 1955 crée vingt et une **régions économiques de programme** articulées autour de villes charnières et un autre, du 2 juin 1960, les transforme en **circonscriptions d'action régionale**. Leurs limites sont les limites actuelles, avec fusion des « régions » Rhône et Alpes et transfert de certains départements pyrénéens d'une « région » à une autre.

Les décrets du 14 mars 1964 créent vingt et un préfets de région. Parallèlement, sont mises en place des commissions de développement économique régionales (CODER), instances consultatives composées des représentants des intérêts socioprofessionnels ou territoriaux, chargées d'émettre un avis sur toutes les questions relatives au développement économique et à l'aménagement du territoire dans la circonscription régionale.

En 1969, un projet de projet de réforme du Sénat et de création des régions est soumis à référendum par le général de Gaulle. Le titre I^{er} de ce projet devait constitutionnaliser l'existence des régions comme collectivités territoriales. Il devait s'agir des circonscriptions d'action régionale créées en 1960, plus la Corse. On connaît l'issue de ce référendum et les conséquences politiques qui en ont résulté.

De fait, l'histoire de la régionalisation s'est mise en sommeil jusqu'à la loi du 5 juillet 1972, transformant les circonscriptions d'action régionale en « ***établissements publics régionaux*** », leur conférant ainsi la personnalité juridique et l'autonomie budgétaire mais sans leur donner le statut de collectivités locales, en particulier sans leur donner une représentation démocratique directe.

La véritable création des régions intervient dans les lois de décentralisation du 1er Gouvernement Mauroy et, plus précisément, lors de la promulgation de la loi du 2 mars 1982, dite « **loi relative aux droits et libertés des communes, des départements et des régions** », qui crée la région en tant que nouvelle collectivité locale. Celle-ci sera dotée d'un exécutif élu au suffrage universel direct, le conseil régional, dont les premières élections eurent lieu pour la 1ère fois en 1986, puis tous les 6 ans jusqu'en 2010.

La décentralisation a été cristallisée par la loi constitutionnelle du 28 mars 2003 (promulguée le 13 août 2004) relative à l'organisation décentralisée de la République qui consacre le principe de décentralisation et reconnaît aux régions un statut de collectivité territoriale de plein droit, à l'instar des communes et des départements.

[5] Sources : Association des Régions de France et *alii*

Dispositions de la LOI n°2015-29 du 16 janvier 2015 - art. 1

Sans préjudice des dispositions applicables aux régions d'outre-mer et à la collectivité territoriale de Corse, les régions sont constituées des régions suivantes, dans leurs limites territoriales en vigueur au 31 décembre 2015 :

- Alsace, Champagne-Ardenne et Lorraine ;

- Aquitaine, Limousin et Poitou-Charentes ;

- Auvergne et Rhône-Alpes ;

- Bourgogne et Franche-Comté ;

- Bretagne ;

- Centre (renommée Centre - Val de Loire) ;

- Île-de-France ;

- Languedoc-Roussillon et Midi-Pyrénées ;

- Nord - Pas-de-Calais et Picardie ;

- Basse-Normandie et Haute-Normandie ;

- Pays de la Loire ;

- Provence-Alpes-Côte d'Azur.

Principes généraux

La méthode mise en œuvre est celle de la ***Classification ascendante hiérarchique sous contrainte de contiguïté (CAH),*** dont les principes généraux et les conditions de programmation ont été exposés dans [JMS 2000] et [JMS 2012]. On rappellera ci-dessous les principaux éléments de méthode et les principaux résultats, tout en développant certains aspects qui n'avaient pas été évoqués dans les papiers cités.

2 Modèle de base : rappels

2.1 Centre de gravité, Inertie

Le cadre général est rappelé ci-après. Les démonstrations des résultats figurent dans [JMS 2012].

Considérons une population finie $\mathcal{P}$ de cardinal *N*, composée d'éléments appelés *unités statistiques*, indexées par un indice i. Sur chaque unité statistique, on suppose définis :

- Une variable d'intérêt x_i. Celle-ci est supposée *numérique* à ce stade.
- Un poids α_i (> 0).

Il est d'usage de définir :

- le *centre de gravité* (pondéré) de la population $\mathcal{P}$ relativement aux variables d'intérêt x_i, qui sera la valeur : $\boxed{g = \frac{\sum_{i \in P} \alpha_i x_i}{\sum_{i \in P} \alpha_i}}$.

  On a donc la relation : $(\sum_{i \in P} \alpha_i) g = \sum_{i \in P} \alpha_i x_i$, d'où : $\boxed{\sum_{i \in P} \alpha_i (x_i - g) = 0}$.

- La *variance ou inertie*[6] de la population $\mathcal{P}$ relativement à ces mêmes variables :

$$\boxed{I = \frac{\sum_{i \in P} \alpha_i (x_i - g)^2}{\sum_{i \in P} \alpha_i}}.$$

2.2 Effet d'une partition de la population de référence

Supposons que la population $\mathcal{P}$ soit partitionnée en K *sous-populations* ou *classes*, notées $\mathcal{P}_1$, $\mathcal{P}_2$, …, $\mathcal{P}_K$. Sur chaque classe $\mathcal{P}_k$, on définit :

- le centre de gravité $g_k = \frac{\sum_{i \in P_k} \alpha_i x_i}{\sum_{i \in P_k} \alpha_i}$ [vérifiant la relation : $\sum_{i \in P_k} \alpha_i (x_i - g_k) = 0$]
- et l'inertie $I_k = \frac{\sum_{i \in P_k} \alpha_i (x_i - g_k)^2}{\sum_{i \in P_k} \alpha_i}$.

[6] Le terme « Inertie » sera préféré par la suite car il permet de traiter des cas plus généraux que celui du calcul de la décomposition de la variance d'une variable numérique définie sur les unités de la population $\mathcal{P}$.

Dans ce cas, l'inertie I relative à la population $\mathcal{P}$ sera qualifiée d'*inertie totale* pour la distinguer des inerties I_k.

On dispose alors de la classique équation d'**analyse de la variance** (ou théorème de HUYGHENS-KOENIG) :

$$\boxed{I = \sum_{k=1}^{K}\left[\frac{\omega_k}{\omega}[I_k + (g_k - g)^2]\right]}, \textbf{ (1)}$$

en notant : $\omega = \sum_{i \in P} \alpha_i$ et $\omega_k = \sum_{i \in P_k} \alpha_i$.

- Le terme $\boxed{I^a = \sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k}$ constitue la variance (ou inertie) *intra-classe* : il s'agit d'une moyenne pondérée des inerties propres à chacune des classes $\mathcal{P}_k$, les poids étant les ω_k.

  On peut montrer que la variance intra-classe s'écrit aussi sous la forme :

$$\boxed{I^a = \frac{1}{2\omega}\sum_{k=1}^{K}\frac{1}{\omega_k}\left[\sum_{i,j \in P_k} \alpha_i \alpha_j (x_i - x_j)^2\right]}. \textbf{ (2)}$$

  Dans cette expression, le terme $\left|x_i - x_j\right|$ peut s'interpréter comme une distance[7] entre les deux unités i et j : $d_{i,j} = \left|x_i - x_j\right|$.

- Le terme $\sum_{k=1}^{K} \frac{\omega_k}{\omega}(g_k - g)^2$ représente la variance (ou inertie) *inter-classes*, d'où la relation :

  Inertie totale = Inertie intra-classe + Inertie inter-classes.

Ainsi, l'inertie totale étant une *constante de la population* (vis-à-vis du choix des variables x_i), la décomposition en les deux termes ci-dessus dépendra de la manière dont est construite la partition ; une partition pourra avoir une inertie intra-classe plus élevée qu'une autre (donc une inertie inter-classes *plus faible*) ou vice-versa.

Pour les cas limites, une inertie intra-classe nulle signifie une *homogénéité parfaite* de chacune des classes (au sein de chaque classe, toutes les variables x_i prennent la même valeur), tandis qu'une inertie inter-classes nulle exprime que les centres de gravité de toutes les classes sont identiques. On peut alors parler de *similarité parfaite* des classes, celle-ci étant mesurée par la distance du centre de gravité d'une classe par rapport au centre de gravité de la population prise dans son ensemble.

Plus les classes sont dissemblables, plus la variance inter-classes est forte et vice-versa.

[7] A strictement parler, on n'obtient de vraie distance qu'en la considérant définie sur l'ensemble-quotient de la population $\mathcal{P}$ par la relation d'équivalence : $i \sim j \Leftrightarrow \left|x_i - x_j\right| = 0.$ On convient ici que deux unités i et j ne sont pas discernables dès lors que $\left|x_i - x_j\right| = 0.$

2.3 Agrégation.

Supposons que l'on agrège les deux classes $\mathcal{P}_{k_1}$ et $\mathcal{P}_{k_2}$, les autres restant inchangées. Notons $\mathcal{P}_{k_1k_2}$ la classe résultant de cette agrégation, c'est-à-dire : $\mathcal{P}_{k_1k_2} = \mathcal{P}_{k_1} \cup \mathcal{P}_{k_2}$.

La variation d'inertie intra-classe qui en résulte (elle ne met en jeu que les classes $\mathcal{P}_{k_1}$ et $\mathcal{P}_{k_2}$) est :

$$\Delta I^a = \frac{\omega_{k_1} + \omega_{k_2}}{\omega} I_{k_1k_2} - \frac{\omega_{k_1}}{\omega} I_{k_1} - \frac{\omega_{k_2}}{\omega} I_{k_2}.$$

Elle peut s'exprimer sous la forme :

$$\boxed{\Delta I^a = \frac{(g_{k_1} - g_{k_2})^2}{\omega} \frac{\omega_{k_1}\omega_{k_2}}{\omega_{k_1} + \omega_{k_2}}} \quad \textbf{(3)}$$

Conséquence :

Par agrégation de deux classes, **la variation d'inertie intra-classe est toujours positive ou nulle, c'est-à-dire que l'inertie intra-classe ne peut qu'augmenter dans une agrégation, et l'inertie inter-classes diminuer.**

Partant d'une partition donnée, on peut chercher à construire une nouvelle partition moins fine (et de nombre de classes fixé), **minimisant l'inertie intra-classe** (ou maximisant l'inertie inter-classes). On mettra pour cela en œuvre un algorithme par agrégations successives **réalisant à chaque étape l'agrégation de deux classes** préexistantes à l'étape antérieure**, de telle sorte que la variation d'inertie intra-classe en résultant soit la plus faible possible.**

Si l'on cherche au contraire au maximiser l'inertie intra-classe, l'algorithme **réalisera à chaque étape l'agrégation de deux classes qui permettra d'obtenir la variation d'inertie intra-classe en résultant la plus forte possible.**

Dans tous les cas, cette variation d'inertie intra-classe donnée par **(3)** s'interprète comme une distance entre les classes $\mathcal{P}_{k_1}$ et $\mathcal{P}_{k_2}$ **.**

▲ Prendre garde au fait que cette distance fait intervenir les poids.

La minimisation (resp. la maximisation) de l'inertie intra-classe conduit donc à agréger à chaque étape les classes les plus proches (resp. les plus éloignées) au sens de cette distance.

3 Extensions du modèle de base

3.1 Décomposition en cascade.

L'équation d'analyse de la variance peut s'écrire : $I = \sum_{k=1}^{K} \frac{\omega_k}{\omega}(g_k - g)^2 + \sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k$.

Le premier terme de cette somme (variance inter-classes) s'interprète comme la variance des variables g_k définies sur une population dont les éléments sont les $\mathcal{P}_1$, $\mathcal{P}_2$, …, $\mathcal{P}_K$, munis des poids

ω_k , g étant le barycentre des valeurs g_k affectées de ces poids. Le second terme (variance intra-classe) ne fait intervenir que des quantités relatives à ces éléments $\mathcal{P}_k$.

Si l'on considère cette fois l'univers $\mathcal{F} = \{ \mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_K \}$, on peut lui appliquer les mêmes principes de décomposition de l'inertie lorsqu'on partitionne $\mathcal{F}$ en p sous-familles $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_p\}$.

L'inertie initiale de cette famille considérée comme univers de référence, dont les unités sont dorénavant les $\mathcal{P}_k$, a la même expression que lorsqu'on la définit sur les unités statistiques de base : elle est définie à partir des valeurs g_k affectées aux $\mathcal{P}_k$ et d'un système de poids ω_k , que l'on peut définir également de manière exogène indépendamment de l'interprétation sous-jacente. Elle s'exprimera toujours comme une moyenne pondérée des écarts au carré par rapport au barycentre, les poids des unités construites étant recalculés de manière cohérente à partir de ces ω_k lors de toute opération d'agrégation.

Finalement, il revient au même de considérer l'inertie totale à partir de la population initiale des unités statistiques considérées ou l'inertie du système de classes constituées à partir d'une certaine agrégation : la différence des deux inerties réside dans le terme $\sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k$. Dans la mesure où l'on raisonnera toujours en considérant des variations d'inertie, la présence de cette quantité (*constante si l'on fixe une première agrégation de référence une fois pour toutes*) n'a aucune importance.

3.2 Conventions de vocabulaire sur les unités statistiques

Ainsi, dans ce problème, il y a trois types d'unités statistiques :

- des « *unités statistiques élémentaires* » : ménages, individus, établissements, parcelles agricoles ou cadastrales, communes …
- des « *agrégats de base* » constitués par regroupements d'unités statistiques élémentaires : communes, cantons, arrondissements, départements, …, entreprises..
- des « *agrégats composites* » : regroupements des agrégats de base.

Dans la pratique, **on travaillera sur des unités intermédiaires, c'est-à-dire les agrégats de base**, sans revenir jamais aux données définies sur les unités statistiques élémentaires. Cependant, il est important de savoir quelles elles sont pour savoir, lors d'une agrégation d'agrégats de base, comment on calcule des valeurs « synthétiques » sur les agrégats composites obtenus.

En revanche, il faut choisir le système de poids de manière pertinente, permettant des interprétations socio-économiques valides.

3.3 Réflexion sur la signification des variables et le système de poids.

Il faut bien comprendre la signification des g_k :

- si $\alpha_k = 1$, les g_k sont des moyennes arithmétiques des valeurs x_i au sein de l'unité $\mathcal{P}_k$, et les poids ω_k sont les *tailles* de ces éléments $\mathcal{P}_k$, c'est-à dire les effectifs en nombre d'unités statistiques initiales.

- plus généralement, les g_k sont construits et s'interprètent comme des moyennes pondérées ou barycentres des valeurs x_i affectées des poids α_i .

- Mais on peut définir directement les valeurs g_k sur les agrégats Φ_k et attribuer à ces derniers des poids ω_k, en omettant leur interprétation à partir des x_i : i.e. sans se soucier de leur construction à partir des valeurs et des poids affectés aux unités statistiques initiales.

Mais ceci va poser problème dans l'agrégation. En effet, après agrégation, **on recalcule un nouveau centre de gravité à partir des poids. Cela implique une nécessaire cohérence entre les g_k et les poids.** Les g_k ne peuvent pas être des grandeurs quelconques. Ils doivent s'interpréter nécessairement, pour assurer la compatibilité avec les poids :

- comme des *moyennes pondérées* sur les unités statistiques élémentaires
- comme des *proportions* (cas particulier d'une moyenne)
- ou comme des *ratios (exemple : densité de la population au km², taux de chômage)*. Dans ce cas, les poids adéquats sont les dénominateurs des ratios.

En particulier, les g_k ne peuvent être des *totaux*, des *valeurs médianes* (exemple : revenu médian des ménages résidant dans l'agrégat Φ_k [8]) ni des gr*andeurs sans dimension* (altitude du point culminant, température maximale relevée un mois donné…).

Choix des poids

Dans un certain nombre de cas, les poids sont imposés du fait de la nature des unités statistiques élémentaires : si l'on s'intéresse au revenu des ménages, les poids seront les nombres de ménages. Toute agrégation d'unités conduira à affecter à la nouvelle classe créée le revenu moyen des ménages appartenant à la classe.

Mais parfois, le système de poids n'est pas imposé d'emblée. Si les variables définies sur les unités statistiques élémentaires sont des grandeurs sans dimension, par exemple (ex. : température maximale relevée un mois donné dans une commune), on pourra pondérer par la superficie si l'on s'intéresse plutôt aux aspects climatiques ou environnementaux ou bien par la population si l'on s'intéresse aux effets négatifs ressentis par les individus ou impactant leur santé.

Avec une seule variable quantitative s'interprétant comme une moyenne ou un taux, il n'y a aucun problème théorique pour la question de l'agrégation (exemple : revenu moyen, âge moyen, proportion des plus de 50 ans, taux de chômage…). La seule contrainte est que les poids soient cohérents avec ces variables.

La question de l'agrégation et des poids se pose dès lors qu'on raisonne avec plusieurs variables de natures différentes.

[8] Cela pose problème avec le revenu médian : le barycentre des revenus médians n'est pas égal au revenu médian d'une nouvelle classe formée par agrégation de deux classes. En fonction de la disponibilité des données, on peut toutefois faire une approximation assimilant le revenu médian à un revenu moyen et traiter ce dernier comme une variable justiciable du calcul de barycentre lors d'une agrégation.

Exemples :

Unité statistique élémentaire (i)	**Variable x_i**	**Agrégat de base (k)**	**Variable g_k**	**Poids ω_k**
Ménage	Revenu	Commune	Revenu moyen des ménages de[9] la commune	Nombre de ménages de la commune
Individu	Age	Commune	Age moyen des individus de la commune	Nombre d'individus de la commune
Individu	1 si l'individu i a plus de 50 ans, 0 sinon	Commune	Proportion des individus de plus de 50 ans résidant dans la commune	Nombre d'individus de la commune
Électeur	1 si l'individu i a voté pour le candidat A aux élections présidentielles, 0 sinon	Commune (d'inscription sur les listes électorales)	Proportion des individus inscrits sur les listes électorales de la commune k (ou ayant voté et exprimé un suffrage) ayant voté pour le candidat A	Nombre d'individus inscrits (ou ayant voté et exprimé un suffrage) sur les listes électorales de la commune
Individu en emploi	Catégorie socio-professionnelle	Commune du lieu de travail	Répartition des personnes en emploi sur une commune par catégorie socio-professionnelle	Nombre d'individus en emploi dans la commune
Individu de 15 ans ou plus.	Statut d'activité (actif occupé / chômage / en formation / autre inactif)	Commune	Répartition des personnes de 15 ans ou plus par statut d'activité	Nombre de personnes de 15 ans ou plus dans la commune
Femme	Age	Commune	Répartition des femmes par âge	Nombre de femmes dans la commune

Les deux derniers exemples montrent une difficulté intrinsèque : selon les variables d'intérêt que l'on prend sur les unités statistiques élémentaires, on a des définitions différentes des systèmes de poids associés aux agrégats de base. Cette difficulté va se répercuter toutes les fois que l'on cherchera à travailler sur des variables composites.

Exemple : si l'on pondère les votants par la taille des ménages, on arrive à des paradoxes gênants :

Soit un ménage comprenant 4 personnes dont 2 électeurs :
- si l'un des électeurs a voté pour FH et l'autre pour NS, cela reviendrait à affecter à ce ménage 2 votes pour FH et 2 pour NS
- si les deux ont voté pour NS, on affecterait au ménage 4 votes pour NS.

[9] Résidant dans…

3.4 Utilisation conjointe de plusieurs variables quantitatives

Lorsqu'on dispose de deux ou plusieurs variables numériques **associées à des systèmes de poids identiques**, la solution adoptée dans ce papier consiste à :

- normaliser les variables en les divisant par leur écart-type empirique dans la population de départ ;
- additionner les inerties relatives à chacune de ces variables normalisées.

La normalisation a pour but de ***rendre comparables les différentes variables***, en se ramenant à des grandeurs sans dimension, ce qui légitime l'additivité des inerties. En particulier, des problèmes d'unités et d'ordres de grandeur rendraient inadaptée une addition brute des composantes (exemple : utilisation des variables revenu annuel du ménage / nombre d'individus du ménage).

Mais l'addition des inerties ne permet pas la prise en compte d'éventuelles corrélations entre les variables. D'un certain point de vue, l'existence de corrélations indique que les variables peuvent avoir un contenu commun en termes d'information apportée, qui sera compté plusieurs fois si l'on ajoute les inerties. L'addition des ineries simplement et se justifie par le fait de ***donner la même importance à chacune des variables***.

Notations : si l'on travaille sur les agrégats de base avec p variables dont les valeurs sur l'agrégat k sont notées $g_{1,k}, g_{2,k}, ..., g_{p,k}$, chaque agrégat étant muni d'un poids ω_k, l'inertie totale considérée sera : $\boxed{I = \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{K}\frac{\omega_k}{\omega}(g'_{i,k} - g'_i)^2}$, avec :

- $g'_{i,k} = \dfrac{g_{i,k}}{\sqrt{\sum_{l=1}^{K}\frac{\omega_l}{\omega}(g_{i,l} - g_i)^2}} = \dfrac{g_{i,k}}{\sigma_i}$ (***variable réduite***),
- σ_i = écart-type de la $i-ème$ variable
- $g'_i = \frac{1}{\omega}\sum_{k=1}^{K}\omega_k g'_{i,k}$ (centre de gravité de la $i-ème$ variable).

Cette formule fournit **l'inertie totale initiale** (c'est-à-dire avant toute agrégation : c'est **l'inertie totale initiale des agrégats de base**), le facteur $\frac{1}{p}$ lui donnant la valeur conventionnelle 1, et c'est à partir de ces valeurs $g'_{i,k}$ que l'on recalculera les distances entre classes ou les variations d'inertie intra-classes dans le processus d'agrégation.

D'autres solutions dont possibles si l'on veut tenir compte de la corrélation entre les variables et se ramener à des variables non corrélées entre elles :

- construire de nouvelles variables non corrélées entre elles, par exemple en mettant en œuvre une version aléatoire de l'algorithme de SCHMIDT pour construite une base orthogonale dans un espace vectoriel préhilbertien (voir annexe).

- plus généralement, construire une famille de nouvelles variables non corrélées deux à deux par une procédure de ***réduction générale***.

Celle s'obtient de la manière suivante : si l'on considère un vecteur aléatoire X à valeurs dans $\boldsymbol{R}^p$, dont les composantes sont p variables aléatoires de carré intégrable, $X_1, X_2, ..., X_p$, et si l'on note Σ sa matrice de variance-covariance (supposée *inversible*), alors le vecteur $Y = \Sigma^{-1/2} X$ a pour variance I_p : en particulier, ses composantes sont deux à deux non corrélées.

La matrice $\Sigma^{-1/2}$ se calcule sous la forme : $\Sigma^{-1/2} = P\Delta^{-1/2}P'$, où P est la matrice de changement de base vers les vecteurs propres de Σ (P' sa transposée) et $\Delta^{-1/2}$ la matrice diagonale dont les éléments sont $\frac{1}{\sigma_i}$.

Ici, les variables aléatoires X_i sont des variables discrètes prenant les valeurs $g_{i,k}$ sur chaque classe $\mathcal{P}_k$, élément de la population des agrégats statistiques de base, avec des probabilités (= *poids*) ω_k.

Cela revient à considérer une métrique non plus euclidienne canonique mais s'appuyant sur la matrice Σ^{-1}.

- faire de l'**ACP**. Cela revient à « résumer » les variables initiales par un nombre plus petit de variables, les ***composantes principales***, ayant le « pouvoir discriminant » le plus fort.
  Les composantes principales sont les variables $Y_{u_i} = u'_i X$, où les u_i sont les vecteurs propres de Σ (éléments de $\boldsymbol{R}^p$).

  Les résultats classiques montrent que le vecteur aléatoire Y de composantes Y_{u_i} a pour matrice de variance-covariance : $\begin{pmatrix} \lambda_1 & 0 & . & 0 \\ 0 & \lambda_2 & . & . \\ . & . & . & 0 \\ 0 & . & 0 & \lambda_p \end{pmatrix}$, où les λ_i sont les valeurs propres de Σ.

  Finalement, *les composantes principales de X sont des variables aléatoires Y_u, combinaisons linéaires des composantes de X, non corrélées entre elles et dont les variances sont les valeurs propres de la matrice de variance-covariance de X.*

  Les composantes principales correspondant à la plus petite (resp. la plus grande) valeur propre sont les combinaisons linéaires des composantes de X de plus faible (resp. de plus grande variance).

3.5 Généralisation à plusieurs variables quantitatives

On peut généraliser l'approche théorique précédente avec plusieurs variables quantitatives, même associées à des systèmes de poids différents.

On additionnera les inerties résultant de chacun des groupes de variables, chacun avec leur système de poids propre. On normalise à 1 l'inertie de chacune des variables.

Dans ce cas, les valeurs des variables relatives à l'agrégat constitué sont calculées comme barycentre des valeurs relatives aux agrégats initiaux, affectés des poids correspondant à chaque variable.

3.6 Cas des variables catégorielles.

Cadre général :

On considère une population finie $\mathcal{P}$ constituée de K agrégats de base : $k = 1,...,K$. Chaque agrégat k contient n_k unités statistiques élémentaires. Le nombre total de ces unités (taille totale de la population) est :

$$N = \sum_{k=1}^{K} n_k \ .$$

Considérons une variable catégorielle à Q modalités définie sur les unités statistiques élémentaires : par exemple, la classe d'âge ou la CS pour la répartition des individus, l'APET pour la répartition des établissements …

Pour un agrégat de base k, on note $n_{k,q}$ le nombre d'unités élémentaires de l'agrégat k appartenant à la modalité q de la variable.

Il est clair que la distance entre deux agrégats au vu de cette variable ne peut pas reposer sur les quantités $n_{k,q}$, qui sont les effectifs « absolus » ; en effet, si l'on a deux agrégats avec une répartition d'une variable en 3 modalités, l'un de 6 unités réparties (1, 2, 3), l'autre de 6000 unités réparties (1000, 2000, 3000), les répartitions sont identiques dans les deux unités (17 %, 33 %, 50 %) alors que la distance euclidienne calculée à partir des valeurs brutes serait grande.

Suivant ici la théorie et la pratique de l'analyse des données (cf. LEBART, MORINEAU, TABARD), on va utiliser une **distance du χ^2** pour comparer deux agrégats :

On définit, pour un agrégat k : $p_{k,q} = \dfrac{n_{k,q}}{n_k}$, la proportion d'unités de l'agrégat correspondant à la modalité q de la variable, où $n_k = \sum_{q=1}^{Q} n_{k,q}$ est **l'effectif (en nombre d'unités élémentaires)** de l'agrégat k. On a évidemment : $p_{k,\bullet} = \sum_{q=1}^{Q} p_{k,q} = 1$.

La distance entre les agrégats k et l sera alors : $\boxed{d^2_{k,l} = \sum_{q=1}^{Q} \frac{(p_{k,q} - p_{l,q})^2}{\pi_q}}$, avec :

$\boxed{\pi_q = \dfrac{1}{N}\sum_{k=1}^{K} n_{k,q} = \dfrac{n_{\bullet,q}}{N}}$,

qui représente la proportion empirique, dans la population des unités élémentaires, de celles qui correspondent à la modalité q de la variable.

▲ **K est le nombre d'agrégats de base, N le nombre d'unités statistiques élémentaires.**

On peut écrire cette distance sous la forme : $\boxed{d^2_{k,l} = \sum_{q=1}^{Q} \left(\frac{p_{k,q}}{\sqrt{\pi_q}} - \frac{p_{l,q}}{\sqrt{\pi_q}}\right)^2}$.

C'est cette distance qui sera utilisée pour calculer une inertie intra-classe dans le cas de variables catégorielles, en se référant à la formulation (2).

Il s'agit donc d'une distance euclidienne calculée sur une nouvelle variable de $\boldsymbol{R}^Q$ qui, pour l'agrégat générique k, a pour $q-ième$ composante : $\boxed{Z_{k,q} = \frac{p_{k,q}}{\sqrt{\pi_q}} = \frac{n_{k,q}}{n_k}\frac{\sqrt{N}}{\sqrt{n_{\bullet,q}}}}$.

Ceci permettra de traiter des variables catégorielles comme des variables quantitatives.

- Moyenne de cette variable :

  On utilise ici des **poids des agrégats proportionnels à leurs tailles, en nombre d'unités statistiques élémentaires**, soit : $\alpha_k = \frac{n_k}{N}$.

$$\boxed{\overline{Z}_q = \frac{1}{N}\sum_{k=1}^{K} n_k Z_{k,q} = \frac{1}{N}\sum_{k=1}^{K} n_k \frac{n_{k,q}}{n_k}\frac{\sqrt{N}}{\sqrt{n_{\bullet,q}}} = \frac{1}{N}\frac{\sqrt{N}}{\sqrt{n_{\bullet,q}}}\underbrace{\sum_{k=1}^{K} n_{k,q}}_{=n_{\bullet,q}} = \sqrt{\frac{n_{\bullet,q}}{N}}}.$$

- Variance :

$$\begin{aligned}
\Sigma_q^2 &= \frac{1}{N}\sum_{k=1}^{K} n_k (Z_{k,q} - \overline{Z}_q)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k Z_{k,q}^2 - \overline{Z}_q^2 \\
&= \frac{1}{N}\sum_{k=1}^{K} n_k \left(\frac{n_{k,q}^2}{n_k^2}\frac{N}{n_{\bullet,q}}\right) - \left(\sqrt{\frac{n_{\bullet,q}}{N}}\right)^2 \\
&= \frac{1}{N}\sum_{k=1}^{K} \frac{n_{k,q}^2}{n_k}\frac{N}{n_{\bullet,q}} - \frac{n_{\bullet,q}}{N} \\
&= \frac{1}{n_{\bullet,q}}\sum_{k=1}^{K} \frac{n_{k,q}^2}{n_k} - \frac{n_{\bullet,q}}{N}.
\end{aligned}$$

- Inertie totale :

  On part de l'expression rappelée ci-dessus (formule **(2)**, en la transposant pour une inertie totale) : $I = \frac{1}{2\omega^2}\sum_{k=1}^{K}\sum_{l=1}^{K} \alpha_k \alpha_l d_{k,l}^2$, avec :

$$\alpha_k = \frac{n_k}{N} \text{ (poids de l'agrégat } k\text{)}, \ \omega = \sum_{k=1}^{K} \alpha_k = 1 \text{ et :}$$

$$d_{k,l}^2 = \sum_{q=1}^{Q} (Z_{k,q} - Z_{l,q})^2 = N\sum_{q=1}^{Q} \frac{1}{n_{\bullet,q}}\left(\frac{n_{k,q}}{n_k} - \frac{n_{l,q}}{n_l}\right)^2.$$

  On a donc :

$$\sum_{k=1}^{K}\sum_{l=1}^{K} \alpha_k \alpha_l d_{k,l}^2 = N\sum_{k=1}^{K}\sum_{l=1}^{K} \frac{n_k}{N}\frac{n_l}{N}\left[\sum_{q=1}^{K} \frac{1}{n_{\bullet,q}}\left(\frac{n_{k,q}}{n_k} - \frac{n_{l,q}}{n_l}\right)^2\right]$$

$$= N\sum_{q=1}^{Q}\left[\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{n_k}{N}\frac{n_l}{N}\frac{1}{n_{\bullet,q}}\left(\frac{n_{k,q}}{n_k}-\frac{n_{k,q}}{n_l}\right)^2\right]$$
$$= \frac{1}{N}\sum_{q=1}^{Q}\frac{1}{n_{\bullet,q}}\left[\sum_{k=1}^{K}\sum_{l=1}^{K}n_k n_l\left(\frac{n_{k,q}}{n_k}-\frac{n_{l,q}}{n_l}\right)^2\right].$$

Or :

$$\sum_{k=1}^{K}\sum_{l=1}^{K}n_k n_l\left(\frac{n_{k,q}}{n_k}-\frac{n_{l,q}}{n_l}\right)^2 = \sum_{k=1}^{K}\sum_{j=1}^{K}n_k n_l\left(\frac{n_{k,q}^2}{n_k^2}+\frac{n_{l,q}^2}{n_l^2}-2\frac{n_{k,q}}{n_k}\frac{n_{l,q}}{n_l}\right)$$
$$= \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n_l\frac{n_{k,q}^2}{n_k}+n_k\frac{n_{l,q}^2}{n_l}-2n_{k,q}n_{l,q}\right).$$

Le terme $\sum_{k=1}^{K}\sum_{l=1}^{K}n_l\frac{n_{k,q}^2}{n_k}$ s'écrit : $\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}\left(\sum_{l=1}^{K}n_l\right) = N\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}$.

Par symétrie : $\sum_{k=1}^{K}\sum_{l=1}^{K}n_k\frac{n_{l,q}^2}{n_l} = N\sum_{l=1}^{K}\frac{n_{l,q}^2}{n_l} = N\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}$.

Et : $-2\sum_{k=1}^{K}\sum_{l=1}^{K}n_{k,q}n_{l,q} = -2\sum_{k=1}^{K}n_{k,q}\underbrace{\left(\sum_{l=1}^{K}n_{l,q}\right)}_{=n_{\bullet,q}} = -2n_{\bullet,q}\underbrace{\left(\sum_{k=1}^{K}n_{k,q}\right)}_{=n_{\bullet,q}} = -2n_{\bullet,q}^2$.

Par suite : $\sum_{k=1}^{K}\sum_{l=1}^{K}n_k n_l\left(\frac{n_{k,q}}{n_k}-\frac{n_{l,q}}{n_l}\right)^2 = 2N\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}-2n_{\bullet,q}^2$.

On en déduit :

$$I = \frac{1}{2N}\sum_{q=1}^{Q}\frac{1}{n_{\bullet,q}}\left[2N\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}-2n_{\bullet,q}^2\right]$$
$$= \sum_{q=1}^{Q}\frac{1}{n_{\bullet,q}}\left[\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}-\frac{1}{N}n_{\bullet,q}^2\right] = \sum_{q=1}^{Q}\frac{1}{n_{\bullet,q}}\left(\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_k}\right)-\frac{1}{N}\underbrace{\sum_{q=1}^{Q}\frac{1}{n_{\bullet,q}}n_{\bullet,q}^2}_{=\sum_{q=1}^{Q}n_{\bullet,q}=N},$$

d'où :

$$\boxed{I = \sum_{q=1}^{Q}\sum_{k=1}^{K}\frac{n_{k,q}^2}{n_{\bullet,q}n_k}-1.}\ \textbf{(4)}$$

C'est la formule pour l'inertie totale. Elle peut se généraliser sans problème pour calculer l'inertie intra-classe.

- Effet d'une agrégation.

  Si l'on agrège deux agrégats k et l, on peut recalculer les variables $Z_{k+l,q}$ relatives au nouvel agrégat constitué :

$$Z_{k+l,q} = \frac{p_{k+l,q}}{\sqrt{\pi_q}} = \frac{1}{\sqrt{\pi_q}} \frac{n_{k+l,q}}{n_{k+l}} = \frac{1}{\sqrt{\pi_q}} \frac{n_{k,q} + n_{l,q}}{n_{k+l}} = \frac{1}{\sqrt{\pi_q}} \frac{n_k \sqrt{\pi_q} Z_{k,q} + n_l \sqrt{\pi_q} Z_{l,q}}{n_{k+l}},$$

  soit :

$$Z_{k+l,q} = \frac{n_k Z_{k,q} + n_l Z_{l,q}}{n_{k+l}} = \frac{n_k}{n_{k+l}} Z_{k,q} + \frac{n_l}{n_{k+l}} Z_{l,q},$$

  ou :

$$\boxed{Z_{k+l,q} = \frac{n_k}{n_k + n_l} Z_{k,q} + \frac{n_l}{n_k + n_l} Z_{l,q}}.$$

  On obtient le barycentre des variables $Z_{k,q}$ et $Z_{l,q}$ avec des coefficients égaux aux poids relatifs des agrégats initiaux en termes de nombre d'unités statistiques élémentaires (= effectif des agrégats).

3.7 Cas de plusieurs variables catégorielles

On crée autant de familles de variables quantitatives qu'il y a de variables qualitatives, selon le procédé ci-dessus. Mais on renormalisera les variables $Z_{k,q}$ de façon que l'inertie résultante (calculée en **(4)**) soit égale à 1, pour chacune des variables qualitatives.

3.8 Cas mixte : mélange de variables catégorielles et quantitatives

Les variables catégorielles sont converties en variables quantitatives selon la méthodologie exposée ci-dessus en renormalisant les variables $Z_{k,q}$: ainsi, l'inertie résultante est égale à 1, pour chacune des variables qualitatives. De même, l'inertie relative à chaque variable quantitative est prise à la valeur 1 par normalisation adéquate.

4 Mise en œuvre informatique

4.1 Description de l'algorithme

On rappelle ici les principes généraux de l'algorithme mis en œuvre. La description détaillée figure dans [JMS 2012].

L'algorithme se déroule en deux phases :

- La première phase vise à effectuer une première agrégation de la population de référence en un nombre de classes connexes choisi par l'utilisateur, respectant du mieux possible (cf. infra sur le relâchement des contraintes) les contraintes de taille fixées et maximisant ou minimisant la variance intra-classe.
- La seconde étape améliore la partition trouvée lors de la première étape en procédant à des *échanges* pour améliorer si nécessaire le respect des contraintes de taille et en optimisant la

variance intra-classe. Cette étape est rendue nécessaire parce que la méthode utilisée dans la première partie n'est pas optimale.

4.1.1 La première phase : une classification ascendante hiérarchique contiguë

La méthode utilisée ici est une méthode directement inspirée par la Classification Ascendante Hiérarchique (CAH) qui consiste, à une étape *t* de l'algorithme, à regrouper les deux classes précédemment créées les plus « proches », c'est-à-dire minimisant l'augmentation de la variance intra-classe liée à leur agrégation, ou les plus « distantes » (si l'on veut maximiser l'augmentation de la variance intra-classe).

Par rapport à la CAH classique, cet algorithme a été modifié comme suit :

- L'agrégation ne peut avoir lieu qu'entre classes ***contiguës***, puisque l'on veut obtenir une partition en classes connexes ;
- Selon le choix de l'utilisateur, les classes agrégées correspondent soit à une augmentation d'inertie maximale (si l'utilisateur souhaite maximiser l'inertie intra-classe), soit minimale (dans le cas contraire).
- Le traitement des contraintes de taille sera décrit ci-dessous.

Cet algorithme récursif est initialisé avec la partition formée de tous les singletons correspondant à chaque agrégat de base. Si au cours d'une itération, l'agrégation optimale (du point de vue de la variation d'inertie) conduit à constituer une classe de taille supérieure à la taille maximale, cette agrégation n'a pas lieu et on recherche l'agrégation optimale *seulement sous contrainte d'éligibilité de la taille résultante*. Ainsi, on privilégie les contraintes de taille par rapport à l'optimalité de l'inertie.

Dans le cas où la procédure ne peut se poursuivre à cause des contraintes de taille, un message est adressé à l'utilisateur et une taille maximale plus grande de 25% est fixée temporairement, puis l'algorithme se poursuit avec ce nouveau paramètre[10].

4.1.2 La seconde phase : échanges ou transfert d'unités entre classes

Sous l'ensemble des contraintes mentionnées ci-dessus (recours à un algorithme de type « CAH), contraintes de contiguïté et de taille), un optimum local est atteint.

Toutefois, comme nous l'avons indiqué ci-dessus, les contraintes initiales de taille fixées par l'utilisateur peuvent ne pas être respectées.

La seconde phase va donc chercher à réduire, en premier lieu, l'écart des tailles effectives des classes constituées par rapport aux contraintes de taille, puis ensuite à optimiser l'inertie intra-classe. Elle va procéder par transferts d'unités d'une classe à une autre ou par échanges de deux unités appartenant à deux classes différentes.

Pour tester les contraintes de taille, une fonction additive, dite « critère », va être utilisée. Elle vaut, pour chaque classe :

- 0 si la taille de la classe vérifie les contraintes
- ou la valeur absolue de la différence entre la taille de la classe et la borne de taille la plus proche, soit, pour une classe G ne vérifiant pas les contraintes :

[10] A nouveau, ce relâchement de contraintes ne s'applique pas si la contrainte survient à une étape où l'on a déjà réduit le nombre de classes obtenu par rapport au nombre souhaité.

$$CRIT(G) = \begin{cases} (taille(G) - T_{\max})^2 & si\ taille(G) > T_{\max} \\ ou & \\ (T_{\min} - taille(G))^2 & si\ taille(G) < T_{\min} \end{cases}$$

La méthode est relativement simple : pour tous les échanges ou transferts possibles d'unités, les influences sur les contraintes de taille et la variation de l'inertie sont calculées, puis on procède à l'échange qui diminue le plus le critère ou, si les contraintes de taille sont vérifiées, qui optimise le plus l'inertie intra-classe.

Naturellement, un échange ou un transfert ne sont licites que si les classes en résultant sont tous les deux connexes. Il convient donc de vérifier cette connexité avant de calculer les variations de critère ou de variance intra-classe.

Cette seconde phase s'arrête lorsqu'il n'est plus possible d'améliorer le critère ou l'inertie intra-classe. Chacune de ces itérations pouvant être longue, l'utilisateur peut en fixer un nombre maximal (plusieurs centaines) ou un nombre d'itérations dépendant du temps de calcul maximal admissible.

4.2 Mise en œuvre de la méthode

Les programmes mis en œuvre permettent de simuler la construction de « régions » sous les contraintes suivantes :

- on se limite à un découpage de la ***France continentale***, les notions de contiguïté n'ayant pas de sens dans un contexte géographique plus large ou nécessitant une autre définition que celle de la contiguïté terrestre.

- le nombre d'unités à construire est fixée à ce que définit l'actuelle loi votée par le Parlement, soit **12**

- des simulations sont construites sous les critères de maximisation ou de minimisation de l'inertie.

- *contraintes de taille* : exprimées sous la forme de % de la population des régions construites par rapport à la population totale. Les seuils utilisés sont [4%, 19%], correspondant aux extrema pour les régions continentales créées par le Parlement.

- différentes variables servant à la définition de l'inertie sont utilisées, soit séparément, soit combinées.

- les unités de base qui sont agrégées peuvent être : les régions actuelles (c'est ainsi qu'a procédé le Parlement), les départements (hypothèse évoquée pour le futur : des départements pourraient changer de région), les arrondissements, les zones d'emploi ou, au niveau le plus élémentaire, les cantons, voire les communes.

- en réalité, l'agrégation des régions actuelles aura surtout pour but de comparer les solutions obtenues à celle votée par le Parlement. A l'inverse, travailler avec une maille trop ténue, donc un nombre d'unités de base très important (notamment, avec les communes) conduit actuellement à des temps de calcul prohibitifs. Il s'agit là très certainement d'une limite et d'une voie d'amélioration de l'outil pour le futur. **Aussi la décision a-t-elle été prise de se limiter à des unités de niveau d'agrégation supérieur ou égal à celui du canton.** On construit donc des « régions » par agrégation de cantons-villes (3646)[11], d'arrondissements (325), de zones d'emploi (297), de départements (94) ou à partir des régions actuelles (21)[12].

- le programme produit des cartes, montrant les « régions » construites , en comparaison avec les contours des actuelles régions. Il produit également des statistiques, d'une part sur les

[11] Il s'agit des ***anciens cantons*** et non de ceux qui ont servi pour les élections départementales des 22 et 29 mars 2015…

[12] Paris est alors à la fois un département, un arrondissement et un canton.

variables qui ont servi à la constitution de ces « régions », d'autre part sur d'autres variables (% de votants N. Sarkozy / F. Hollande au 2ème tour de l'élection présidentielle de 2012).

Les données utilisées sont obtenues des sources suivantes :

- données sur le revenu (sources fiscales 2010)
- données sur la répartition de la population en 6 classes d'âge : 0-14 ans / 15-29 / 30-44 / 45-59 / 60-74 / ≥ 75 (recensement 2010).

Ces données sont d'abord définies et récupérées au niveau communal puis recalculées au niveau des agrégats statistiques élémentaires.

4.3 Traitement de la contiguïté.

La contiguïté est traitée d'abord au niveau élémentaire (communal). Pour la définir entre deux entités géographiques quelconques (ensemble de communes), on se réfère au cadre général développé dans [JMS 2012] :

Soit E un ensemble non vide muni d'une relation *R* réflexive et symétrique, dite *relation de contiguïté*.

Cette relation, définie sur E, peut s'étendre à deux *parties* de E.

On dit que deux parties A et B de E, non vides, sont *contiguës* si et seulement si :

$$\exists\, x \in A, \exists\, y \in B : x\ R\ y\ .$$

Ainsi, deux entités seront dites contiguës si chacune d'elles contient au moins une commune contiguë à une commune de l'autre.

4.4 Quelques contraintes pratiques.

Les premiers essais de simulation mis en œuvre ont montré que les zones obtenues avaient souvent une forme « serpentine » confinant à l'absurde pour les « régions » recréées. Pour limiter cet effet pervers, une autre contrainte a été introduite : **deux unités ne peuvent être agrégées que si la longueur de la diagonale du rectangle, dont les sommets sont les points de coordonnées respectivement égales au min ou au max des abscisses et ordonnées des centres des communes composant l'agrégat envisagé, est inférieure ou égale à un seuil. Cette contrainte ne joue que lors de la 1ère phase. Différents tests ont été réalisés avec des seuils allant de 0 à 500 Km par tranche de 100.**

Commentaire sur les seuils : il est clair qu'un seuil trop petit bloque l'algorithme très vite qui, pour continuer, doit être autorisé à violer le seuil : ainsi seules les premières unités agrégées ont respecté la contrainte de seuil. A l'inverse, un seuil trop grand ne joue évidemment pas. Il y a donc très certainement un seuil optimum, à déterminer empiriquement, pour que la contrainte joue au bout d'un nombre d'itérations assez avancées et évite la construction de macro-régions trop étirées en longueur.

4.5 Mise en œuvre informatique.

Des macros SAS ayant permis de réaliser les simulations de construction de régions décrites dans ce papier sont disponibles sur demande. Elles peuvent être utilisées avec différents paramétrages. Elles peuvent être réemployées soit dans le même objectif, en variant les contraintes ou les variables mises en jeu, soit pour des objectifs différents mais participant de la même logique, à des niveaux géographiques éventuellement plus fins (partition d'une ville en quartiers à partir des îlots par exemple).

5 Les résultats

Les résultats sont produits sous forme de cartes, en jouant sur les différents paramètres suivants :

- maximisation ou minimisation de l'inertie
- variables mises en jeu : revenu seul, répartition de la population en classes d'âge seule, mélange des deux
- avant / après optimisation
- avec seuil de 0, 100, 200, …, 500 Km
- agrégats de base : régions actuelles, départements, cantons-villes, zones d'emploi

Les cartes ne sont pas insérées dans cette version préliminaire du papier. Elles le seront dans le diaporama présenté lors de la session des JMS et seront incluses ultérieurement dans les versions révisées du papier.

6 Conclusion.

Les simulations de constructions de régions présentées dans ce papier restent des exercices d'école, qui ne sauraient se substituer aux décisions de la représentation nationale.

Elles mettent en évidence plusieurs difficultés :

- d'une part définir des critères statistiques objectivables

- les mettre en œuvre de façon simple : algorithmes complexes, ne conduisant qu'à des optima locaux, temps de calcul augmentant rapidement lorsqu'on travaille sur des agrégats statistiques de base plus petits et plus nombreux

- des contraintes souvent incompatibles : comment gérer simultanément l'optimisation de l'inertie, le respect des contraintes de taille, l'obtention de « régions » pas trop serpentines mais bornées dans un rectangle adéquat … C'est de l'optimisation multicritères dans laquelle le choix de l'opérateur de privilégier tel ou tel critère au détriment d'un autre est prépondérant .. ou alors il faut les pondérer mais là encore l'arbitraire dans le choix des poids est irréductible. Dans la méthode présentée ici, c'est le processus d'échange qui assure cet arbitrage implicite entre inertie et contraintes de taille.

- les solutions sont instables lorsque l'on fait bouger les paramètres, en témoignent les cartes obtenues très différentes selon les spécifications retenues.

Voies de progrès et d'amélioration :

- optimiser l'algorithmique et le temps de calcul
- régler de manière théorique générale la prise en compte simultanée de plusieurs variables es, qualitatives et quantitatives
- étendre la méthode au cas de comparaison sur des statistiques non linéaires : par exemple, fonder la distance entre classes sur la comparaison de distributions de variables quantitatives telles que le revenu ou la comparaison de quantiles.

Mais surtout :

La méthode ou la recherche de solutions méthodologiques à un problème statistique complexe et multi-critères n'a de sens que si l'on est au clair sur les objectifs poursuivis. Cherche-t-on à faire des régions qui soient des « petites France », c'est-à-dire chacune d'elles un modèle réduit de la France : auquel cas, c'est bien l'approche en termes de *maximisation* de l'inertie intra-classe qui est la bonne modélisation. Ou bien cherche-t-on à faire des régions très homogènes mais se différenciant fortement ? Cette optique justifie alors la *minimisation* de l'inertie.

Nul doute que les parlementaires n'avaient pas nécessairement de préoccupations statistiques, ou alors, si c'était le cas, soit elles étaient *non dites* (arrière-pensées électoralistes par exemple), soit elles étaient *difficilement exprimables en termes statistiques* (respecter des équilibres socio-démographiques, culturels, historiques..).

Force est de constater que la solution qu'ils ont votée réalise un compromis entre les différentes contraintes statistiques que ce papier cherchait à mettre en évidence et résoudre.

■ ■ ■ ■ ■ ■

7 RÉFÉRENCES BIBLIOGRAPHIQUES.

C. ROCHE : « exemple de classification hiérarchique avec contraintes de contiguïté : le partage d'Aix-en-Provence en quartiers homogènes », les Cahiers de l'Analyse des Données, Vol III - 1978 - n°3.

L. LEBART : « Programme d'agrégation avec contraintes », les Cahiers de l'Analyse des Données, Vol III - 1978 - n°3

ANNEXE 1

Encadré : Algorithme de GRAM-SCHMIDT aléatoire

Considérons K variables aléatoires de carré intégrable, $X_1, X_2, ..., X_K$. On note σ_k leurs écarts-types : $\sigma_k = \sqrt{VX_k}$, supposés > 0. L'espace L_2 des variables aléatoires de carré intégrable est muni de l'orthogonalité au sens de l'opérateur covariance.

Nota : pour appliquer au problème considéré, il faut comprendre que les variables aléatoires X_i sont des variables discrètes prenant les valeurs $g_{i,k}$ sur chaque classe $\mathcal{P}_k$ de la population des agrégats statistiques de base avec des probabilités (= *poids*) ω_k.

- on commence par normer chacune des variables en posant : $X'_k = \dfrac{X_k}{\sigma_k}$, d'où :

$$\boxed{VX'_k = 1}.$$

  Cette opération a pour but de rendre « comparables » les différentes variables, en gommant les effets d'échelle ou d'unités de mesure.

- on construit alors par récurrence les variables aléatoires suivantes :

$$\boxed{\begin{cases} Y_1 = X'_1 \\ Y_k = X'_k - P^{\perp}_{(Y_1,...,Y_{k-1})}(X'_k),\ pour\ k = 2,..,K \end{cases}}.$$

  A la $k-ième$ étape de l'algorithme, $P^{\perp}_{(Y_2,...,Y_{k-1})}(X'_k)$ est le projeté orthogonal de la variable X'_k sur le sous-espace vectoriel engendré par les $k-1$ variables construites aux étapes précédentes : $(Y_1,...,Y_{k-1})$; il est clair que ce sous-espace vectoriel est aussi celui engendré par les variables $(X_1,...,X_{k-1})$.

  Ce projeté orthogonal a pour expression :

$$\boxed{P^{\perp}_{(Y_1,...,Y_{k-1})}(X'_k) = \sum_{j=1}^{k-1} \frac{Cov(X'_k, Y_j)}{VY_j} Y_j},$$

  La quantité $\dfrac{Cov(X_k, Y_j)}{VY_j} Y_j$ s'interprète comme la projection orthogonale de la variable aléatoire X_k sur la variable Y_j. Par suite :

$$\begin{cases} Y_1 = \dfrac{X_1}{\sigma_1} \\ Y_k = \dfrac{X_k}{\sigma_k} - \sum_{j=1}^{k-1} \dfrac{Cov(X_k, Y_j)}{\sigma_k VY_j} Y_j = \dfrac{1}{\sigma_k}\left[X_k - \sum_{j=1}^{k-1} \dfrac{Cov(X_k, Y_j)}{VY_j} Y_j \right], \text{ pour } k = 2,..,K \end{cases}.$$

Pour construire les variables Y_k, on défalque donc de chaque variable X'_k ses projections orthogonales sur les variables Y_j construites aux étapes antérieures et on ne conserve que le « résidu ».

Les variables Y_k sont alors, par construction de proche en proche, non corrélées deux à deux. Intuitivement, on conserve de chaque variable X_k la composante qui n'est pas corrélée aux variables déjà construites.

Démonstration : par récurrence.

- $Y_2 = \dfrac{1}{\sigma_2}\left[X_2 - \dfrac{Cov(X_2, Y_1)}{VY_1} Y_1 \right]$, d'où :

$$Cov(Y_2, Y_1) = \frac{1}{\sigma_2}\left[Cov(X_2, Y_1) - \frac{Cov(X_2, Y_1)}{VY_1} VY_1 \right] = 0 .$$

- Si les Y_j sont deux à deux non corrélées pour $1 \le j \le k-1$, alors, pour $1 \le i \le k-1$:

$$Cov(Y_k, Y_i) = \frac{1}{\sigma_k}\left[Cov(X_k, Y_i) - \sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} \underbrace{Cov(Y_j, Y_i)}_{=0 \text{ si } j \neq i} \right]$$

$$= \frac{1}{\sigma_k}\left[Cov(X_k, Y_i) - \frac{Cov(X_k, Y_i)}{VY_i} VY \right] = 0 .$$

On a alors, pour k **fixé** ≥ 2 :

$$VY_k = \frac{1}{\sigma_k^2} V\left[X_k - \sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} Y_j \right].$$

$$V\left[X_k - \sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} Y_j \right] = VX_k - 2Cov[X_k, \sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} Y_j] + V\left[\sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} Y_j \right]$$

$$= \sigma_k^2 - 2\sum_{j=1}^{k-1} \frac{Cov^2(X_k, Y_j)}{VY_j} + V\left[\sum_{j=1}^{k-1} \frac{Cov(X_k, Y_j)}{VY_j} Y_j \right]$$

$$= \sigma_k^2 - 2\sum_{j=1}^{k-1} \frac{Cov^2(X_k, Y_j)}{VY_j} + \sum_{j=1}^{k-1} \left[\frac{Cov(X_k, Y_j)}{VY_j} \right]^2 VY_j$$

[puisque les Y_j sont deux à deux non corrélées pour $1 \leq j \leq k-1$]

$$= \sigma_k^2 - \sum_{j=1}^{k-1} \frac{Cov^2(X_k, Y_j)}{VY_j}.$$

Par suite :

$$\boxed{VY_k = 1 - \frac{1}{\sigma_k^2} \sum_{j=1}^{k-1} \frac{Cov^2(X_k, Y_j)}{VY_j}}. \textbf{ (1)}$$

Calcul des covariances $Cov(X_k, Y_j)$.

Notons $\rho_{i,j}$ le coefficient de corrélation entre X_i et X_j.

- Pour $k = 2$ et $j = 1$: $Cov(X_2, Y_1) = \frac{1}{\sigma_1} Cov(X_2, X_1) = \rho_{1,2}\sigma_2$.

- Pour $k \geq 3$ et $j = 1$: $Cov(X_k, Y_1) = \frac{1}{\sigma_1} Cov(X_k, X_1) = \rho_{1,k}\sigma_k$.

- Pour $k \geq 3$ et $2 \leq j \leq k-1$:

$$Cov(X_k, Y_j) = Cov\left(X_k, \frac{1}{\sigma_j}\left[X_j - \sum_{i=1}^{j-1} \frac{Cov(X_j, Y_i)}{VY_i} Y_i \right]\right)$$

$$= \frac{1}{\sigma_j}\left[Cov(X_k, X_j) - \sum_{i=1}^{j-1} \frac{Cov(X_j, Y_i)}{VY_i} Cov(X_k, Y_i) \right]$$

$$= \frac{1}{\sigma_j}\left[\rho_{j,k}\sigma_j\sigma_k - \sum_{i=1}^{j-1} \frac{Cov(X_j, Y_i)}{VY_i} Cov(X_k, Y_i) \right]$$

Ces différentes formules permettent, de manière récursive, de calculer les variances VY_k.

Ainsi, par exemple :

- $VY_2 = 1 - \frac{1}{\sigma_2^2} \frac{Cov^2(X_2, Y_1)}{\underbrace{VY_1}_{=1}} = 1 - \frac{1}{\sigma_2^2} \frac{Cov^2(X_2, X_1)}{\sigma_1^2}$ soit :

  $\boxed{VY_2 = 1 - \rho_{1,2}^2}$, en notant $\rho_{1,2}$ le coefficient de corrélation entre X_1 et X_2.

- $VY_3 = 1 - \frac{1}{\sigma_3^2}\left[\frac{Cov^2(X_3, Y_1)}{\underbrace{VY_1}_{=1}} + \frac{Cov^2(X_3, Y_2)}{VY_2} \right]$

$$= 1 - \frac{1}{\sigma_3^2}\left[\frac{Cov^2(X_3, X_1)}{\sigma_1^2} + \frac{Cov^2(X_3, Y_2)}{VY_2}\right]$$

$$= 1 - \rho_{1,3}^2 - \frac{1}{\sigma_3^2}\frac{Cov^2(X_3, Y_2)}{VY_2}.$$

Or : $Cov(X_3, Y_2) = \frac{1}{\sigma_2}\left[Cov(X_3, X_2) - \frac{Cov(X_2, Y_1)}{\underbrace{VY_1}_{=1}} Cov(X_3, Y_1)\right]$

$$= \frac{1}{\sigma_2}\left[Cov(X_3, X_2) - \frac{Cov(X_2, X_1)}{\sigma_1^2} Cov(X_3, X_1)\right]$$

$$= \frac{1}{\sigma_2}\left[\rho_{2,3}\sigma_2\sigma_3 - \frac{\rho_{1,2}\sigma_1\sigma_2}{\sigma_1^2}\rho_{1,3}\sigma_1\sigma_3\right]$$

$$= \sigma_3(\rho_{2,3} - \rho_{1,2}\rho_{1,3}).$$

D'où :

$$\boxed{VY_3 = 1 - \rho_{1,3}^2 - \frac{(\rho_{2,3} - \rho_{1,2}\rho_{1,3})^2}{1 - \rho_{1,2}^2}}.$$

On peut alors calculer une **inertie totale** :

$$\boxed{I = \sum_{k=1}^{K} VY_k}.$$

Il est facile de voir, à partir de l'égalité **(1)**, que : $\boxed{I \leq K = \sum_{k=1}^{K} X'_k}$.

ANNEXE 2

Comparaison des structures de deux partitions.

Univers U de taille N, deux partitions P_1 et P_2.

Pour tout $i \in U$, on note : $C_1(i) =$ ensemble des éléments de U classés dans la même partie de la partition P_1 que i. Idem pour $C_2(i)$.

On peut alors construire un *indice de similarité des deux partitions* :

$$I = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} 1_{j \in C_1(i)} 1_{j \notin C_2(i)} + \sum_{k=1}^{N} 1_{k \in C_2(i)} 1_{k \notin C_1(i)} \right).$$

En d'autres termes, pour chaque $i \in U$, on compte le nombre d'éléments de U classés avec i dans la partition P_1 mais pas dans la partition P_2 et on symétrise en permutant les deux permutations, puis on fait la somme sur tous les $i \in U$.

Il est clair que, si les deux partitions sont identiques, l'indice est nul.

L'indice peut se calculer quel que soit le système de désignation ou de numérotation des parties des deux partitions et peut même avoir du sens si les deux partitions n'ont pas le même nombre de parties.

Généralisation

Si les individus de l'univers se distinguent les uns des autres par des « importances » différentes, mesurées par des poids α_i, on peut donner une formule plus générale :

$$I = \sum_{i=1}^{N} \alpha_i \left(\sum_{j=1}^{N} \alpha_j 1_{j \in C_1(i)} 1_{j \notin C_2(i)} + \sum_{k=1}^{N} \alpha_k 1_{k \in C_2(i)} 1_{k \notin C_1(i)} \right).$$

Variante :

On raisonne en termes de *proportion relative* d'éléments de U non classés avec i dans la partition P_2 par rapport à ceux classés avec i dans la partition P_1 et vice-versa.

$$I^* = \sum_{i=1}^{N} \alpha_i \left(\frac{\sum_{j=1}^{N} \alpha_j 1_{j \in C_1(i)} 1_{j \notin C_2(i)}}{\sum_{j=1}^{N} \alpha_j 1_{j \in C_1(i)}} + \frac{\sum_{k=1}^{N} \alpha_k 1_{k \in C_2(i)} 1_{k \notin C_1(i)}}{\sum_{k=1}^{N} \alpha_k 1_{k \in C_2(i)}} \right)$$

$$= \sum_{i=1}^{N} \alpha_i \left(\frac{\sum_{j \in C_1(i)} \alpha_j 1_{j \notin C_2(i)}}{\sum_{j \in C_1(i)} \alpha_j} + \frac{\sum_{k \in C_2(i)} \alpha_k 1_{k \notin C_1(i)}}{\sum_{k \in C_2(i)} \alpha_k} \right)$$

ANNEXE 3

Extension de la méthode : cas de statistiques non linéaires

Il peut être intéressant de comparer deux agrégats statistiques non plus à partir de la considération de la variation d'inertie résultant de leur agrégation mais à partir de statistiques non linéaires. Un exemple théorique est traité ici avec les **quantiles**.

Supposons que l'on compare deux agrégats statistiques selon la distribution de revenus. L'indicateur le plus simple est la ***médiane***. Ces deux agrégats seront considérés comme proches si les revenus médians de ces deux agrégats sont proches. On peut donc construire un algorithme, de principe similaire à celui décrit dans les sections précédentes, agrégeant à chaque étape les unités dont les revenus médians sont les plus proches (si l'on vise l'homogénéité) ou les plus distants (si l'on vise l'hétérogénéité). L'unité résultant d'une agrégation se verra ensuite affecter le revenu médian résultant.

L'une des difficultés qui se présente est que, si l'on ne connaît pas les valeurs du revenu sur les unités statistiques élémentaires constituant les agrégats considérés, on ne peut pas calculer exactement le revenu médian résultant.

On propose alors **l'estimation paramétrique suivante** :

Considérons deux agrégats de base i et j, sur chacun desquels est définie une variable numérique R (respectivement R_i et R_j) s'interprétant comme la *médiane* de quantités définies sur les unités statistiques élémentaires constituant ces agrégats. Le problème est d'estimer la médiane R_{i+j} relative à l'agrégat composite issu de l'agrégation de i et de j, *sans connaître les données relatives aux unités statistiques élémentaires*.

Pour cela, on va faire l'hypothèse que, dans chacun des agrégats, **les données relatives aux unités statistiques élémentaires suivent une distribution lognormale.**

Rappel :

La densité d'une distribution lognormale est :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(Ln\,x - m)^2]\frac{1}{x}1_{IR^{+*}}(x)\,.$$

Son espérance est : $e^{m+\frac{\sigma^2}{2}}$ et sa variance : $e^{2m}(e^{2\sigma^2} - e^{\sigma^2})$.

Sa fonction de répartition vaut : $F(x) = H\left(\frac{Ln\,x - m}{\sigma}\right)$ pour $x > 0$, où H est la fonction de répartition de la loi ***N*** (0, 1).

Ses quantiles d'ordre α sont : $q_\alpha = e^{m+\sigma H^{-1}(\alpha)}$, d'où la médiane : $\boxed{q_{1/2} = e^{m+\sigma H^{-1}(1/2)} = e^m}$.

On notera que, si l'on connaît la médiane et, par exemple, un quartile de la distribution, on peut en déterminer les paramètres :

$$\begin{cases} m = Ln\ q_{1/2} \\ \sigma = \dfrac{1}{H^{-1}(1/4)} Ln\ (\dfrac{q_{1/4}}{q_{1/2}}) = \dfrac{1}{H^{-1}(3/4)} Ln\ (\dfrac{q_{3/4}}{q_{1/2}}) \end{cases}.$$

On obtient alors l'espérance de la loi : $e^{m+\frac{\sigma^2}{2}} = q_{1/2} \exp \frac{1}{2}\left[\frac{1}{[H^{-1}(3/4)]^2} Ln^2 (\frac{q_{3/4}}{q_{1/2}}) \right]$.

- 1ère approche :

  Lorsqu'on agrège deux agrégats de base i_1 et i_2, d'effectifs respectifs n_1 et n_2 (en nombre d'unités statistiques élémentaires), la distribution des données élémentaires sera un « *mélange* » des deux distributions d'origine, c'est-à-dire une loi de probabilité admettant pour densité $\sum_{l=1}^{2} \alpha_l f_l$, avec : $\alpha_l = \dfrac{n_l}{n_1 + n_2}$ et les f_l sont les densités des lois lognormales relatives à chacun des deux agrégats de base.

  La fonction de répartition de cette loi sera : $\sum_{l=1}^{2} \alpha_l F_l$, où les F_l sont les fonctions de répartition associées aux densités f_l.

  La médiane de cette loi est le réel μ défini par : $\sum_{l=1}^{2} \alpha_l F_l(\mu) = \dfrac{1}{2}$, soit :

  $$\sum_{l=1}^{2} n_l F_l(\mu) = \frac{n_1 + n_2}{2}.$$

  En remplaçant les F_l par leurs expressions explicites (en fonction de paramètres m_l et σ_l calculés à partir de deux quantiles des lois correspondantes), on peut mettre cette équation sous l'une ou l'autre des deux formes suivantes :

  $$\sum_{l=1}^{2} n_l H\left(\frac{Ln\mu - m_l}{\sigma_l} \right) = \frac{n_1 + n_2}{2} \quad \text{ou :} \quad \sum_{l=1}^{2} n_l H\left(\frac{Ln\dfrac{\mu}{\mu_l}}{\sigma_l} \right) = \frac{n_1 + n_2}{2},$$

  en notant μ_l la médiane de la loi de fonction de répartition F_l : $\mu_l = e^{m_l}$.

  On peut aussi écrire :

  $$\sum_{l=1}^{2} n_l \int_0^{\mu} \frac{1}{\sigma_l \sqrt{2\pi}} \exp[-\frac{1}{2\sigma_l^2} (Ln\, x - m_l)^2] \frac{1}{x} dx = \frac{n_1 + n_2}{2},$$

  soit :

  $$\int_0^{\mu} \frac{1}{x} \left[\sum_{l=1}^{2} \frac{n_l}{\sigma_l} \exp[-\frac{1}{2\sigma_l^2} (Ln\, x - m_l)^2] \right] dx = \sqrt{\frac{\pi}{2}} (n_1 + n_2)$$

$$\boxed{\int_0^{\mu} \frac{1}{x}\left[\sum_{l=1}^{2} \frac{n_l}{\sigma_l} \exp[-\frac{1}{2\sigma_l^2}(Ln\frac{x}{\mu_l})^2]\right] dx = \sqrt{\frac{\pi}{2}}(n_1+n_2).}$$

- 2ème approche :

L'approche précédente, outre les problèmes de calcul numérique qu'elle pose ainsi que la nécessité d'estimer les σ_l, présente une difficulté vis-à-vis de l'itération de la procédure d'agrégation. En effet, si l'on peut faire l'hypothèse que la distribution des données élémentaires est lognormale au sein de chaque agrégat de base, cela ne sera plus vrai dès lors qu'on aura construit des agrégats composites.

Pour itérer la procédure, il faut se mettre dans une situation où la distribution des données élémentaires est lognormale au sein de chaque agrégat composite construit.

Pour cela, on va approximer la loi mélange relative à une agrégation de deux agrégats de base par une loi lognormale et remplacer la vraie loi par cette approximation lognormale.

Le critère d'approximation retenu sera celui de l'information de KULLBACK.

Rappel :

L'écart entre une loi de probabilité P et une loi de probabilité P_0 servant de référence est : $I_{P/P_0} = \int -Ln(\frac{f}{f_0})\, f_0\, d\nu$, où f et f_0 sont les densités respectives des lois P et P_0 par rapport à une même mesure ν.

Ici : P_0 est la loi mélange et P la loi lognormale cherchée, définie par des paramètres m et σ.

Il s'agit donc de déterminer les paramètres m et σ minimisant

$$I_{P/P_0} = \int -Ln(\frac{f}{f_0})\, f_0\, d\nu = \int -(Ln\ f)\, f_0\, d\nu + \int (Ln\ f_0)\, f_0\, d\nu .$$

Le problème équivaut à : $Max \int (Ln\ f)\, f_0\, d\nu$.

Ici : $f_0 = \sum_{l=1}^{2} \alpha_l f_l$ et : $Ln\, f(x) = -Ln\,\sigma - Ln\sqrt{2\pi} - \frac{1}{2\sigma^2}(Ln\,x - m)^2 - Ln\,x$ pour $x > 0$.

On a donc la succession de problèmes équivalents suivants :

$$Max \int_0^{+\infty} [-Ln\,\sigma - Ln\sqrt{2\pi} - \frac{1}{2\sigma^2}(Ln\,x - m)^2 - Ln\,x][\sum_{l=1}^{2} \alpha_l f_l(x)]\, dx$$

$$\Leftrightarrow Min \sum_{l=1}^{2} \alpha_l \left[\int_0^{+\infty} [Ln\,\sigma + Ln\sqrt{2\pi} + \frac{1}{2\sigma^2}(Ln\,x - m)^2 + Ln\,x]\, f_l(x)\, dx\right]$$

$$\Leftrightarrow Min \sum_{l=1}^{2} \alpha_l \left[\int_0^{+\infty} [Ln\,\sigma + \frac{1}{2\sigma^2}(Ln\,x - m)^2]\, f_l(x)\, dx\right]$$

[élimination des termes ne dépendant pas des paramètres]

$$\Leftrightarrow Min\left[Ln\,\sigma \sum_{l=1}^{2} \alpha_l \underbrace{\int_0^{+\infty} f_l(x)\,dx}_{=1} + \sum_{l=1}^{2} \alpha_l \int_0^{+\infty} \frac{1}{2\sigma^2} (Ln\,x - m)^2 \; f_l(x)\,dx \right]$$

$$\Leftrightarrow Min\left[Ln\,\sigma + \frac{1}{2\sigma^2} \sum_{l=1}^{2} \alpha_l \int_0^{+\infty} (Ln\,x - m)^2 \; f_l(x)\,dx \right].$$

L'intégrale $\int_0^{+\infty} (Ln\,x - m)^2 \; f_l(x)\,dx$ s'interprète comme l'espérance de $(Ln\,X - m)^2$, où X est une variable aléatoire suivant une loi lognormale de densité f_l. Mais, dans ce cas : $Ln\,X$ ~> ***N*** (m_l, σ_l^2), d'où :

$$E(Ln\,X - m)^2 = V(Ln\,X) + [E(Ln\,X) - m]^2 = \sigma_l^2 + (m_l - m)^2 .$$

Finalement, le problème équivaut à :

$$\boxed{Min\left[Ln\,\sigma + \frac{1}{2\sigma^2} \sum_{l=1}^{2} \alpha_l [\sigma_l^2 + (m_l - m)^2] \right]}.$$

On écrit les conditions du 1er ordre :

$$\frac{\partial}{\partial m}\left[Ln\,\sigma + \frac{1}{2\sigma^2} \sum_{l=1}^{2} \alpha_l [\sigma_l^2 + (m_l - m)^2] \right] = \frac{1}{2\sigma^2} \sum_{l=1}^{2} \alpha_l\, 2(m - m_l) ,$$

la solution à l'optimum m^* vérifie donc : $\frac{1}{\sigma^2} \sum_{l=1}^{2} \alpha_l (m^* - m_l) = 0$, d'où : $\boxed{m^* = \sum_{l=1}^{2} \alpha_l m_l}$.

Puis : $\frac{\partial}{\partial \sigma}\left[Ln\,\sigma + \frac{1}{2\sigma^2} \sum_{l=1}^{2} \alpha_l [\sigma_l^2 + (m_l - m)^2] \right] = \frac{1}{\sigma} - \frac{1}{\sigma^3} \sum_{l=1}^{2} \alpha_l [\sigma_l^2 + (m_l - m)^2]$.

La solution à l'optimum σ^* vérifie donc :

$$\boxed{\sigma^{*2} = \sum_{l=1}^{2} \alpha_l [\sigma_l^2 + (m_l - m^*)^2]}.$$

Notons enfin que :

$$\begin{cases} m_1 - m^* = m_1 - (\alpha_1 m_1 + \alpha_2 m_2) = \underbrace{(1 - \alpha_1)}_{=\alpha_2} m_1 - \alpha_2 m_2 = \alpha_2 (m_1 - m_2) \\ m_2 - m^* = \alpha_1 (m_2 - m_1) \end{cases},$$

d'où :

$$\boxed{\sigma^{*2} = \sum_{l=1}^{2} \alpha_l \sigma_l^2 + 2\alpha_1\alpha_2 (m_1 - m_2)^2}.$$

Une fois qu'on a déterminé les paramètres de la loi lognormale approximante, on peut prendre comme médiane de l'agrégat composite celle de cette loi lognormale obtenue, soit : $\mu^* = e^{m^*} = e^{\sum_{l=1}^{2} \alpha_l m_l} = e^{\alpha_1 m_1} e^{\alpha_2 m_2} = (e^{m_1})^{\alpha_1} (e^{m_2})^{\alpha_2}$, soit :

$$\boxed{\mu^* = \mu_1^{\alpha_1} \mu_2^{\alpha_2}}.$$

On obtient un résultat particulièrement simple et élégant, s'interprétant comme un moyenne höldérienne des médianes des agrégats de base, avec des poids égaux aux poids relatifs, en termes de nombres d'unités statistiques élémentaires, de ces agrégats. On n'a d'ailleurs pas besoin d'estimer σ^{*2} ni de connaître les σ_l.

Généralisation possible :

Comparer deux agrégats au vu de la distribution des revenus, résumée par la considération d'un nombre prédéfini de quantiles : prendre par exemple une distance euclidienne entre quantiles.