TAUX DE CHOMAGE ET TYPOLOGIE DE ZONES D'EMPLOI

B. GELEIN-DOUKKALI
INSEE-POITOU-CHARENTES, Service Études et Diffusion

1. Introduction

1.1 Objectif de l'étude

L'objectif de l'étude consiste en la réalisation d'une typologie de zones d'emploi construite sur une distinction véritable entre la variable à expliquer - le taux de chômage - et un ensemble de variables explicatives.

1.2 Des méthodes encore peu utilisées à l'INSEE

Un certain nombre de travaux ont déjà eu pour objectif la réalisation d'une typologie de zones ou de bassins d'emploi parmi lesquels :
- « Les cousins des bassins d'emploi bas-normands », Cent pour Cent, Insee basse Normandie, 1999.
- « Le chômage par zones d'emploi en France », INSEE Première, n° 577, 1998.
- « Typologie des zones d'emploi sensibles aux risques de chômage »,
Les Dossiers de la DARES, numéros 3-4, La Documentation Française, 1997.

La classification ascendante hiérarchique est la technique la plus couramment utilisée. Pourtant, elle relève du seul domaine du descriptif. Elle ne permet pas de privilégier une variable - à expliquer - par rapport aux autres variables -explicatives. D'où la nécessité de recourir aux techniques de **segmentation** (analyse discriminante non paramétrique ou régression non paramétrique) qui allient **l'explication et la recherche d'une partition.**

Par ailleurs, l'utilisation d'une **analyse factorielle multiple,** en amont de la segmentation, **permet[1] de regrouper les variables actives par thème tout en équilibrant l'influence de ces différents thèmes**. Ce traitement simultané de plusieurs thèmes semble particulièrement adapté à l'étude d'un phénomène à causes multiples tel que le chômage.

1.3 Enchaînement des méthodes

Cet enchaînement s'inspire de celui d'une analyse factorielle simple[2] et d'une classification.

- Sélection et recodage des variables explicatives continues afin de tenir compte d'éventuelles relations non linéaires : obtention de variables explicatives qualitatives.

- Analyse factorielle multiple sur les variables recodées et classées par thème.

- Segmentation (régression non paramétrique):
  -variable à expliquer : le taux de chômage (variable continue)
  -variables explicatives : les composantes principales en sortie de l'AFM

- Interprétation de la typologie obtenue par la segmentation grâce à un retour aux variables initiales : comparaison de moyennes et de pourcentages.

[1] Contrairement aux analyses factorielles simples (Analyse des Correspondances Multiples ou Analyse en Composantes Principales)
[2] Analyse des correspondances multiples ou Analyse en composantes principales

2. Sélection et recodage des variables

Les variables utilisées ont été sélectionnées parmi celles présentes dans l'Atlas des zones d'emploi et complétées par des données issues des DADS. Le taux de chômage -variable à expliquer - est le taux de chômage annuel moyen pour 1997, calculé au sens du BIT. Les variables explicatives[3] ont été regroupées en 9 thèmes :

THEME 1 : Répartition sectorielle
THEME 2 : Taille des établissements
THEME 3 : Gestion, ressources humaines
THEME 4 : Démographie des établissements
THEME 5 : Niveau de formation
THEME 6 : Profil des chômeurs
THEME 7 : Population, démographie
THEME 8 : Population active
THEME 9 : Revenu, conditions de vie

Seuls les huit premiers thèmes ont joué un rôle actif dans l'analyse. Le thème 9, traité en élément illustratif, a permis d'enrichir l'interprétation de la typologie.

Les variables continues initiales ont été discrétisées afin d'appliquer une AFM sur variables qualitatives, et ce, pour deux raisons :

- mise en évidence de liaisons non linéaires entre variables
- réduction de l'influence des valeurs extrêmes

Ce codage présente 2 inconvénients. Pour chaque variable discrétisée, on a :

- une perte d'information précise sur la variable
- une introduction artificielle d'une distance entre deux individus dont les valeurs tombent de part et d'autre de la frontière.

TRAVAUX FUTURS ENVISAGÉS
Les techniques de ***codage flou*** permettant de pallier ces deux inconvénients, une macro SAS mettant en oeuvre la technique du codage barycentrique pourrait être élaborée.

Un codage automatique reposant sur le calcul des quartiles ne tient pas compte des véritables ruptures dans la distribution d'une variable. C'est pourquoi, la détermination des limites de classes s'est faite après examen des histogrammes de chaque variable, et donc de façon non automatique. L'interactivité du module SAS/INSIGHT a permis d'accélérer cette phase (utilisation des graphiques dynamiques).

[3] La liste de ces variables est donnée en annexe.

3. L'analyse factorielle multiple

L'analyse factorielle multiple s'applique aux tableaux dans lesquels les individus sont décrits par plusieurs groupes de variables. Ces derniers correspondent, pour la présente étude, aux 9 thèmes déjà évoqués auxquels s'ajoute un groupe contenant une seule variable - le taux de chômage. Ce dernier a été traité en élément supplémentaire puisque cette étape a pour but **la création et la sélection de variables synthétiques explicatives.**

3.1 Avantages de cette méthode

Le grand intérêt de l'AFM est de pouvoir **traiter simultanément différents groupes de variables en équilibrant leur influence**. Ce point est important dans la mesure où un groupe de variables peut jouer un rôle prépondérant dans une analyse globale traditionnelle pour 2 raisons :

- le nombre de variables du groupe (plus ce nombre est élevé, plus forte est l'influence du groupe)
- la structure du groupe (plus ses variables sont liées, plus son influence sera grande dans la détermination des axes principaux de l'analyse globale).

Par ailleurs, l'analyse factorielle classique permettrait de réaliser des analyses séparées des groupes (exemple de l'article « Les cousins des bassins d'emploi bas-normands »). Mais leurs résultats, obtenus indépendamment, sont difficilement comparables : l'identité de sous-espaces factoriels peut être masquée par des rotations. Une véritable comparaison de ces typologies séparées nécessite **la construction d'un référentiel commun**. L'AFM répond à ce besoin en offrant une représentation factorielle où les observations (ici les zones d'emploi) sont décrites par chaque groupe de variables séparément. Ainsi, on obtient la projection sur un même plan des 9 images d'une même zone d'emploi observée, d'une part, au travers de chacun des 8 thèmes actifs[4], et, d'autre part, du point de vue de l'ensemble des variables actives. On peut donc savoir si une zone d'emploi se distingue des autres globalement ou bien sur un thème particulier.

De plus, l'étude est enrichie (par rapport à une analyse traditionnelle) par la possibilité de comparer les groupes de variables eux-mêmes. On est alors en mesure **de voir si le taux de chômage est davantage lié à tel thème qu'à tel autre**.

[4] Les aspects théoriques de l'AFM, développés en annexe, expliquent pourquoi seuls les individus partiels correspondant aux groupes actifs sont représentés.

L'AFM peut traiter en même temps des groupes de variables qualitatives et des groupes de variables quantitatives. Par contre, au sein d'un même groupe, les variables doivent être de même nature.

3.2 Le principe

L'AFM se comporte comme une ACP pour les variables quantitatives et comme une ACM pour les variables qualitatives.

L'AFM équilibre l'influence des groupes de variables en donnant à chaque variable un poids - identique au sein de chaque groupe afin de conserver la structure interne de chacun des groupes. Le poids donné par l'AFM à chacune des variables d'un groupe est égal à l'inverse de l'inertie du premier axe principal de ce groupe (obtenue par des ACM ou ACP séparées). On a donc une normalisation des nuages représentant les individus pour chaque groupe.

Cette pondération rend possible la présence simultanée de variables continues et nominales parmi les éléments actifs. Elle autorise également la présence simultanée de variables continues centrées réduites et de variables continues non réduites.

3.3 Résultats en sortie de l'AFM

On obtient une table avec comme observations les 348 zones d'emploi et comme variables une sélection des premiers axes principaux. On retient les variables synthétiques qui donnent le plus d'information sur la distribution du nuage global initial et sont les mieux corrélées avec la variable à expliquer. Cette table sera utilisée en entrée de la segmentation.

L'interprétation des listings permet notamment :

- d'identifier les thèmes les plus liés au niveau du taux de chômage,

- de repérer les proximités entre zones d'emploi,

- de distinguer les zones globalement atypiques,

- dans le cadre d'une étude régionale, de pointer les thèmes qui mettent en exergue les particularités de certaines zones.

Une exploitation complète des résultats - très riches - de l'AFM offre donc matière à une étude à deux niveaux géographiques : national et régional.

4 La segmentation

La segmentation permet de résoudre les problèmes de discrimination (variable à expliquer nominale) ou de régression (variable à expliquer quantitative) en segmentant de façon progressive un échantillon pour obtenir un **arbre de décision binaire**. Elle conduit, après élagage de cet arbre[5], à la **création d'une partition des zones d'emploi.**

4.1 Avantages de cette méthode :

Tout comme la classification, la segmentation aboutit à l'élaboration d'une typologie mais avec ceci d'essentiel : la distinction entre variable expliquée et variables explicatives. La segmentation permet donc en plus de **faire de la prévision**.

Ce pouvoir prédictif est d'un intérêt majeur pour les décideurs. En effet, il permet d'aider à déterminer quelle action mettre en oeuvre pour contribuer à la baisse du taux de chômage (agir par exemple sur la part des non diplômés dans la population active). La segmentation offre donc **un véritable tableau de bord.**

Elle fait concurrence aux méthodes plus classiques (régression multiple, analyse discriminante, régression logistique).

Elle présente en effet les avantages suivants :

- lisibilité des règles d'affectation,
- la technique étant non paramétrique, on est peu contraint par la nature des données,
- on obtient d'office la sélection des variables à utiliser en tenant compte d'éventuelles interactions,
- robustesse vis-à-vis de données erronées ou de valeurs aberrantes,
- gestion des données manquantes.

On ne peut néanmoins faire l'économie d'une comparaison des résultats de la segmentation avec ceux obtenus par une simple régression ou une discrimination.

4.2 Principe :

[5] Ce qui correspond, en classification, au choix d'une partition par coupure de l'arbre.

La méthode - divisive et descendante - partitionne itérativement la population.

On recherche parmi les variables explicatives (ici les axes principaux en sortie de l'AFM) celle qui est le plus corrélée avec la variable à expliquer (ici le taux de chômage). Cette variable explicative définit une première division de l'échantillon en deux parties - deux segments. Cette procédure est réitérée pour chacun de ces deux segments. A chaque étape du partitionnement, on définit des sous-populations les plus homogènes possibles vis-à-vis de la variable à expliquer. Dans le cas de la régression non paramétrique, on minimise donc la variance intra-classes.[6] On obtient ainsi un arbre binaire complet pour lequel chaque segment terminal contient un seul individu.

La méthode détermine ensuite un sous-arbre « optimal », après élagage de l'arbre binaire complet. Cette procédure minimise l'estimation de l'erreur théorique d'affectation d'une valeur du taux de chômage à une zone d'emploi. La minimisation de la variance intra-classes s'effectue soit à l'aide d'un échantillon test, soit par validation croisée.

4.3 Résultats en sortie de la segmentation :

Les zones d'emploi sont classées en différents groupes ou segments, définis par des règles de décision appliquées aux variables synthétiques retenues (axes principaux de l'AFM).

[6] Si la variable à expliquer est nominale et possède k modalités, on cherche à réduire le mélange de ces modalités au sein des segments.

5. Complémentarité entre AFM et segmentation

- ***Lissage des données*** :

L'utilisation des premiers facteurs principaux uniquement, en entrée de la segmentation, permet d'éliminer les fluctuations aléatoires qui souvent constituent l'essentiel de la variance recueillie dans les directions des derniers axes[7].

- ***Difficultés d'interprétation des axes des analyses factorielles :***

Un segment peut être typique d'un axe de rang élevé et aider à son interprétation. Il est plus facile de décrire des segments qu'un espace continu, même à deux dimensions :

-La notion de segment est accessible à l'intuition : il est défini à partir de règles de décision simples

-La description des segments peut s'effectue aussi à partir de simples comparaisons de moyennes ou de pourcentages.

-Il est possible de remplacer les nombreux individus par quelques centres de gravité de segments d'où une amélioration de la qualité de représentation.

- ***Robustesse imparfaite des analyses factorielles :***

La segmentation est moins sensible aux points marginaux isolés.

- ***Faculté descriptive des axes de l'AFM :***

La segmentation ne permet pas toujours de détecter l'importance de certaines tendances d'où l'utilité du positionnement des segments sur les axes factoriels.

- ***Détection d'effets croisés.***

La segmentation, de par son aspect séquentiel, tient compte d'effets croisés - ce que l'analyse factorielle ne fait pas.

[7] Variations non systématiques contenues dans les données.

6. Résultats globaux

Les disparités spatiales du taux de chômage et la décentralisation des politiques d'emploi plaident en faveur d'une analyse localisée des marchés du travail. Caractérisés par des flux entre emploi, chômage et inactivité, ces marchés ne présentent pas tous la même efficience dans le processus d'appariement entre offre et demande de travail. Les caractéristiques socio-économiques qui structurent le territoire expliquent ces différences. Le taux de chômage dépend de facteurs que l'on peut classer en trois grandes catégories :

- **les facteurs d'offre de travail**

(ce sont les hommes qui proposent aux entreprises leur force de travail)

- **les facteurs de demande de travail**

(ce sont les entreprises qui expriment un besoin en main d'œuvre)

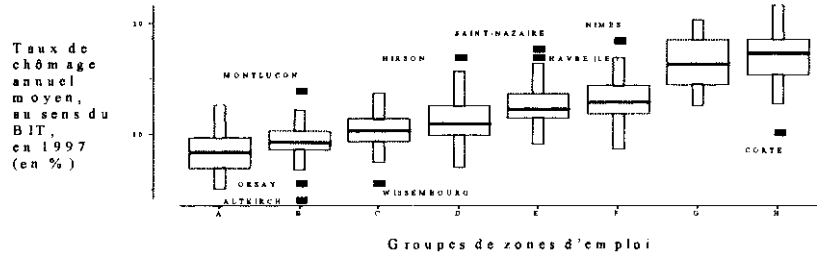
- **les facteurs d'environnement du marché du travail.**

Les zones d'emploi peuvent être regroupées en huit grands ensembles relativement homogènes tant au regard de ces critères, qu'à l'aune de ce que l'on cherche à expliquer : le taux de chômage.

Les facteurs ayant le plus contribué à expliquer le taux de chômage et à déterminer les huit groupes de zones d'emploi sont ceux relatifs :

- **au profil des chômeurs,**
- **au niveau de formation,**
- **à la démographie des entreprises,**
- **à la population active.**

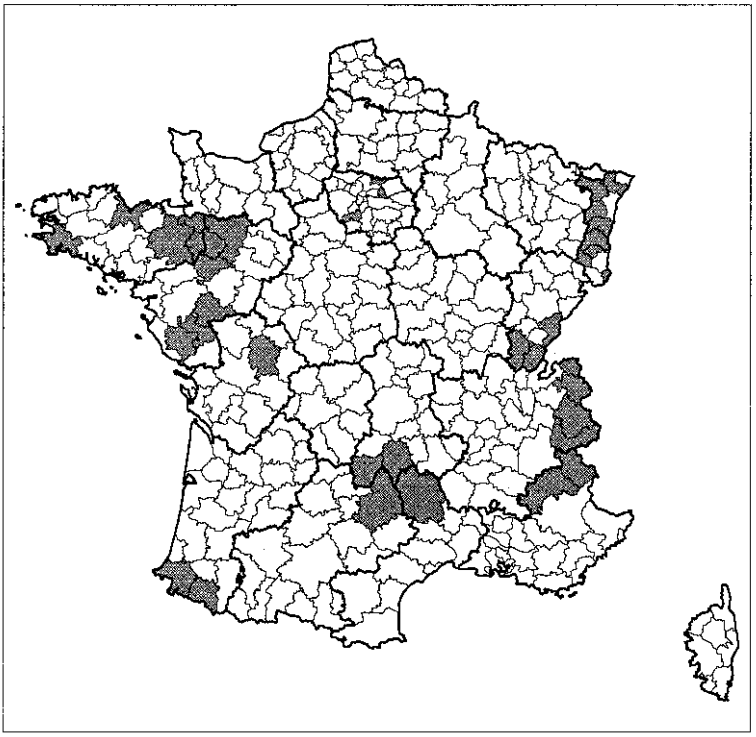
Le graphique ci-dessous permet de visualiser la distribution du taux de chômage dans chacun de ces groupes ou segments.



Cette simple série de box-plots permet d'avoir une idée de la ***différenciation des segments*** les uns par rapport aux autres, au regard du phénomène à expliquer. Il offre aussi une visualisation de la plus ou moins grande ***homogénéité*** de chaque segment vis-à-vis du taux de chômage. Certains segments présentent sensiblement un même taux de chômage, mais les **déterminants de ce taux de chômage diffèrent.**

- **Segment A : taux de chômage très inférieur à la moyenne nationale**

Un premier ensemble, constitué de 38 zones d'emploi, affiche un taux de chômage très inférieur à la moyenne nationale. Construit essentiellement sur la base de facteurs décrivant le profil des chômeurs, il comprend aussi bien des zones placées le long des frontières de l'Est de la France, que des zones d'Ile-de-France, Bretagne, Pays-de-la-Loire ou d'autres encore, situées plus au sud.

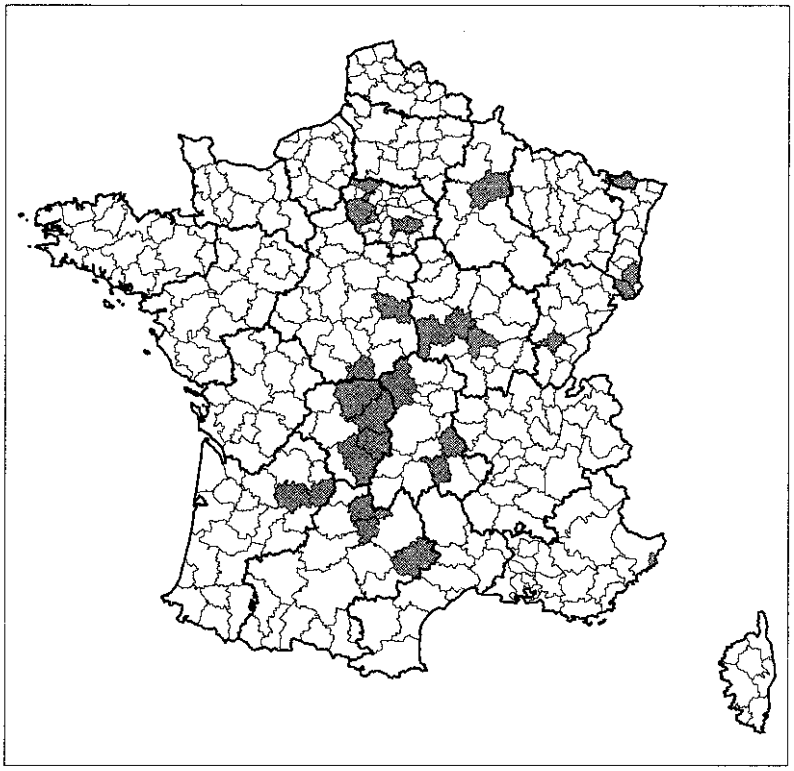


Segment A :

- Très bon taux de retour à l'emploi
- Peu de chômeurs de longue durée dans les DEFM.
- Sur-représentation des personnes ayant un CAP ou un BEP.
- Très bon taux de survie à 5 ans des reprises d'entreprises.
  Peu de bas niveaux de formation parmi les jeunes entrant au chômage.

- **Segment B : taux de chômage bien inférieur à la moyenne nationale**

Globalement, ce groupe de 27 zones connaît un faible taux de chômage. Alors qu'Orsay et Altkirch bénéficient d'un très faible taux de chômage, Montluçon enregistre, néanmoins, une valeur plutôt élevée sur cet indicateur. Les zones de cet ensemble présentent des similitudes non seulement sur le profil des chômeurs, mais aussi sur des données relevant de la démographie ou des ressources humaines. Elles sont principalement situées en Alsace, Ile-de France ou plus au centre du territoire métropolitain.

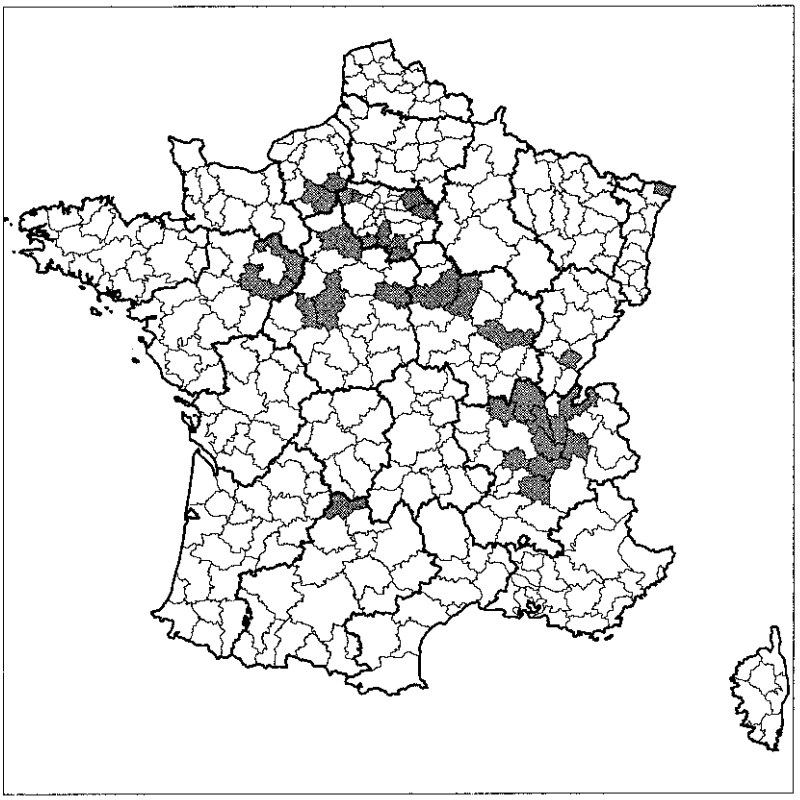


Segment B :

- Peu de créateurs d'entreprises individuelles de moins de 35 ans
- Forte proportion d'entrepreneurs individuels de 55 ans et plus.
- Peu de chômeurs de longue durée parmi les DEFM.
- Faible taux d'évolution annuel moyen de la population totale dû au solde naturel.

- **Segment C : taux de chômage inférieur au niveau national**

Les 32 zones de ce segment connaissent, en moyenne, un taux de chômage inférieur au niveau national. Situées notamment en Alsace et en Ile-de-France, ces zones présentent une certaine continuité géographique autour du Massif Central. La typicité de ces zones est essentiellement relative à la population active et au profil des demandeurs d'emploi. Le haut niveau du taux de dépendance à la région des entreprises joue également un rôle important. Les zones de ce groupe ne sont pas totalement homogènes : Wissembourg se distingue par un taux de chômage très faible par rapport aux autres zones de cet ensemble.

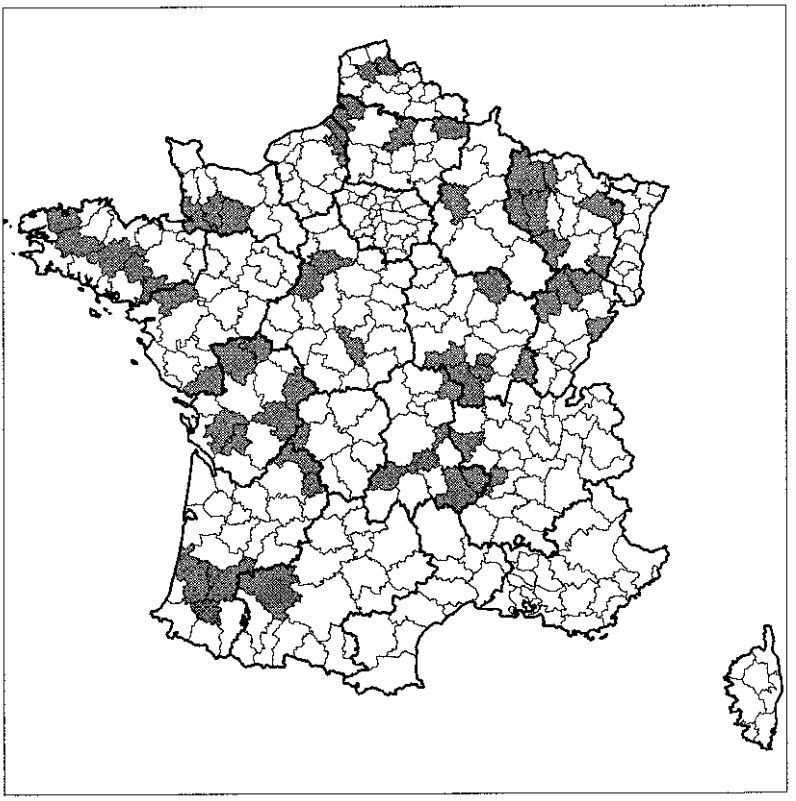


Segment C :

- Très fort taux d'activité des 15 ans et plus, et, en particulier, des 50-64 ans.
- Taux de dépendance des entreprises à la région, relativement élevé.
- Forte progression du nombre de DEFM, surtout pour les 50 ans et plus.

- **Segment D : taux de chômage peu inférieur au niveau national**

Réparties sur l'ensemble du territoire métropolitain, les 60 zones d'emploi de ce groupe affichent un taux de chômage en moyenne légèrement inférieur au niveau national. La zone d'Hirzon, néanmoins, connaît un taux de chômage élevé. Les zones de ce segment se distinguent sur de nombreux thèmes - la démographie d'entreprises et la répartition sectorielle en premier lieu.



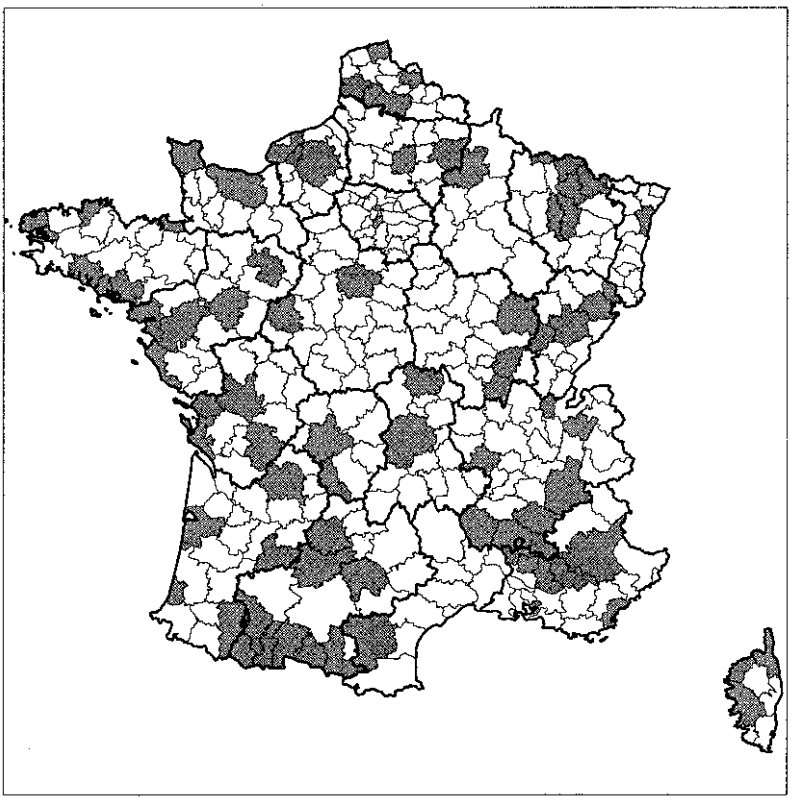
Segment D :

- Très peu de créations pures d'établissements mais des taux de survie parmi les plus élevés.
- Sous-représentation des établissements de moins de 5 ans.
- Très forte proportion de l'emploi dans l'industrie.
- Sous-représentation de l'emploi tertiaire.

- **Segment E : taux de chômage peu supérieur à la moyenne nationale**

Avec un taux de chômage légèrement supérieur à la moyenne nationale, cet ensemble de 81 zones (dont Paris) se répartit sur tout le territoire métropolitain. Il est particulièrement présent sur le littoral atlantique et autour de Toulouse. Il enserre également la zone de Gap. Deux zones affichent un taux de chômage bien plus élevé que les autres : Le Havre et Saint-Nazaire.

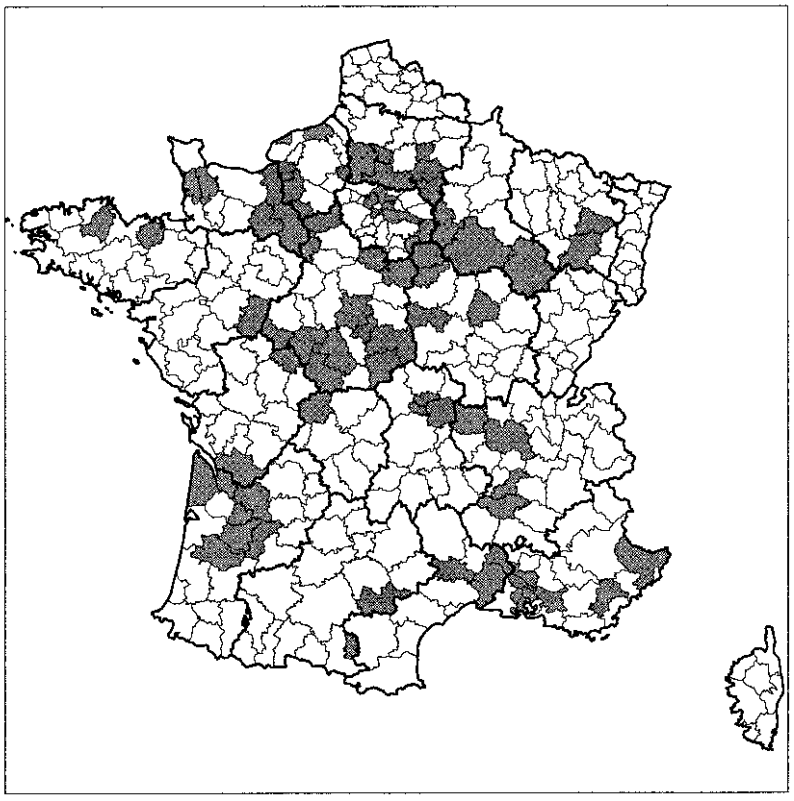
Ce sont surtout les indicateurs décrivant la répartition sectorielle, la population active et le niveau de formation qui permettent de caractériser ce groupe.



Segment E :

- Effectifs salariés importants dans le commerce de détail.
- Emploi tertiaire généralement sur-représenté contrairement à l'emploi industriel.
- Sur-représentation des employés et professions intermédiaires mais peu d'ouvriers.
- Bon niveau de formation.
- Peu de non-qualifiés parmi les DEFM
- **Segment F : taux de chômage supérieur à la moyenne nationale**

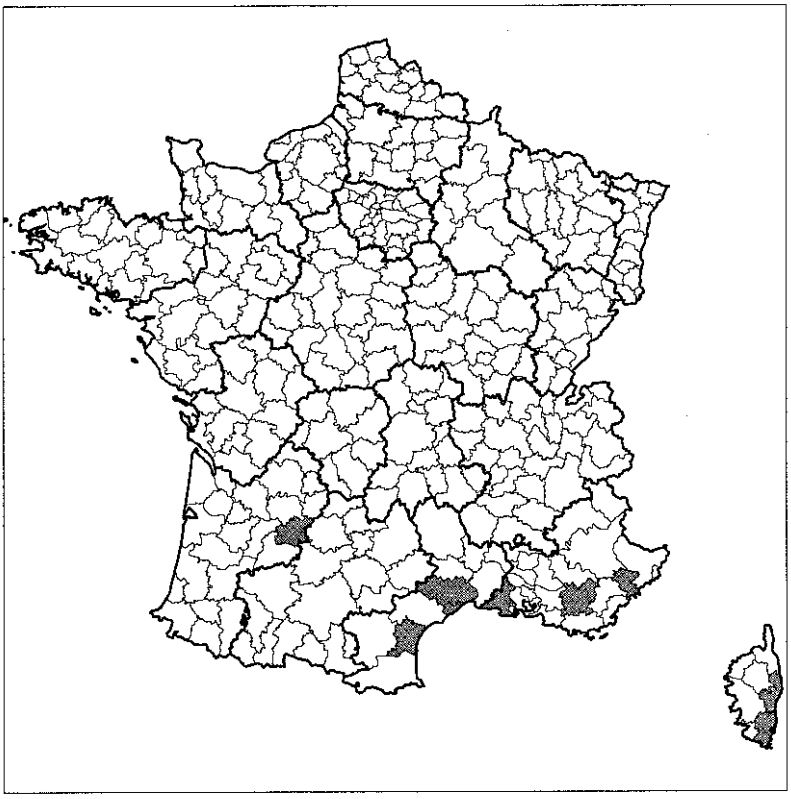
Ce groupe rassemble 78 zones d'emploi présentes dans de nombreuses régions. Elles forment une couronnes autour de Paris, de l'Ile de France et de Bordeaux. Elles constituent également un ensemble important dans la partie sud de la région Centre et au-dessus du littoral méditerranéen. Le taux de chômage y est supérieur à la moyenne nationale, et ce, surtout à Nîmes. La majorité des indicateurs caractérisant ce groupe de zones porte sur le profil des chômeurs.



Segment F :

- Forte proportion de chômeurs de longue durée parmi les DEFM.
- Très faible taux de retour à l'emploi.
- Sur-représentation des bas niveaux de formation parmi les jeunes entrant au chômage.
- Beaucoup de personnes de 50 ans et plus parmi les DEFM.
- Peu de créateurs d'entreprises individuelles de moins de 35 ans.
- De nombreux résidents travaillent dans une autre zone
- **Segment G : taux de chômage bien supérieur à la moyenne nationale**

Ce groupe de 9 zones traduit un phénomène essentiellement méditerranéen. Le taux de chômage y est très élevé. Il se distingue du segment H par, notamment, une faible taille des établissements et un solde migratoire plus fort.

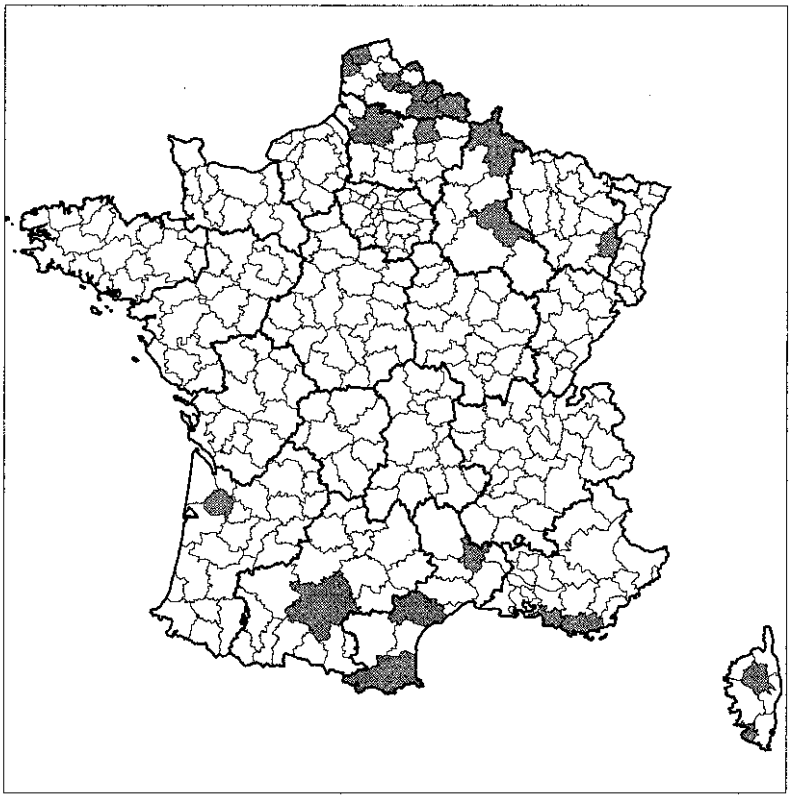


Segment G :

- Un taux d'évolution annuel moyen de la population totale, dû au solde migratoire, très élevé.
- De très nombreux établissements sans salarié.
- Faible taux d'activité des femmes.
- Sur-représentation des artisans-commerçants et chefs d'entreprises
- Bas niveaux de formation

- **Segment H : taux de chômage très supérieur à la moyenne nationale**

Ces 23 zones d'emploi révèlent des similitudes entre le nord et le sud de la France. En dehors de Corte, le taux de chômage y atteint un niveau bien supérieur à la moyenne nationale. Les trois thèmes sur lesquels ces zones se distinguent le plus du niveau national moyen sont : la population active, le profil des demandeurs d'emploi, la démographie.



Segment H :

- Faible taux d'activité des femmes et des 50-64 ans.
- Forte proportion de chômeurs de longue durée dans les DEFM.
- Bas niveau de formation très nombreux parmi les jeunes entrant au chômage.
- Sur-représentation des familles monoparentales

BIBLIOGRAPHIE

Escofier B., Pagès J. (1998), « Analyses factorielles simples et multiples », Dunod.

Groupe d'étude et de réflexion interrégional - GERI - (1996), « L'analyse des données évolutives : méthodes et applications », Technip.

Lavit C. (1988), « Analyse conjointe de tableaux quantitatifs », Masson.

Lebart L., Morineau A., Piron M. (1997), « Statistique exploratoire multidimensionnelle », Dunod.

Pour davantage de détails sur les résultats obtenus dans cette étude :

INSEE, Document de travail de la DDAR n° H004 (2000), « Journée régionale de méthodologie statistique du 06.12.00 ».

ANNEXE 1 - ASPECTS THÉORIQUES DE L'AFM

Cette annexe s'inspire essentiellement de l'ouvrage de B. Escofier et J. Pagès.

Les éléments théoriques portent, dans un premier temps, sur l'application de la méthode à des variables quantitatives. Sont ensuite traités les variables qualitatives et les éléments supplémentaires.

Les données portent sur I individus décrits par K variables formant J groupes. A chaque groupe correspond un tableau X_j, j = 1...J. Mais l'ensemble des J tableaux initiaux peut être traité comme un unique tableau structuré en sous-tableaux. En effet, les J tableaux portent sur les mêmes individus et peuvent, par juxtaposition, former un tableau croisant les I individus et les K variables.

1 - Notations

On note X le tableau complet, I l'ensemble des individus, K l'ensemble des variables, J l'ensemble des sous-tableaux, K_j l'ensemble des variables du groupe J. Les symboles I, J, K ou K_j désignent donc à la fois l'ensemble et son cardinal.
On considère les trois espaces suivants :

R^K dans lequel est situé le nuage des individus.

R^I dans lequel est situé le nuage des variables

R^{I^2} dans lequel est situé le nuage des groupes de variables

Soit D la matrice diagonale des poids des individus. Elle a pour terme diagonal p_i, avec i =1...I et $\sum_i pi = 1$.

Soit M*j* la matrice diagonale des poids des variables du groupe j. Elle a pour terme diagonal m_k^j le poids de la variable v_k appartenant au groupe j. Les variables numériques ont souvent un poids égal à 1.

Soit M la matrice diagonale des poids de l'analyse globale. Elle a pour terme diagonal $m_k = m_k^j / \lambda_1^j$ pour k $\in$ K_j, en notant λ_1^j la première valeur propre de l'analyse factorielle séparée du groupe j. **Cette pondération permet d'équilibrer le rôle des groupes.**

2 - Les individus

Dans l'espace R^K, on cherche :

- une représentation du nuage N_I des individus caractérisés par les K variables,
- une représentation superposée des J nuages N_I^j (j=1...J) d'individus caractérisés chacun par un groupe de variables.

R^K peut se décomposer en somme directe de J sous-espaces orthogonaux deux à deux isomorphes aux espaces R^{K_j} engendrés chacun par un des J groupes de variables : $R^K = \oplus R^{K_j}$

Les coordonnées des points de N_I^j sont contenues dans le tableau X_j. Les coordonnées de ces points, dans R^K, sont contenues dans le tableau $\widetilde{X}_j$ de dimensions I et K : X_j complété par des zéros pour les variables n'appartenant pas au groupe j.

$$\widetilde{X}_j = \begin{array}{|c|c|c|} \hline 0 & X_j & 0 \\ \hline \end{array}$$

Dans R^K, on a N_I^* le nuage des centres de gravité, notés i^*, des J points représentant le même individu i dans les N_I^j (J individus partiels). N_I^* se déduit de N_I par une homothétie de rapport 1/J : il constitue un nuage moyen pour les N_I^j.

On effectue une ACP du tableau X, l'objectif étant de projeter le nuage N_I (et donc N_I^*) sur un sous-espace de faible dimension tel que le nuage projeté ressemble le plus possible au nuage initial N_I (respectivement N_I^*).

La représentation superposée des J nuages définis par chaque groupe de variables s'obtient par l'ACP du tableau X en traitant en éléments supplémentaires les individus partiels. On projette donc les nuages N_I^j sur les axes factoriels de N_I. Cette projection permet d'obtenir :

- une bonne représentation de chaque N_I^j.
- une ressemblance entre les nuages N_I^j (éviter que les symétries, rotations ou homothéties ne masquent les ressemblances).

3 - Les variables

Dans l'espace R^I, on veut représenter:

- les K variables initiales,
- les composantes principales obtenues par des ACP séparées de chacun des groupes.

La représentation des variables s'obtient par l'ACP du tableau X. Duale de l'image de N_I obtenue dans R^K, elle offre une aide à l'interprétation de la représentation du nuage des individus. Elle constitue également une représentation optimale des corrélations entre variables. Elle permet donc une comparaison fine des groupes de variables.

La pondération des groupes par $1/\lambda_1^j$ est telle que :

- aucun groupe ne détermine à lui seul le premier axe (sauf situation de symétrie).
- un groupe contribue à la détermination de nombreux axes s'il est composé de nombreuses variables indépendantes (il est de grande « dimensionalité »).

On obtient la représentation des composantes principales de chaque groupe en les introduisant en éléments supplémentaires dans l'ACP de X. On peut ainsi repérer des facteurs communs à plusieurs groupes de variables.

4 - Les groupes de variables

Dans R^{I^2}, on cherche à comparer globalement les groupes. Pour les représenter, on construit un nuage de J points, noté N_J. Dans R^{I^2}, un groupe j est représenté par la matrice de dimensions I et I : $X_j M_j X_j' D = W_j D$. Cette matrice de I^2 scalaires est considérée comme élément de l'espace vectoriel R^{I^2} de dimension I^2 et muni du produit scalaire :

$\langle W_j D , W_l D \rangle$ = trace($W_j D \; W_l D$) pour les éléments $W_j D$ et $W_l D$.

En faisant apparaître explicitement la pondération de l'AFM, on obtient une mesure de liaison entre les groupes j et l :

$$L(K_j, K_l) = \langle \frac{W_j D}{\lambda_1^j}, \frac{W_l D}{\lambda_1^l} \rangle$$

Pour comparer les groupes, on décrit les proximités entre les $W_j D$ en les projetant sur un espace de faible dimension de R^{I^2}. On cherche, dans R^{I^2}, un repère orthonormé qui ajuste au mieux les $W_j D$ et dont les axes sont des éléments symétriques de rang 1. Ces éléments sont de la forme $z_s z'_s D$ (les z_s sont les composantes principales issues de l'ACP de X). Ils s'obtiennent par maximisation de la somme des projections des $W_j D$ sur $z_s z'_s D$. Cette somme est égale à l'inertie dans R^I des K variables projetées sur z_s : $\sum_j \langle W_j D , z_s z'_s D \rangle$.

La suite des éléments recherchés est associée à celle des composantes principales z_s. L'orthonormalité des z_s dans R^I est équivalente à celle des $z_s z'_s D$ dans R^{I^2}. La coordonnée de $W_j D$ sur $z_s z'_s D$ est la contribution du groupe j à l'inertie de z_s.

5 - Variables qualitatives et tableaux mixtes

L'AFM s'applique aux tableaux disjonctifs complets en raison de l'équivalence existant entre l'ACM d'une part et l'ACP appliquée aux variables indicatrices pondérées de façon appropriée d'autre part :

- on considère, dans R^I, le nuage des indicatrices non centrées mais divisées par leur écart-type,
- on affecte à chaque indicatrice k le poids $(I - I_k)/I$.

L'ACP normée des indicatrices, réalisée en AFM, est telle que les projections des colonnes correspondent aux corrélations entre les indicatrices et les facteurs sur I. En ACM, elles représentent par contre les centres de gravité des classes d'individus (définies par les modalités). Certains programmes d'AFM fournissent ces deux représentations.

L'AFM permet aussi de traiter simultanément des tableaux de variables quantitatives et des tableaux de variables indicatrices.

6 - Eléments supplémentaires

Les individus, affectés d'un poids nul, peuvent intervenir en tant qu'éléments supplémentaires. Sont alors calculées les projections de leur représentant dans N_I^* et dans les N_I^j.

Un groupe de variables peut également être traité en élément supplémentaire. On effectue alors :

- la normalisation du nuage N_K^j (défini par le groupe j dans R^I), semblable à celle des autres groupes, de manière à pouvoir le comparer aux autres nuages.
- la projection des composantes principales du groupe afin de comparer la forme générale de N_K^j à celle du nuage moyen N_K et à celles des autres nuages aux autres groupes actifs
- la projection des $W_j D$. La coordonnée d'un élément supplémentaire $W_j D$ sur l'axe de rang s donne la mesure de la liaison entre le facteur s et le groupe j, à savoir l'inertie des variables du groupe j le long de la direction s (on ne l'interprète plus comme une contribution).

Il n'existe pas de représentation superposée des nuages d'individus associée à des groupes de variables supplémentaires : cela reviendrait à projeter un nuage N_I^j sur un axe de R^K orthogonal au sous-espace contenant N_I^j.

ANNEXE 2 - ASPECTS THÉORIQUES DE LA SEGMENTATION

Seule la régression multiple par arbre de prédiction binaire avec la méthode CART sera abordée : c'est la méthode utilisée dans l'élaboration de la typologie de zones d'emploi.

1 - Notations, vocabulaire

Le tableau de données contient une variable continue *y* à expliquer en fonction d'autres variables x_j j=1 à K. On peut utiliser simultanément des variables explicatives continues, ordinales et nominales.

La segmentation s'appuie sur la construction d'un arbre de décision binaire obtenu par des divisions successives de l'échantillon en deux sous-ensembles ou segments.
On appelle :

- segments intermédiaires ou nœuds, les segments engendrant deux segments descendants immédiats,
- segments terminaux, les segments non divisés,
- branche d'un segment t, l'ensemble des segments descendant de t - t non inclus,
- arbre binaire complet (noté A_{max}), l'arbre pour lequel chaque segment terminal ne contient qu'un seul individu,
- sous-arbre A, l'arbre obtenu par élagage d'une ou plusieurs branches de A_{max}.

La méthode CART donne, à partir de A_{max}, une séquence de sous-arbres obtenue par la suppression successive des branches les moins informatives en terme d'explication de y. Un sous-arbre optimal est sélectionné par minimisation de l'estimation de l'erreur théorique de prévision en utilisant un échantillon-test.

2 - Construction de l'arbre de décision binaire

Les étapes de l'algorithme sont :

1. Un seul segment contient l'ensemble des individus.
2. On cherche pour chaque variable explicative x_j la meilleure division d_j^* du segment, au sens de la variance résiduelle (critère de division). Parmi les

K divisions d_j^*, on choisit celle correspondant au **critère de division** (voir infra).

3. On réitère la procédure pour chacun des deux segments descendants.
4. La procédure s'arrête si tous les segments sont terminaux (segments constitués d'un seul individu ou bien de taille inférieure à un effectif donné).

3 - Les divisions possibles

Pour une variable x_j **continue** donnée, on observe toutes les divisions d_j^a possibles de la forme $x_j < a$ où a est une valeur contenue dans l'étendue de x_j. Le segment descendant de gauche contient les individus pour lesquels $x_j < a$. Le segment descendant de droite contient les individus pour lesquels $x_j \geq a$.

Une variable x_j **nominale** à deux modalités m_j^1 et m_j^2 fournit une seule division : un des segments descendants contient les individus vérifiant $x_j = m_j^1$ et le second segment descendant contient les individus vérifiant $x_j = m_j^2$. Une variable x_j **nominale à k modalités ordonnées** fournit k-1 divisions. Une variable x_j **nominale à k modalités non ordonnées** fournit $2^{k-1}-1$ divisions.

4 - Critère de division : la variance résiduelle

Pour toute division d_j^a d'un nœud t par une variable x_j, on calcule la moyenne pondérée des variances de y à l'intérieur de chacun des segments descendants t_g et t_d, soit la variance résiduelle du nœud t :

$$\text{var}\left(d_j^m, t\right) = \left(\frac{n_g}{n_t} s_g^2\right) + \left(\frac{n_d}{n_t} s_d^2\right)$$

Les segments t_g, t_d et t ont respectivement pour effectifs n_g, n_d et n_t. Les variances de la variable continue y à l'intérieur de t_g et t_d sont notées s_g^2 et s_d^2.

Pour chaque variable x_j, la division retenue d_j^* minimise la variance résiduelle :

$$\text{var}(d_j^*, t) = \min_{m \in d_j} \{\text{var}(d_j^m, t)\}$$

avec d_j l'ensemble des divisions du nœud t qu'il est possible de réaliser à partir de la variable x_j.

La division d^* finalement retenue pour le nœud t s'effectue à partir de la variable x_j telle que :

$$\text{var}(d^*, t) = \min_{j=1...K} \{\text{var}(d_j^*, t)\}$$

5 - Règle d'affectation

Pour un individu *i* dont on ne connaît pas la valeur prise sur *y*, on cherche à prévoir cette valeur y_i. La règle d'affectation s'obtient en faisant descendre l'individu *i* dans l'arbre. La valeur estimée de y_i est égale à la moyenne de *y* calculée dans le segment auquel *i* aura été affecté. L'écart-type sera égal à celui calculé sur ce segment.

6 - Erreur Apparente de Prévision

À chaque segment terminal t de l'arbre A, on associe l'erreur R_t :

$$R_t = \frac{n_t}{n} \times s_t^2$$

où n est le nombre total d'individus, n_t est le nombre d'individus du segment t, s_t^2 est la variance de la variable *y* à l'intérieur du segment t.

L'erreur apparente de prévision (EAP) associée à l'arbre A est :

$$EAP_{(A)} = \sum_{t \in A} R_t$$

Plus on divise, plus la variance résiduelle diminue. On a $EAP_{(A\max)} = 0$.

7 - Sélection du sous-arbre optimal

On cherche le sous-arbre vérifiant les 3 critères :
- le nombre de segments terminaux le plus faible possible,
- une erreur apparente de prévision minimale,
- une estimation correcte de l'erreur théorique.

Un nombre trop faible de segments terminaux entraîne une forte EAP qui sur-estime l'erreur théorique. A contrario, un nombre trop élevé est associé à une EAP faible mais sous-estimant l'erreur théorique.

8 - Échantillon-test

L'utilisation d'un **échantillon-test** permet de déterminer le sous-arbre optimal sans utiliser de tests statistiques pour définir une règle d'arrêt de la procédure de division. Elle fournit une estimation précise de l'erreur théorique de prévision ou de classement.
On divise l'échantillon de base en 2 parties : l'échantillon d'apprentissage et l'échantillon-test.

On construit $A_{\max}$ avec l'échantillon d'apprentissage. Par élagage, on sélectionne une séquence optimale de sous-arbres $\{A_H,\ldots,A_h,\ldots A_1\}$. A_H correspond à $A_{\max}$. A_h est le sous-arbre optimal ayant h segments terminaux. A_1 coïncide avec l'échantillon total. Soit S_h l'ensemble des sous-arbres de $A_{\max}$ ayant h segments terminaux, on a l'erreur apparente de A_h :

$$EA(A_h) = \min_{A \in S_h} \{EA(A)\}$$

Les individus de l'échantillon-test parcourent chaque sous-arbre de la séquence optimale $\{A_H,\ldots,A_h,\ldots A_1\}$. On retient le sous-arbre A^* qui présente la plus faible erreur théorique ***ET*** :

$$ET(A^*) = \min_{1 \le h \le H} \{ET(A_h)\}$$

Cette erreur théorique ***ET*** est estimée, pour chaque sous-arbre A, par une erreur théorique de prévision en calculant, sur l'échantillon-test :

$$\hat{ETP}(A) = \sum_{t \in A} \tilde{R}_t$$

avec $\widetilde{R}_t = \frac{\widetilde{n}_t}{\widetilde{n}} \times \widetilde{S}_t^2$, $\widetilde{n}$ est la taille de l'échantillon-test, $\widetilde{n}_t$ est le nombre d'individus de l'échantillon-test appartenant au segment t et $\widetilde{S}_t^2$ est la variance de y dans le segment t.

À noter, les règles de décision déterminées lors de la construction de $A_{\max}$ et donc de A_H, sont élaborées à partir de l'échantillon de base. Elles aboutiront donc généralement au classement de plusieurs éléments de l'échantillon-test dans un même segment terminal t de A_H (on aura alors $\widetilde{S}_t^2$ non nulle).

ANNEXE 3 - LISTE DES VARIABLES CLASSÉES PAR THÈME

GROUPE	LIBELLE DU GROUPE	VARIABLE	LIBELLE DE LA VARIABLE
1	Répartition sectorielle	DC	Effectifs salariés dans le commerce de détail au 31.12.95 pour 10 000 habitants
	Répartition sectorielle	CH	Capacité d'accueil des hôtels rapportée à la population (en nombre de chambre pour 100 habitants)
	Répartition sectorielle	CC	Capacité d'accueil des campings rapportée à la population (en nombre d'emplacements pour 100 habitants)
	Répartition sectorielle	SCR	Indicateur de spécificité locale dans les conseils et la recherche-développement au 31.12.95
	Répartition sectorielle	SOP	Indicateur de spécificité locale dans les services opérationnels au 31.12.95
	Répartition sectorielle	TER	Part de l'emploi tertiaire en 1995 (en %)
	Répartition sectorielle	INDU	Part de l'emploi industriel en 1995 (en %)
	Répartition sectorielle	BAT	Part de l'emploi dans le bâtiment en 1995 (en %)
2	Taille des établissements	P9	Part des établissements de 1 à 9 salariés au 01.01.97 (en %)
	Taille des établissements	P49	Part des établissements de 10 à 49 salariés au 01.01.97 (en %)
	Taille des établissements	PP50	Part des établissements de 50 salariés et plus au 01.01.97 (en %)
	Taille des établissements	P9I	Part des établissements industriels de 1 à 9 salariés au 01.01.97 (en %)
	Taille des établissements	CR9	Part des établissements de 1 à 9 salariés dans le total des créations pures sur la période 1993-1996 (en %)
	Taille des établissements	CRP10	Part des établissements de plus de 10 salariés dans le total des créations pures sur la période 1993-1996 (en %)
3	Gestion, ressources humaines	TDR	Taux de dépendance à la région au 01.01.97 (en %)
	Gestion, ressources humaines	PIM	Part des entrepreneurs individuels de moins de 35 ans au 01.01.97 (en %)
	Gestion, ressources humaines	PIP	Part des entrepreneurs individuels de 55 ans et plus au 01.01.97 (en %)
	Gestion, ressources humaines	PCRJI	Part des créateurs d'entreprises individuelles de moins de 35 ans sur la période 1993-1996 (en %)
	Gestion, ressources humaines	TCDI	Taux d'encadrement dans l'industrie au 31.12.95 (en %)
	Gestion, ressources humaines	TCDS	Taux d'encadrement dans les services aux entreprises au 31.12.95(en %)

GROUPE	LIBELLE DU GROUPE	VARIABLE	LIBELLE DE LA VARIABLE
4	Démographie des établissements	TCP	Taux moyen de création pure d'établissements sur la période 1993-1996 (en %)
	Démographie des établissements	TCR	Taux moyen de création d'établissements par reprise sur la période 1993-1996 (en %)
	Démographie des établissements	TSP	Taux de survie à 5 ans des créations pures d'entreprises de la génération 1990 (en %)
	Démographie des établissements	TSR	Taux de survie à 5 ans des reprises d'entreprises de la génération 1990 (en %)
	Démographie des établissements	SOLD	Solde relatif des transferts d'établissements sur la période 1993-1996
	Démographie des établissements	PM5A	Part des établissements de moins de 5 ans dans le stock des établissements au 01.01.97 (en %)
5	Niveau de formation	SUP	Part des bac+2 et plus dans la population ayant achevé ses études en 1990 (en %)
	Niveau de formation	BAC	Part des personnes ayant uniquement le bac dans la population ayant achevé ses études en 1990 (en %)
	Niveau de formation	CAP	Part des personnes ayant un CAP ou un BEP dans la population ayant achevé ses études en 1990 (en %)
	Niveau de formation	NOD	Part des non diplômés dans la population ayant achevé ses études en 1990 (en %)
6	Description du chômage	DE	Taux d'évolution annuel moyen des DEFM entre 1991 et 1996 (en %)
	Description du chômage	EJ	Taux d'évolution annuel moyen des DEFM de moins de 25 ans entre 1991 et 1996 (en %)
	Description du chômage	EG	Taux d'évolution annuel moyen des DEFM des 50 ans et plus entre 1991 et 1996 (en %)
	Description du chômage	CL	Part des chômeurs de longue durée dans les DEFM (catégories 1 et 6) au 31.12.96 (en %)
	Description du chômage	JE	Part des moins de 25 ans dans les DEFM (catégories 1 et 6) au 31.12.96 (en %)
	Description du chômage	PV	Part des 50 ans et plus dans les DEFM (catégories 1 et 6) au 31.12.96 (en %)
	Description du chômage	NQ	Part des non qualifiés dans les DEFM (catégories 1 et 6) au 31.12.96 (en %)
	Description du chômage	TR	Taux de retour à l'emploi en 1996 (en %)
	Description du chômage	FC	Part des fins de contrats précaires dans les entrées au chômage en 1996 (en %)
	Description du chômage	JB	Part des bas niveaux de formation parmi les jeunes entrant au chômage (catégorie 1) en 1996 (en %)

GROUPE	LIBELLE DU GROUPE	VARIABLE	LIBELLE DE LA VARIABLE
7	Population, démographie	UU	Part de la population urbaine dans la population totale en 1990 (en %)
	Population, démographie	SNAN	Taux d'évolution annuel moyen de la population totale dû au solde naturel entre 1990 et 1995 (en %)
	Population, démographie	SMAN	Taux d'évolution annuel moyen de la population totale dû au solde migratoire entre 1990 et 1995 (en %)
	Population, démographie	PMOY	Part des 20 à 64 ans dans la population totale en 1995 (en %)
	Population, démographie	MONO	Part des familles monoparentales dans l'ensemble des ménages en 1990 (en %)
	Population, démographie	PJE	Part des moins de 20 ans dans la population totale en 1995 (en %)
	Population, démographie	CP3	Part des couples ayant au moins 3 enfants de moins de 25 ans dans l'ensemble des ménages en 1990 (en %)
8	Population active	A	Taux global d'activité des 15 ans et plus en 1990 (en %)
	Population active	AJ	Taux d'activité des 15_24 ans en 1990 (en %)
	Population active	AG	Taux d'activité des 50_64 ans en 1990 (en %)
	Population active	AF	Taux d'activité des femmes de 25 à 49 ans en 1990 (en %)
	Population active	EV	Taux d'évolution annuel moyen de la population active entre 1975 et 1990 (en %)
	Population active	E	Taux d'entrée (nb postes travaillant dans la ZE mais résidant ailleurs / nb de postes travaillant dans la ZE)
	Population active	S	Taux de sortie (nb postes travaillant dans 1 autre ZE mais résidant dans cette ZE / nb postes résidant dans cette ZE)
	Population active	AR	Part des artisans-commerçants-chefs d'entreprises dans la population active ayant un emploi en 1990 (en %)
	Population active	O	Part des ouvriers dans la population active ayant un emploi en 1990 (en %)
	Population active	C	Part des cadres dans la population active ayant un emploi en 1990 (en %)
	Population active	EP	Part des employés dans la population active ayant un emploi en 1990 (en %)
	Population active	IN	Part des professions intermédiaires dans la population active ayant un emploi en 1990 (en %)
9	Revenus, conditions de vie	ER	Taux d'évolution annuel moyen du revenu imposable total entre 1984 et 1994 (en %)
	Revenus, conditions de vie	I	Part des foyers imposés en 1994 (en %)
	Revenus, conditions de vie	EM	Taux d'évolution annuel moyen du revenu imposable moyen entre 1984 et 1994 (en %)
	Revenus, conditions de vie	BR	Part des personnes vivant sous le seuil de bas revenu au 31.12.96 dans la population des moins de 65 ans (en %)
	Revenus, conditions de vie	AP	Part des personnes couvertes par l'APE au 31.12.96 (en %)
	Revenus, conditions de vie	R	Revenu net imposable moyen en 1994 (en milliers de francs)
	Revenus, conditions de vie	PRI	Part des résidences principales dans le parc de logements en 1990 (en %)
	Revenus, conditions de vie	PRO	Part des propriétaires au sein des résidences principales en 1990 (en %)
	Revenus, conditions de vie	LM	Part des locataires en HLM au sein des résidences principales en 1990 (en %)
	Revenus, conditions de vie	ID	Part des logements individuels dans les logements construits de 1991 à1996 (en %)
	Revenus, conditions de vie	PA	Part des logements sociaux PLA dans les logements construits de 1991 à1996 (en %)