CLASSIFICATION DES ZONES DE DISTRIBUTION POSTALE A PARTIR D'HISTORIQUES DE FLUX QUOTIDIENS DE COURRIER

Alain Dessertaine

alain.dessertaine@laposte.fr
LA POSTE – Direction du courrier DSCG/DCPES
Cité Descartes - 2 bd Newton
Champs sur Marne
77 453 Marne La Vallée Cedex 2

Cet article fait partie des Actes des 5èmes JOURNEES MODULAD qui ont eu lieu les 16 et 17 novembre 2000 à E.D.F. (Clamart).

1. Introduction

L'objet de cette étude est de montrer comment, dans une phase exploratoire de données, nous avons pu construire une classification de zones de distribution postale à partir d'historiques de flux quotidien de courrier par type d'objets distribués. Cette classification sera confrontée dans l'avenir avec des données externes décrivant nos clients récepteurs (données socio-économiques, démographiques etc. mais aussi des données de production interne) afin de mettre en place une modélisation des pointes de fort trafic d'objets à traiter dans un bureau de Poste distributeur.

Nous présenterons successivement le cadre général dans lequel l'étude est intégrée, les données ainsi que les concepts généraux permettant de mieux appréhender et comprendre notre démarche, et enfin la manière d'envisager le problème d'une classification de courbes de trafics hebdomadaires.

Les questions surgies durant l'exploitation des données que nous avons travaillées seront abordées, puis les éléments de réponse à ces questions, particulièrement sur l'approche classificatrice de courbes multidimensionnelles particulièrement perturbées dans un contexte de mutation de notre réseau de production.

Nous conclurons sur les axes de réflexion suscités par ce travail.

2. Cadre de l'étude présentée

Avec l'ouverture progressive du marché postal à la concurrence, LA POSTE cherche à mieux cerner son activité liée à la distribution du courrier. Cette activité, qui est au cœur du métier du courrier, est primordiale à plusieurs sens :

- La distribution est le dernier point de passage obligé dans le traitement du courrier national et international à l'import; elle fait partie intégrante des services proposés à nos clients émetteurs de courrier.
- Elle est un service rendu à nos clients récepteurs de courrier (soit potentiellement la totalité de la population ménage et entreprise).
- Elle représente une part très importante de nos charges (plus de 30 % de notre masse salariale !)

Des réflexions et des actions sont également menées afin :

- d'améliorer la qualité de service et tenir nos engagements contractuels,
- de proposer de nouveaux services à nos clients émetteurs (courrier suivi, nouveaux produits assurant non pas un délai d'acheminement, mais plutôt une période de livraison etc.)
- de proposer de nouveaux services à nos clients récepteurs (présentation multiple de certains objets...),
- de maîtriser nos coûts de production à la distribution.

Ainsi, l'approche statistique que nous allons présenter se situe au croisement de deux axes de réflexions :

- Etude des données existantes afin d'analyser la variabilité des flux de courrier dans les bureaux distributeurs. L'objectif est d'en saisir l'impact sur la flexibilité de notre réseau, et éventuellement d'aborder une modélisation de ces flux en fonction du temps mais aussi d'autres critères, tels que les descriptifs socio-économiques et géographiques de nos zones de distribution, ou des informations sur les étapes de production en amont de la distribution. Cette modélisation nous permettrait d'effectuer un certain nombre de simulations sur les futures mutations de notre réseau et de nos produits et services, ou encore d'aborder une discrimination des zones vis à vis de la typologie construite. Cette discrimination aura pour but de donner des outils d'aide à la décision sur la flexibilité de notre réseau afin d'améliorer nos coûts, tout en optimisant la qualité des services que nous proposons.

- Mise en place d'un panel de zones géographiques nous permettant de mieux appréhender l'adéquation entre notre organisation, nos charges et les flux de courrier à destination de ces zones. Ainsi, la présente approche statistique donnera des éléments afin de stratifier notre population pour une bonne construction de notre panel.

3. Présentation des données

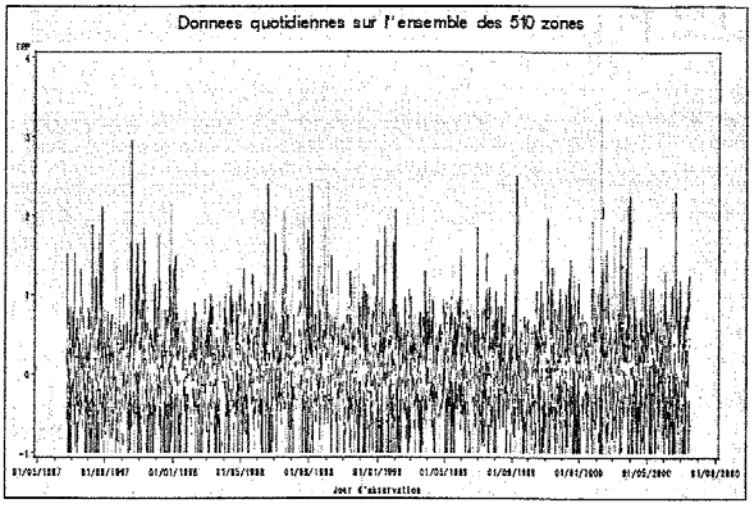
Rappelons, au préalable, qu'une étude semblable avait été commencée au début des années 1990, l'exploration statistique ayant été réalisée, à l'époque, par le CISIA. Seulement, cette étude avait fortement souffert de la qualité des données expérimentales construites pour l'occasion.

Ce problème a été en partie résolu pour notre étude : en effet, nous avons pu récupérer sur un échantillon de 680 bureaux des données concernant le trafic quotidien à traiter, ainsi que d'autres données de production, depuis le début de l'année 1997. Etant en période de fortes instabilités au niveau des organisations des outils de production (avec, entre autres, des « regroupements » ou des « éclatements » de bureaux de Poste ...), nous avons choisi

d'étudier les trafics à traiter au niveau de zones géographiques de distribution desservies par un ou plusieurs bureaux de poste distributeurs.

Après « nettoyage » de la base, et choix ou regroupement d'un certain nombre de bureaux, nous nous retrouvons avec un échantillon de 510 zones sur une période allant du 23 juin 1997 au 15 juillet 2000, soit 960 jours, donc 160 semaines. Nous étudierons plus particulièrement des indicateurs liés au trafic à traiter par type d'objet ou produit. Parmi ceux-ci, les indicateurs donnant le trafic à traiter relatif au trafic quotidien moyen par produit (quotidien ou hebdomadaire) font l'objet de l'étude présentée dans ce document. Ces indicateurs sont nommés ERP(produit), voire ERPTOT, pour celui calculé sur l'ensemble du trafic. Ils ont été calculés sur chacune des 510 zones de l'étude.

Pour montrer simplement la complexité des données, voici les graphes ERPTOT * DATE sur l'ensemble des 510 zones d'étude :



A la vue de ces graphiques, il est facile de comprendre pourquoi il nous a semblé souhaitable, dans un premier temps, de tenter de regrouper certaines courbes afin de mieux appréhender, au niveau de chaque groupe, la complexité due à la variabilité temporelle de ces flux en analysant les courbes moyennes du centre de gravité des classes ainsi construites.

Nous avons, par contre, préféré éclater ce vaste problème en deux :

- Analyse des courbes hebdomadaires (soit l'étude des ERP(produit) hebdomadaires)
- Analyse d'une typologie du trafic quotidien pour chaque couple zone-semaine.

Seul le premier point fait l'objet de cet article.

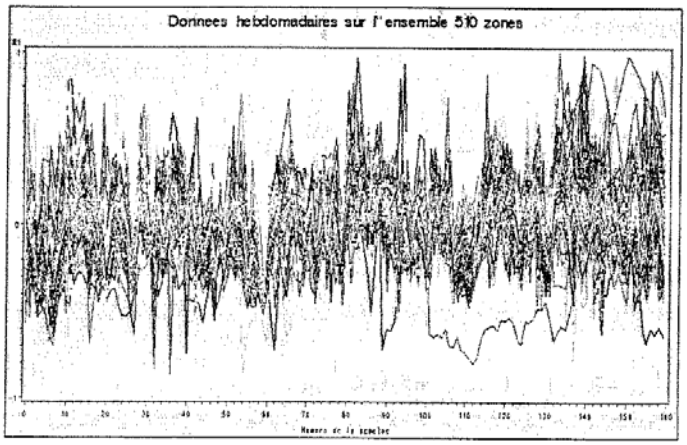
4. Classification des courbes de trafics hebdomadaires par produit commercial

Après avoir calculé sur chaque zone le trafic hebdomadaire pour chaque produit, relatif au trafic moyen hebdomadaire, nous nous retrouvons avec 510 courbes multidimensionnelles (de dimension 4) que nous voulons classifier suivant leurs positions dans l'espace en un instant donné, tout en prenant en compte leurs évolutions pour chacun de ces instants.

Pour ce faire, nous nous sommes inspirés des travaux de A. CARLIER, pour les classifications de trajectoires, et des nouvelles propositions faites pas F. DAZY, présentés dans le livre « L'Analyse des données évolutives – méthodes et applications », aux éditions Technip [DAZ96].

4.1. Pourquoi classifier nos courbes?

Un simple regard sur les courbes hebdomadaires par produit nous montre une grande dispersion des flux de courrier à traiter. Voici, par exemple, l'ensemble des 510 courbes de l'indicateur hebdomadaire ERPTOT :



4.2. Choix d'une distance entre courbes

Quelles que soient les méthodes de classification automatique utilisées, le problème primordial est le choix d'une distance adaptée à nos besoins! Ici, nous voulons construire des classes d'individus proches à la fois en position absolue dans l'espace, et ayant des évolutions hebdomadaires comparables dans cet espace.

Adaptons nos notations pour l'exposé de ce chapitre :

La mesure de l'indicateur X pour un produit p, un individu i et une semaine s sera : $X_{i,s}^{p}$.

A. CARLIER a proposé trois distances pour effectuer les classifications :

1. $d_1(i,i') = \sum_{s=1}^{160} \sum_{p=1}^{4} (X_{i,s}^{p} - X_{i',s}^{p})^2$

2. $d_2(i,i') = \sum_{(s,s') \in U_{160}} \sum_{p=1}^{4} \left((X_{i,s}^{p} - X_{i,s'}^{p}) - (X_{i',s}^{p} - X_{i',s'}^{p}) \right)^2$

   avec U_{160}, un sous-ensemble de $\{1,\ldots,160\} \times \{1,\ldots,160\}$[1].

3. $d_3(i,i')^2 = \beta_1 d_1(i,i')^2 + \beta_2 d_2(i,i')^2$

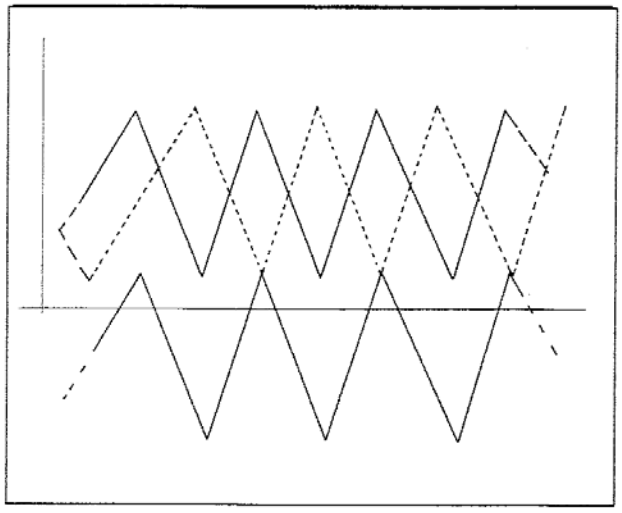
   où β_1 et β_2 sont des coefficients de pondération qui permettent de prendre en compte de manière similaire les distances d_1 et d_2.

Nous remarquons facilement que :

La distance d_1 est la distance euclidienne basée sur les positions des individus dans l'espace de travail. Elle prend des valeurs nulles si et seulement si les courbes sont confondues en chaque point. Par contre, deux courbes présentant des déphasages temporels, peuvent être à même distance que deux autres courbes « parallèles », si les distances entre chaque couple de points sont identiquement distribuées dans les deux cas. L'utilisation de cette distance dans un

[1] Rappelons que 160 est le nombre de semaines disponibles dans nos données.

processus de classification regrouperait des courbes présentant des « déphasages » dans le temps non adaptés à nos objectifs. Voici un exemple illustrant notre propos ; ici, les distances sont calculées à partir des points définissant les arêtes de chaque segment de droite : Les deux courbes en traits pleins sont en phase dans le temps. Mais la courbe supérieure est à égale distance de la courbe en traits pointillés, pourtant en opposition de phase avec les deux autres.



- La distance d_2 est la distance euclidienne basée sur les évolutions des individus sur les 4 dimensions de travail. La classification est une classification sur les évolutions. En choisissant, par exemple, un sous-ensemble U_{160} des 159 couples de semaines successives de la période, nous pouvons construire cette nouvelle distance :

$$d'_2(i,i') = \sum_{s=2}^{160}\sum_{p=1}^{4}\left((X^p_{i,s} - X^p_{i,s-1}) - (X^p_{i',s} - X^p_{i',s-1})\right)^2$$

Il est immédiat alors de constater que si i et i' définissent deux courbes telles que celles qui apparaissent en traits pleins sur la figure précédente, la distance est nulle.

Dans notre problématique, cette distance ne nous intéresse pas totalement ; certes, elle permet par construction de minimiser les regroupements de courbes pourtant en « déphasage » dans le temps, mais elle risque de créer des classes d'individus avec une forte dispersion sur nos 4 variables d'études. Cette remarque reste valable quels que soient les sous-ensembles U_{160} utilisés.

- La distance d_3 est une distance basée sur un compromis entre les deux distances précédentes, donc entre les évolutions et positions. Mais, elle est délicate à mettre en place pour un certain nombre de raisons, dont le choix des coefficients β_1 et β_2 ! Pourtant ce compromis permettrait de minimiser les désagréments causés par les deux premières distances présentées.

Aussi avons-nous cherché à utiliser la distance proposée par F.DASY afin de trouver un compromis entre positions et évolutions. Cette distance est basée sur l'utilisation combinée des coordonnées de chaque point des courbes (que nous appellerons« coordonnée-position ») avec des nouvelles coordonnées décrivant les évolutions par rapport à la date juste antérieure (que nous appellerons « coordonnée-évolution »)[2]. Exposons le calcul de cette distance (avec les notations de notre problème):

Il faut d'abord définir ce que M. DAZY appelle une configuration.

Une configuration est une suite de 160 nombres binaires $(a_s)_{s=1,\ldots,160}=\{a_1,\ldots,a_{160}\}$ telle que :

- $a_1=1$
- $a_s=1$ si l'on considère une « coordonnée-position » à la semaine s
- $a_s=0$ si l'on considère une « coordonnée-évolution » à la semaine s

pour $s \in \{2, \ldots,160\}$:

Ainsi, pour une configuration δ donnée, nous pouvons calculer la distance entre deux courbes de la manière suivante :

$$d_\delta(i,i') = \sum_{s=1}^{160}\sum_{p=1}^{4}\left(a_s(X^p_{i,s} - X^p_{i',s})^2 + (a_s - 1)\left((X^p_{i,s} - X^p_{i,s-1}) - (X^p_{i',s} - X^p_{i',s-1})\right)^2 \right)$$

[2] Notons que nous pouvons utiliser des calculs d'évolution entre deux dates successives, ce que nous allons faire ici, mais nous pouvons aussi utiliser des évolutions prenant en compte des périodicités communes aux courbes.

Le problème consiste simplement à trouver une configuration δ la mieux adaptée à notre problème. F.DASY dit lui-même qu'il faut trouver une configuration δ *« permettant de déterminer une configuration qui possède sa propre optimalité ».*

Dans l'objectif de prendre en compte un compromis réel entre position et évolution, l'auteur propose de calculer les inerties $I\delta_{TOT}$, $I\delta_{POSITION}$ et $I\delta_{EVOLUTION}$ qui sont simplement les inerties de notre nuage de points définies sur la configuration δ, calculées respectivement sur l'ensemble des coordonnées, sur les « coordonnées-position » et sur les « coordonnées-évolutions ». Alors, la configuration « optimale » permettant de prendre en compte de manière semblable les positions et les évolutions est celle qui minimise la valeur absolue de la différence entre $I\delta_{POSITION}$ et $I\delta_{EVOLUTION}$.

Le nombre de configuration possible étant fini (dans notre cas, il en existe 2^{159}...), ce problème de minimisation admet une solution, en général unique.

Se posent alors deux problèmes :

- Comment trouver dans un temps raisonnable la solution ?
- La solution trouvée, répond-elle véritablement à notre problème ?

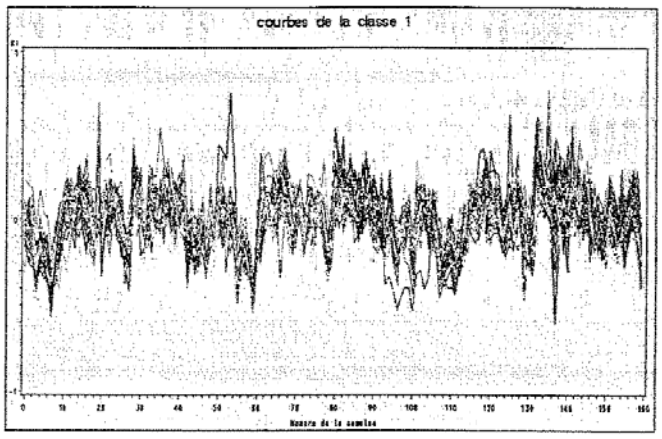
4.3. Adaptation de la méthode et premiers résultats

4.3.1. Premiers constats :

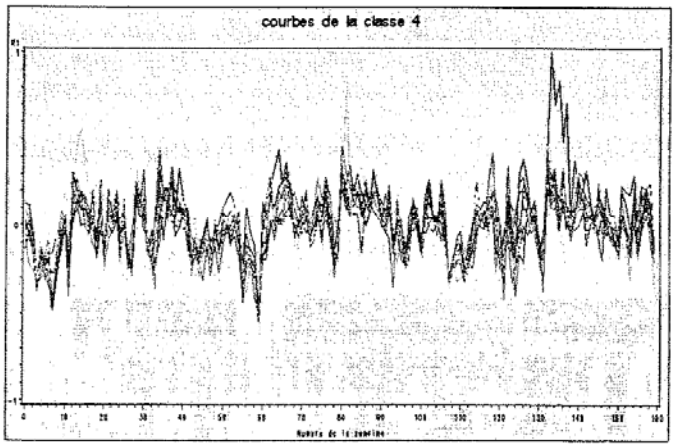
La manière la plus rapide, dans un premier temps, de trouver une configuration proche de la solution « optimale » a consisté à tirer au hasard un certain nombre de configurations δ_i[3], de calculer les inerties $I\delta_{i\ POSITION}$ et $I\delta_{i\ EVOLUTION}$ et de conserver parmi les configurations construites celle qui minimisait la valeur absolue de la différence entre les deux inerties calculées. Une classification mixte (méthode SEMIS de SPAD) a été réalisée sur les coordonnées de cette configuration.

[3] Nous avons testé avec $i \in [1 ;100]$.

Voici pour illustrer nos propos les courbes ERP des individus d'une classe issue de cette classification :



Un simple regard sur cette classe nous montre qu'il existe encore une forte hétérogénéité parmi les nombreux éléments de la classe. Nous voyons aussi que les hétérogénéités proviennent du fait que les courbes sont, par construction, regroupées par rapport à leurs proximités sur certaines semaines, et par rapport à leurs évolutions entre deux semaines consécutives sur d'autres périodes. Une question s'est alors naturellement posée: « Existe-t-il des formes fortes ? » Nous sentons bien, à la visualisation du prochain graphe l'existence probable des groupes de zones bien homogènes :



Il nous faut donc repérer ces « formes fortes » :

4.3.2. Recherche des formes fortes :

Il est clair que la classification construite préalablement était trop liée à la configuration choisie. De ce fait, nous avons abandonné l'idée d'aller à la quête de la configuration « optimale » au sens de F DAZY car elle n'aurait pas répondu suffisamment à notre problématique[4].

Aussi avons-nous eu l'idée de réitérer un certain nombre de fois les étapes précédentes. A chaque étape E, nous avons choisi une configuration δ_E « optimale », et nous avons construit une partition de 20 classes que nous appellerons $\Pi(\delta_E)$. A partir de ces partitions, nous avons alors étudié les formes fortes, en regroupant ensemble des zones se retrouvant fréquemment ensemble dans les partitions $\Pi(\delta_E)$.

La reconnaissance des formes fortes a été réalisée en 3 étapes :

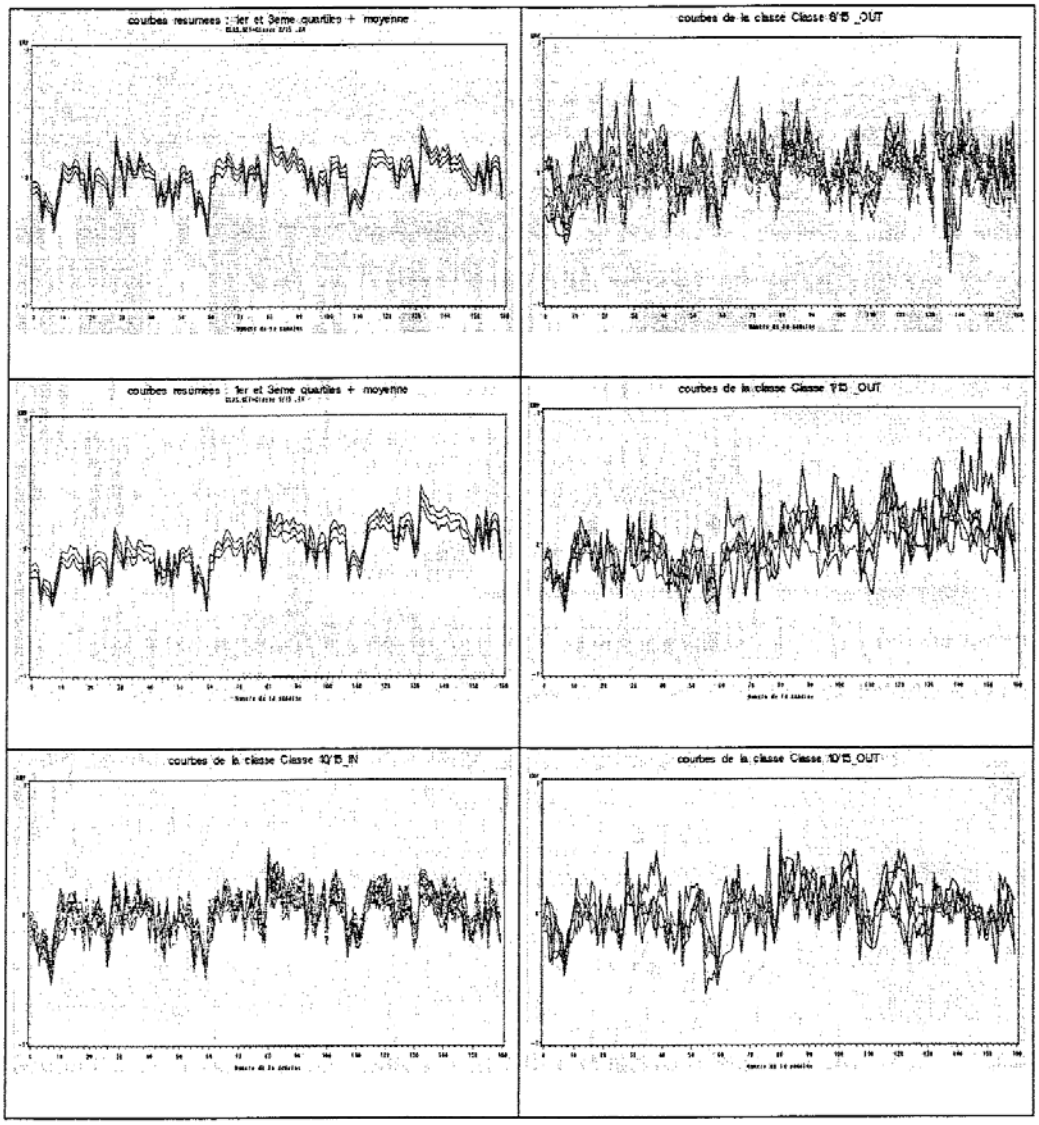
1. Construction du tableau disjonctif complet à partir des modalités correspondant aux classes des partitions $\Pi(\delta_E)$.
2. Classification des zones sur ces nouvelles variables.
3. « Elagage » des zones dont la courbe apportait une inertie relative trop importante dans le calcul de l'inertie intra-classes de sa classe d'appartenance.

Nous avons construit 15 classes correspondant à nos « formes fortes », puis nous avons construit 15 nouvelles « classes » à partir des courbes « élaguées » des 15 premières.

Voici, à titre d'exemple, quelques graphiques illustrant ces regroupements :

---

[4] Notons que la partition que nous avons construite était basée sur une configuration tout à fait intéressante : la différence relative calculée entre les deux inerties étant de l'ordre de 10^{-5}



5. Réflexions et axes de recherche

La classification que nous avons construite nous a permis d'explorer nos données en prenant en compte l'aspect temporel de celles-ci. Il reste, évidemment, un travail important afin de juger de la réelle pertinence de ces regroupements.

En effet, cette classification « expérimentale » dépend de plusieurs paramètres que nous avons choisis de manière arbitraire[5] :

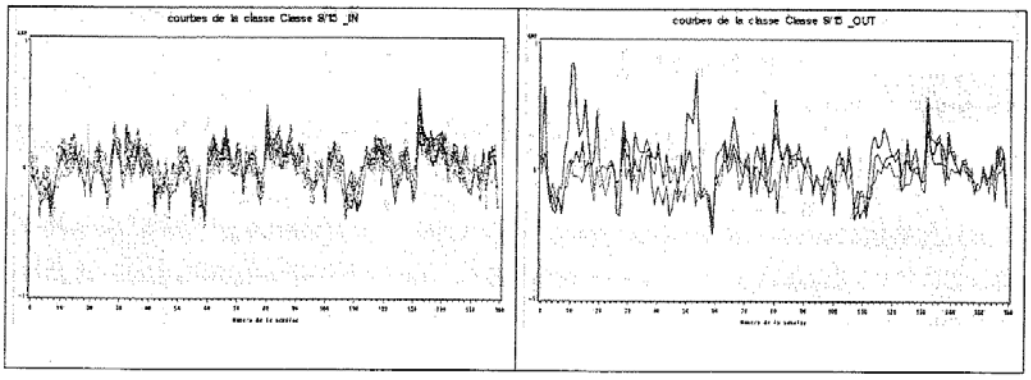
- Le nombre d'itérations pour construire les configurations δ_E de chaque étape.
- Le nombre de classes construites à chaque étape $\Pi(\delta_E)$.
- Le choix de la méthode construction des partitions $\Pi(\delta_E)$ (l'utilisation des méthodes neuronales, et plus particulièrement des cartes de Kohonen est en cours…)
- Le nombre d'étapes.

D'autre part, il est nécessaire d'étudier l'impact du choix de ces paramètres sur :

- le rapport inertie intra-classes sur inertie totale. (une étude sur la construction d'un indicateur externe à la méthode permettant de juger la pertinence des partitions créées est en cours …)
- l'importance de l'élagage final au moment de l'étude des formes fortes.

Nous réfléchissons également sur l'opportunité de prendre en compte de la même manière l'ensemble des dates en notre possession, dans le cadre d'une classification qui permettrait de faciliter les tâches de modélisation. Ne devrions nous pas pondérer plus fortement les périodes récentes ?

Illustrons nos propos avec deux classes 'dissemblables' qui, pourtant, paraissent avoir des comportements moyens proches sur une bonne moitié de la période étudiée, et particulièrement sur la fin de la période :



[5] Ces paramètres ont été choisis afin de rendre « acceptable » notre classification finale…

Nous cherchons également à construire des indices de similitude entre deux zones en appliquant sur l'une d'elle un modèle 'simple' de séries chronologiques construit sur l'autre. L'indice se calculerait à partir de l'erreur quadratique constatée entre la prévision et la valeur réelle. Des premiers travaux prometteurs ont déjà été réalisés, basés sur l'utilisation d'autant de réseaux de neurones prédictifs que nous avons de zones.

D'autres travaux exploratoires sont en cours, consistant à étudier la forme de la courbe quotidienne multidimensionnelle sur chaque couple zone-semaine afin de les regrouper de manière homogène, et d'expliquer ces regroupements en fonction de la position du couple dans le temps (le mois d'appartenance, en début ou en fin de mois, l'année…) mais aussi en fonction de la typologie de la zone d'appartenance.

Des travaux sont en cours, autour de l'approche fréquentielle basée sur des séries de Fourier et sur une analyse harmonique des séries chronologiques (classification des spectres), parallèlement à une classification sur l'approximation de nos signaux à base d'ondelettes (ceci afin de réduire la dimension de l'espace de projection).

BIBLIOGRAPHIE

[DAZ96] DAZY F et LE BARSIC J : L'analyse des données évolutives : Méthodes et applications (éditions TECHNIP, 1996)

[LEB97] LEBART L, MORINEAU A et PIRON M : Statistique exploratoire multidimensionnelle (éditions DUNOD, 1997)

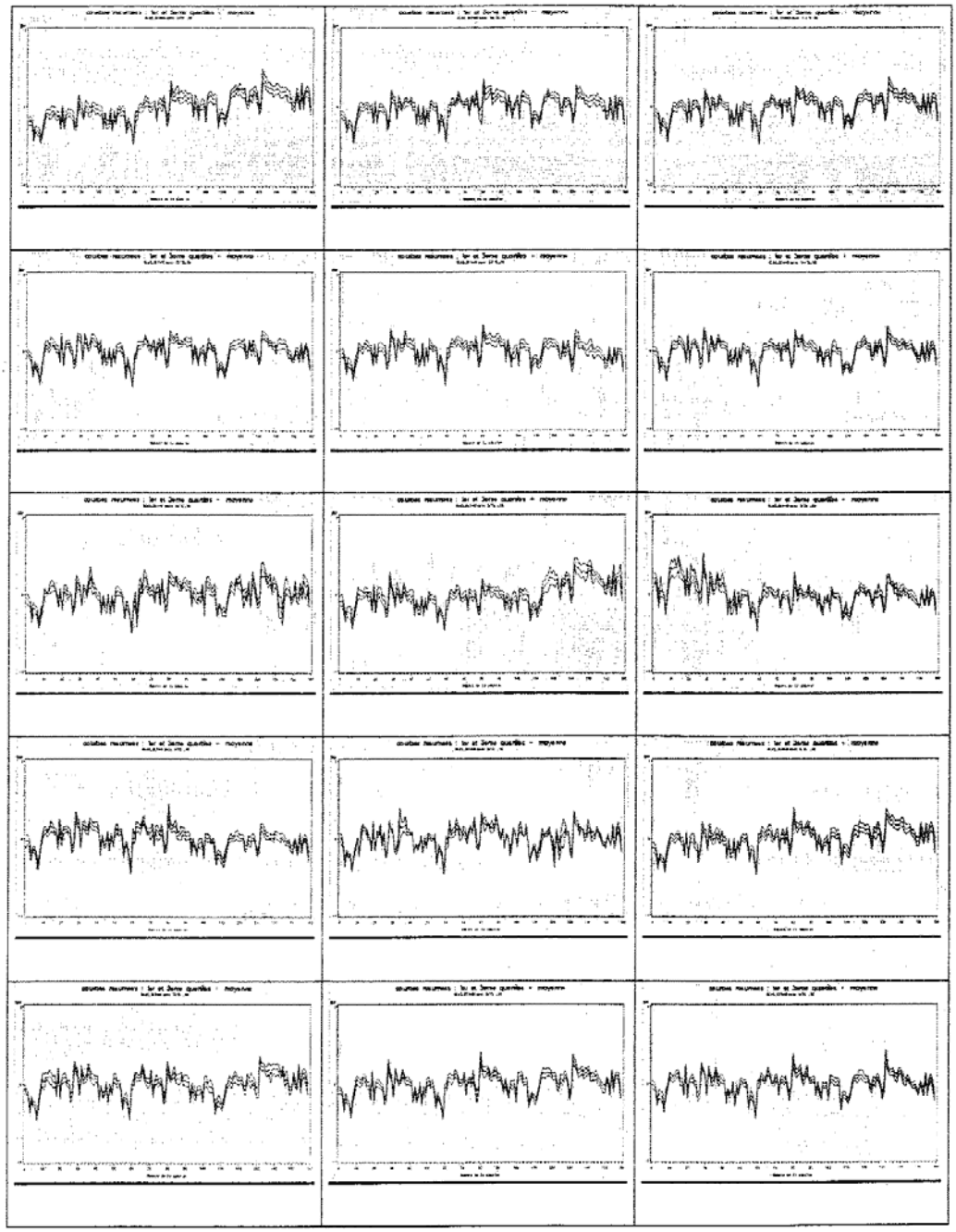
[THI97] THIRIA S, LE CHEVALLIER Y, GASCUEL et CANU S : Statistique et méthodes neuronales (éditions DUNOD, 1997)

[CIS94] CISIA : « Etude du courrier à l'arrivée – modèles de prévision (Structure et saisonnalités - Analyse de la structure) » (rapport d'étude, 1994)

[JAC98] JACQUET G : « Réseaux de neurones formels appliqués à l'étude des variations quotidiennes des parts de trafic distribué en bureaux de Poste – Essai de typologie de séries chronologiques » (rapport d'étude, 1998)

[NOV01] NOVELLI P : Mémoire de DEA : Etude des flux de trafics postaux, Paris 1 Panthéon-Sorbonne (2001)

Annexe : Courbes résumées des 15 classes construites : 'courbe moyenne' encadrée des courbes '1er quartile' et '3eme quartile'



Statistique et logiciels