Représentation géographique du tissu industriel : une application de l'approche MODL

Alexis Bondu[1], Marc Boullé[2], Anne Peradotto[1]

[1] EDF R&D ICAME/SOAD, 1 avenue du Général de Gaulle, 92140 Clamart.
prenom.nom@edf.fr

[2] OrangeLabs, 2 avenue Pierre Marzin, 22300 Lannion.
marc.boulle@orange-ftgroup.com

Résumé :
Cet article propose une nouvelle utilisation de l'approche de regroupement de modalités bivariée MODL sur des données géographiques. Les données exploitées décrivent la localisation et l'activité des entreprises de Paris intra-muros. Les regroupements géographiques et les regroupements d'activités définis conjointement sont visualisés sur des cartes et sur des histogrammes. Cette représentation met en évidence les corrélations entre groupes d'activités et zones géographiques.

Mots-clés : MODL, co-clustering, données géographiques

1 Introduction

Le principal apport de cet article applicatif est de proposer une nouvelle utilisation de l'approche de regroupement de modalités MODL [1] (Boullé, 2007). Les données utilisée décrivent la localisation et l'activité des entreprises de Paris intra-muros. L'objectif de notre étude est double : i) regrouper automatiquement des zones géographiques où le tissu d'activités est similaire ; ii) donner une vue synthétique du tissu d'activités et de sa répartition géographique. Nous modélisons les corrélations entre zones géographiques et les activités professionnelles grâce à un modèle de regroupement de modalités bivarié. Les individus statistiques considérés sont les 605 639 entreprises de paris, caractérisées par deux variables catégorielles : le code NAF et l'identifiant de la "voie" d'appartenance. Ces deux variables présentent toutes les deux un nombre important de modalités : respectivement 734 et 5146. Pour notre étude nous avons choisi d'utiliser l'approche MODL.

1. MODL : *Maximized Optimal Description Length*

2 L'approche MODL

définition : Un modèle de groupement de modalité bivarié est défini par
- un nombre de groupes pour chaque variable à expliquer,
- la partition de chaque variable à expliquer en groupes de valeurs,
- la distribution des individus sur les cellules de la grille de données ainsi définie,
- la distribution des individus de chaque groupe sur les valeurs du groupe, pour chaque variable à expliquer.

notations :
- N : nombre d'individus de l'échantillon
- V_1, V_2 : nombre de valeurs pour chaque variable (connu)
- J_1, J_2 : nombre de groupes pour chaque variable (inconnu)
- $G = J_1 J_2$: nombre de cellules de la grille du modèle
- $j^{(1)}(v_1), j^{(2)}(v_2)$: index du groupe auquel est rattachée la valeur v_1 (resp. v_2)
- $m_{j_1}^{(1)}, m_{j_2}^{(2)}$: nombre de valeurs du groupe j_1 (resp. j_2)
- $n_{v_1}^{(1)}, n_{v_2}^{(2)}$: nombre d'individus pour la valeur v_1 (resp. v_2)
- $N_{j_1}^{(1)}, N_{j_2}^{(2)}$: nombre d'individus du groupe j_1 (resp. j_2)
- $N_{j_1 j_2}$: nombre d'individus de la cellule (j_1, j_2) de la grille

Le meilleur modèle (noté MAP [2]) est sélectionné grâce à une approche Bayesienne visant à maximiser la probabilité $P(M|D) = P(M)P(D|M)/P(D)$ du modèle connaissant les données. A cet effet, une distribution a priori sur les paramètres des modèles est définie (Boullé, 2007). Cette distribution a priori est hiérarchique, et uniforme à chaque étage de la hiérarchie du paramètrage des modèles. La formule de Bayes est ensuite développée sous l'hypothèse d'indépendance des distributions entre les groupes. Finalement, la probabilité d'un modèle connaissant les données est calculée de manière exacte. Un modèle d'estimation de densité par grille suivant un a priori hiérarchique est optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles :

$$\begin{aligned}
&\log V_1 + \log V_2 + \log B(V_1, J_1) + \log B(V_2, J_2) \\
&+ \log \tbinom{N+G-1}{G-1} + \sum_{j_1=1}^{J_1} \log \tbinom{N_{j_1}^{(1)} + m_{j_1}^{(1)} - 1}{m_{j_1}^{(1)} - 1} + \sum_{j_2=1}^{J_2} \log \tbinom{N_{j_2}^{(2)} + m_{j_2}^{(2)} - 1}{m_{j_2}^{(2)} - 1} \\
&+ \log N! - \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \log N_{j_1 j_2}! \\
&+ \sum_{j_1=1}^{J_1} \log N_{j_1}^{(1)}! + \sum_{j_2=1}^{J_2} \log N_{j_2}^{(2)}! - \sum_{v_1=1}^{V_1} \log n_{v_1}^{(1)}! - \sum_{v_2=1}^{V_2} \log n_{v_2}^{(2)}!
\end{aligned} \tag{1}$$

$B(V, J)$ est le nombre de Belle de deuxième espèce dénombrant toutes les répartitions possible de V valeurs explicatives en J groupes éventuellement vides. La première ligne de la formule (1) regroupe des termes d'a priori correspondant au choix des nombres de groupes J_1 et J_2 et à la spécification de la partition de chaque variable à expliquer en groupes de valeurs. La deuxième ligne représente la spécification de la distribution multinômiale des N individus de l'échantillon sur les G cellules de la grille, suivi de la spécification de la distribution des individus de chaque groupe sur les valeurs du groupe. La troisième ligne représente la vraisemblance de la distribution des individus dans les cellules de la grille, au moyen d'un terme du multinôme. La dernière ligne correspond à la vraisemblance des valeurs localement à chaque groupe pour chacune des variables à expliquer.

2. Le "MAP" signifie *Maximum A Posteriori*.

3 Application au regroupement de voies et de codes NAF

La méthode présentée à la Section 2 est appliquée à nos données et un modèle optimal (MAP) est défini. Le MAP est constitué de 50 groupes de codes NAF et de 234 groupes de "voie". En terme de visualisation, le MAP peut être exploité de deux manières : i) en fixant un groupe de codes "Naf" ; ii) en fixant un groupe de "voies".

Visualisation étant donné un groupe de "voies"

Le MAP regroupe les voies dont le tissu d'activités professionnelles est similaire et estime la distribution de probabilité des groupes de codes NAF dans un groupe de "voies" donné. A titre illustratif, considérons le groupe de voies N° 92 qui est constitué de la rue de la paix, la place Vendôme, la rue Montaigne et la rue Balzac. La Figure 1 représente la répartition des entreprises du groupe de "voies" N° 92 (en bleu) sur l'ensemble des groupes de codes "NAF" définis par le MAP. Les groupes de codes "NAF" caractérisent l'activité des entreprises du jeu de données et peut se substituer au code NAF "réel" qui comporte initialement 734 modalités. Le MAP synthétise donc l'information présente dans le jeu de données. La Figure 1 présente également (en rouge) la répartition moyenne des entreprises parisiennes sur les groupes de codes "NAF". Ce diagramme peut être exploité pour identifier des groupes de codes NAF caractéristiques d'un groupe de "voies", que ce soit en terme d'effectif ou en terme d'écart à la répartition moyenne.

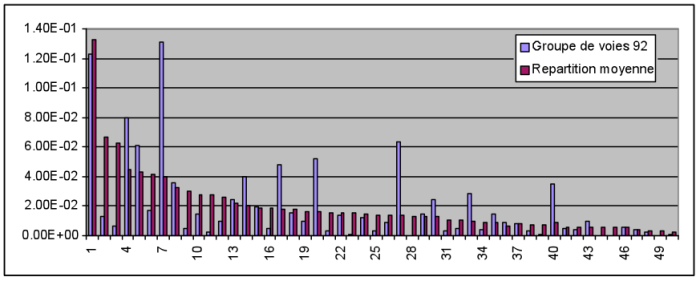

FIGURE 1 – Répartition des groupes de codes NAF au sein du groupe de voies N°92

Visualisation étant donné un groupe de codes "NAF"

La répartition d'un groupe de codes "NAF" peut être visualisée sur une carte géographique. A titre illustratif, nous considérerons le groupe de codes "NAF" N° 3 (labellisé "Arts et Spectacle") qui regroupe les modalités suivantes : arts du spectacle vivant (9001Z), autre création artistique (9003B), création artistique relevant des arts plastiques (9003A). La Figure 2 représente la répartition du groupe de code "NAF" N° 3 sur la carte de Paris. Plus la couleur d'une rue est foncée, plus la probabilité qu'une entreprise de cette rue appartienne au groupe de codes "NAF" N° 3 est importante. Cette carte met en évidence une différence très marquée entre, d'une part les arrondissements 17, 8, 1, 7, 15, 16 et d'autre part le reste de la capitale.

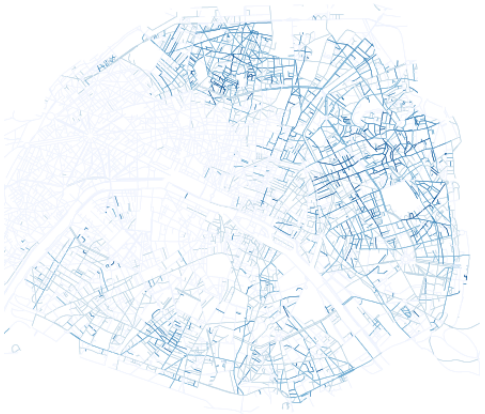

FIGURE 2 – Répartition de l'activité "Arts et spectacles" sur la capitale

4 Conclusion

Cet article propose une nouvelle utilisation de l'approche de regroupement de modalités MODL. Cette approche est appliqué à des données décrivant la localisation (à l'échelle de la voie) et l'activité des entreprises de Paris intra-muros. Les regroupements géographiques et les regroupements d'activités définis conjointement par cette approche sont ensuite visualisés sur des cartes et sur des histogrammes. La représentation proposée est synthétique et met en évidence les corrélations entre groupe d'activités et zones géographiques.

Références

BOCK H. (1979). Simultaneous clustering of objects and variables. In E. DIDAY, Ed., *Analyse des Données et Informatique*, p. 187–203 : INRIA.

BOULLÉ M. (2007). *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. Phd thesis, ENST (Ecole Nationale Supérieur des Télécommunications).

DHILLON I. S., MALLELA S. & MODHA D. S. (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, p. 89–98.

GOVAERT G. & NADIF M. (2006). Classification d'un tableau de contingence et modèle probabiliste. *Revue des Nouvelles Technologies de l'Information*, **2**, 457–462.

HARTIGAN J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67**(337), 123–129.

LECHEVALLIER Y. & VERDE R. (2004). Crossed clustering method : An efficient clustering method for web usage mining. Complex Data Analysis, PÈkin, Chine.

POIRIER D., BOTHOREL C. & BOULLÉ M. (2008). Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires. In *Extraction et gestion des connaissances (EGC'2008)*.