Étude des corrélations spatio-temporelles des appels mobiles en France

Romain Guigourès*,**, Marc Boullé*, Fabrice Rossi**

*Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
{romain.guigoures, marc.boulle}@orange.com
**SAMM - Université Paris 1, 90 rue Tolbiac, 75013 Paris
{fabrice.rossi}@univ-paris1.fr

Résumé. Nous proposons dans cet article de présenter une application d'analyse d'une base de données de grande taille issue du secteur des télécommunications. Le problème consiste à segmenter un territoire et caractériser les zones ainsi définies grâce au comportement des habitants en terme de téléphonie mobile. Nous disposons pour cela d'un réseau d'appels inter-antennes construit pendant une période de cinq mois sur l'ensemble de la France. Nous proposons une analyse en deux phases. La première couple les antennes émettrices dont les appels sont similairement distribués sur les antennes réceptrices et vice versa. Une projection de ces groupes d'antennes sur une carte de France permet une visualisation des corrélations entre la géographie du territoire et le comportement de ses habitants en terme de téléphonie. La seconde phase découpe l'année en périodes entre lesquelles on observe un changement de distributions d'appels sortant des groupes d'antennes. On peut ainsi caractériser l'évolution temporelle du comportement des usagers de mobiles dans chacune des zones du pays.

1 Introduction et présentation du problème

L'étude des populations par l'analyse de détails d'appel téléphonique est un problème auquel s'intéressent les opérateurs de téléphonie mobile depuis quelques années. Différentes études ont été menées sur le sujet, dont certaines s'intéressent à déterminer la structure communautaire des populations grâce au trafic inter-antennes (Blondel et al., 2010), (Guigourès et Boullé, 2011). Ces études de journaux d'appels téléphoniques permettent de mettre évidence une segmentation naturelle du territoire liée à la langue ou aux zones d'influences des grandes métropoles à l'échelle d'un pays, ou encore au profil économique et social des quartiers (bourgeois, populaire, étudiant ...) dans des études plus locales. De telles analyses intéressent les opérateurs de tous les pays, notamment ceux en voie de développement, où les besoins en terme de télécommunications sont amenés à être importants et où les usages demeurent actuellement encore inconnus.

Pour aller plus loin, on peut s'intéresser à délimiter des zones géographiques où le comportement des usagers de téléphones mobiles est différent en fonction de la période de temps étudiée. Une analyse temporelle du trafic au niveau des groupes d'antennes permet de caractériser les zones géographiques grâce aux excès et déficits d'appels dans chacune des périodes étudiées.

Une telle étude apporte aussi bien une information sur les différentes plages horaires qui structurent la journée, la semaine, le mois ou l'année des utilisateurs de téléphones portables, que sur les lieux où les phénomènes temporels sont observés.

Une des contraintes dont il faut tenir compte dans ce type d'études est la volumétrie des données. Les données dont nous disposons et sur lesquelles nous souhaitons étudier les corrélations spatio-temporelles, sont un enregistrement quotidien des appels inter-antennes passés en France métropolitaine entre le 13 Mai et le 13 Octobre 2007. Le nombre d'antennes réparties sur le territoire français est de 17895 entre lesquelles ont transités $1{,}12$ milliards d'appels. Dans la section 2, nous présentons des méthodes adaptées à ce type d'analyse et justifions le choix de la solution utilisée. La section suivante présentera les résultats de l'analyse des corrélations spatiales, puis la section 4 les corrélations temporelles. Une dernière partie conclut cet article en faisant un bilan de l'analyse.

2 État de l'art et choix de la solution

La première question qu'il est important de se poser est la représentation des données. En effet, un appel est décrit par l'antenne source, l'antenne cible et le jour de l'appel. Une précédente étude (Blondel et al., 2010) propose de représenter ces données sous forme d'un graphe non orienté, définissant ainsi un réseau d'antennes liées par le nombre d'appels transitant entre elles. Nous préférons conserver le format tabulaire pour traiter ce problème, et ainsi garder l'orientation naturelle des appels entre les antennes.

2.1 État de l'art

Blondel et al. (2010) proposent de partitionner le réseau d'antennes suivant une approche de maximisation de modularité. La modularité (Newman, 2006) évalue la qualité de la partition d'un graphe en cliques, ou communautés dans la terminologie des réseaux sociaux, qui sont des ensembles de nœuds fortement connectés entre eux. Cette technique possède l'avantage d'être efficacement optimisable en terme de complexité (Blondel et al., 2008) et donc de pouvoir traiter de très grands volumes de données, comme c'est le cas ici. Cependant, l'hypothèse faite lorsqu'on emploie ce type d'approches est très forte en supposant que le réseau est modulaire, c'est-à-dire qu'il peut se subdiviser en groupes de nœuds fortement connectés entre eux. Si le réseau possède en effet une structure communautaire, la maximisation de modularité sera un outil très efficace. Cependant, la structure sous-jacente du réseau dans le cas présent étant inconnue, on ne peut faire d'hypothèse sur la nature des motifs que nous souhaitons faire émerger des données. Nous devons donc trouver une alternative qui permette de considérer toutes les structures possibles du réseau.

Le concept de blockmodeling est à l'origine des premiers travaux d'analyse des structures dans les graphes menés par les sociologues dès les années 1950 dans le contexte de l'analyse des réseaux sociaux (Nadel, 1957). Pour utiliser ce type d'approches, nous devons étudier les données sous leur forme tabulaire. Les lignes et les colonnes de cette matrice représentent respectivement les antennes source et cible, et les valeurs indiquent la fréquence d'appels entre couples d'antennes. Il est ainsi possible de réorganiser les lignes et les colonnes dans le but de découper la matrice en blocs homogènes. Cette technique est appelée *Blockmodeling*. Une fois les blocs extraits, une partition des antennes représentées à la fois par les lignes et

les colonnes peut être réalisée. Cette segmentation simultanée est appelée *coclustering*. En procédant ainsi, nous sommes capables de capturer des structures plus complexes que les structures retrouvées par maximisation de modularité, méthode qui correspond finalement à un blockmodeling diagonal.

De nombreuses méthodes de blockmodeling ont été proposées pour extraire des groupes dans les réseaux. Certaines se basent sur l'optimisation d'un critère (Doreian et al., 2004) permettant d'isoler des blocs homogènes en se focalisant sur les blocs vides comme il est suggéré dans (White et al., 1976). Des approches déterministes plus récentes se sont intéressées à l'optimisation de critères qui mesurent la qualité de la partition en blocs, en termes de résumé des données d'origine (Reichardt et White, 2007). D'autres utilisent le blockmodeling stochastique. Dans ces modèles génératifs, une variable latente indiquant l'appartenance ou non à un cluster est associée à chaque nœud. Conditionnellement à leur variable latente, la probabilité d'observer un arc entre deux acteurs suit une loi de probabilité (Bernoulli dans les cas les plus simples) dont les paramètres dépendent uniquement de la paire de clusters désignés par la variable latente. Les premières approches nécessitaient une paramétrisation du nombre de clusters par l'utilisateur (Nowicki et Snijders, 2001), alors que les méthodes les plus récentes préfèrent le déterminer automatiquement en utilisant un processus de Dirichlet (Kemp et Tenenbaum, 2006).

Outre la diversité des structures pouvant être inférées dans le réseau par les approches de coclustering, il est également possible réaliser un coclustering avec des variables numériques (Nadif et Govaert, 2010),(Boullé, 2012). Des blocs sont extraits des données et produisent une discrétisation de la ou des variables numériques. On peut ainsi appliquer, pour une seconde analyse, un blockmodeling sur des données tabulaires dont les lignes sont les antennes source et les colonnes une variable numérique temporelle. Ainsi, il est possible de trouver des structures temporelles par utilisation de la même méthode.

Dans le cas d'une analyse de comptes rendus d'appels, la technique employée doit présenter plusieurs propriétés :

- **Passage à l'échelle** : avec près de 18000 antennes et $1{,}12$ milliards d'appels, on ne peut pas se permettre d'avoir une complexité algorithmique trop forte, ce qui est souvent le défaut des techniques de blockmodeling et de coclustering.
- **Généricité** : les données traitées peuvent être aussi bien numériques que catégorielles. Ce point est important car l'analyse que nous menons porte à la fois sur des variables catégorielles (les Antennes) que numériques (le temps).
- **Absence de paramétrage** : les données sont complexes et leur structure inconnue, un paramétrage de la structure de coclustering (e.g le nombre de clusters de chaque variable ou la distribution des antennes dans les groupes) serait trop complexe avec un tel jeu de données.
- **Fiabilité** : la méthode utilisée ne doit pas produire de résultats lorsqu'il n'existe aucune structure sous-jacente. Elle doit donc être résistante au bruit et ne pas surapprendre.
- **Finesse et interprétabilité** : l'approche doit capturer toute l'information présente dans les données afin d'extraire des motifs fins. Des outils destinés à l'interprétation des résultats doivent également être proposés afin d'exploiter de manière efficace les résultats.

Étant donné le volume de données, la plupart des méthodes de coclustering ont une complexité algorithmique telle qu'on ne pourrait pas les appliquer directement sur la base de

données. Une idée serait donc de travailler sur un échantillon de données. Cependant, avec 17895 antennes et 1,12 milliards d'appels, le nombre d'appels moyen entre 2 antennes est d'environ $3,5$. Ainsi, un échantillonnage entrainerait une perte d'information importante. Parmi les approches de coclustering, nous retiendrons l'approche MODL (Boullé, 2011) [1] .

2.2 L'approche MODL

Afin de bien formaliser le problème, le tableau 1 liste les caractéristiques des données ainsi que les paramètres de modélisation du coclustering que nous cherchons à déterminer.

$\mathcal{D}$ **: Données**	$\mathcal{M}_S$ **: modèle de coclustering spatial**	$\mathcal{M}_T$ **: modèle de coclustering temporel**
- V_S : Antennes source	- V_S^M : partition de V_S en clusters d'antennes source	- V_S^M : partition de V_S en clusters d'antennes source
- V_C : Antennes cible	- V_C^M : partition de V_C en clusters d'antennes cible	
- V_T : Variable temporelle		- V_T^M : discrétisation de V_T en intervalles de temps
	- k_S : nombre de clusters dans V_S^M	- k_S : nombre de clusters dans V_S^M
	- k_C : nombre de clusters dans V_C^M	- k_T : nombre d'intervalles dans V_T^M
	- $k = k_S k_C$: nombre de biclusters	- $k = k_S k_T$: nombre de biclusters
- n_S : nombre d'antennes source	- $n_{i.}^M$: nombre d'antennes source du i^e cluster de la partition V_S^M	- $n_{i.}^M$: nombre d'antennes sources du i^e cluster de la partition V_S^M
- n_C : nombre d'antennes cible	- $n_{.j}^M$: nombre d'antennes cible du j^e cluster de la partition V_C^M	
- m : nombre total d'appels		
- $m_{i..}$: nombre d'appels sortant de l'antenne source v_i	- $m_{i..}^M$: nombre d'appels sortant du i^e cluster de la partition V_S^M	- $m_{i..}^M$: nombre d'appels sortant du i^e cluster de la partition V_S^M
- $m_{.j.}$: nombre d'appels entrant dans l'antenne cible v_j	- $m_{.j.}^M$: nombre d'appels entrant dans le j^e cluster de la partition V_C^M	- $m_{..t}^M$: nombre d'appels passés pendant le t^e intervalle de temps
- m_{ijt} : nombre d'appels passés de l'antenne v_i à l'antenne v_j au temps v_t	- $m_{ij.}^M$: nombre d'appels passés du i^e cluster d'antennes source vers le j^e cluster d'antennes cible	- $m_{i.t}^M$: nombre d'appels passés du i^e cluster d'antennes source pendant le t^e intervalle

TAB. 1: *Notations.*

1. Outil téléchargeable sur `www.khiops.com`

L'analyse que nous menons est divisée en deux phases. Une première s'intéresse aux corrélations entre antennes source et cible alors que la seconde se focalise sur la dimensions temporelle des appels. C'est pourquoi nous introduisons deux modèles distincts : l'un sera dit spatial $\mathcal{M}_S$ et l'autre temporel $\mathcal{M}_T$. Dans les deux cas, l'approche MODL cherche à déduire les paramètres du modèle $\mathcal{M}_S$ (resp. $\mathcal{M}_T$) à partir des données $\mathcal{D}$.

Dans un premier temps, les deux variables étudiées sont catégorielles, il s'agit des antennes source et cible. Le but de l'étude est de grouper les antennes source dont les appels sont distribués de manière similaire sur les antennes cible et vice-versa. Dans un second temps, une variable est catégorielle, les antennes source, et l'autre numérique, le temps. On connait le nombre d'appels sortant quotidiennement des antennes. On va donc chercher à grouper les antennes et discrétiser en même temps la variable temporelle de manière à ce que le trafic sortant des groupes d'antennes soit stationnaire à l'intérieur de chaque intervalle de temps.

L'approche MODL se base sur l'optimisation d'un critère pour générer la structure de coclustering. La construction du critère, ainsi que l'algorithme d'optimisation et les propriétés asymptotique de l'approche sont détaillés dans Boullé (2011) pour le cas d'un coclustering à deux dimensions catégorielles et dans Boullé (2012) pour le cas de données mixtes, i.e numériques et catégorielles. Il s'agit d'un critère construit suivant une approche MAP (Maximum A Posteriori), constitué d'une probabilité a priori (ou prior) sur le modèle de coclustering et de la vraisemblance du graphe connaissant les paramètres du modèle :

- **le prior** : noté $P(\mathcal{M}_S)$ (resp. $P(\mathcal{M}_T)$), il pénalise le modèle en spécifiant la distribution a priori des paramètres de ce dernier. Il est construit hiérarchiquement et uniformément à chaque étape afin d'être non-informatif (Jaynes, 2003).
- **la vraisemblance** : Une fois les paramètres du modèle spécifiés, la vraisemblance $P(\mathcal{D}|\mathcal{M}_S)$ (resp. $P(\mathcal{D}|\mathcal{M}_T)$) est définie comme la probabilité d'observer les données initiales connaissant les paramètres du modèle étudié.

Le produit du prior et de la vraisemblance resulte en la probabilité a posteriori du modèle. Le logarithme négatif de cette dernière probabilité est utilisé pour construire le critère.

Définition (Coût du Modèle spatial)**.** *Le modèle spatial $\mathcal{M}_S$, représentation synthétique des données $\mathcal{D}$ est optimal s'il minimise le critère suivant :*

$$
\begin{aligned}
c(\mathcal{M}_S) &= -\log\left[P(\mathcal{M}_S)\right] - \log\left[P(\mathcal{D}|\mathcal{M}_S)\right] \qquad (1)\\
&= \log n_S + \log n_C + \log B(n_S, k_S) + \log B(n_C, k_C) + \log \binom{m+k-1}{k-1}\\
&+ \sum_{c_i \in X_S^M} \log \binom{m_{i..}^M + n_{i.}^M - 1}{n_{i.}^M - 1} + \sum_{c_j \in X_C^M} \log \binom{m_{.j.}^M + n_{.j}^M - 1}{n_{.j}^c - 1}\\
&+ \log m! - \sum_{\substack{c_i \in X_S^M\\ c_j \in X_C^M}} \log m_{ij.}^M! + \sum_{c_j \in X_C^M} \log m_{.j.}^M! - \sum_{v_j \in V_C} \log m_{.j.}!\\
&+ \sum_{c_i \in V_S^M} \log m_{i..}^M! - \sum_{v_i \in V_S} \log m_{i..}!
\end{aligned}
$$

Définition (Coût du Modèle temporel)**.** *Le modèle temporel $\mathcal{M}_T$, représentation synthétique des données $\mathcal{D}$ est optimal s'il minimise le critère suivant :*

$$c(\mathcal{M}_T) = -\log\left[P(\mathcal{M}_T)\right] - \log\left[P(\mathcal{D}|\mathcal{M}_T)\right] \tag{2}$$
$$= \log n_S + \log m + \log B(n_S, k_S) + \log \binom{m+k-1}{k-1} + \sum_{c_i \in X_S^M} \log \binom{m_{i..}^M + n_{i.}^M - 1}{n_{i.}^M - 1}$$
$$+ \log m! - \sum_{\substack{c_i \in X_S^M \\ c_t \in X_T^M}} \log m_{i.t}^M! + \sum_{c_t \in V_T^M} \log m_{..t}^M! + \sum_{c_i \in V_S^M} \log m_{i..}^M! - \sum_{v_i \in V_S} \log m_{i..}!$$

$B(|V_S|, K_S) = \sum_{k=1}^{K_S} S(|V_S|, k)$ est une somme des nombres de Stirling de second ordre, c'est-à-dire le nombre de manières de partitionner $|V_S|$ éléments en k sous ensembles non-vides.

Les deux premières lignes de l'équation 1 et la première de l'équation 2 représentent le prior alors que la dernière représente la vraisemblance dans les deux cas. D'un point de vue théorie de l'information, un logarithme négatif de probabilité correspond à une longueur de codage. Ainsi, le logarithme négatif du prior est la longueur de codage du modèle alors que le logarithme négatif de la vraisemblance est la longueur de description des données pour une paramétrisation du modèle donnée. Minimiser la somme de ces deux termes a donc une interprétation naturelle en terme de MDL (Minimum Description Length) (Grünwald, 2007). D'un point de vue algorithmique, l'optimisation est réalisée à l'aide d'une heuristique gloutonne ascendante démarrant du clustering le plus fin (une antenne par cluster) et réalisant à chaque étape la fusion de clusters qui décroit le plus le critère. Une post-optimisation améliore cette heuristique en effectuant des permutations au sein des clusters. Cet algorithme, de complexité en $\mathcal{O}(m\sqrt{m}\log m)$, est détaillé dans Boullé (2011).

3 Analyse des corrélations spatiales

La première étape est une analyse des appels entre antennes source et antennes cible. On obtient un total 2141 clusters d'antennes source et 2107 clusters d'antennes cible. Le nombre d'antennes par cluster est compris entre huit et neuf, ce qui est très fin. La difficulté est donc de savoir interpréter les résultats. À l'échelle de la France, le nombre de clusters ne permet pas d'avoir une vision synthétique du regroupement d'antennes. Cependant, à huit antennes par cluster, des résultats de cette finesse peuvent être un atout pour une analyse locale.

3.1 Analyse à l'échelle nationale

Dans un premier temps, nous nous intéressons à une analyse à l'échelle du pays. Le niveau de finesse du modèle obtenu ne permet pas d'avoir une vue synthétique de la structure de coclustering à l'échelle du pays. C'est pourquoi nous proposons de construire une classification hiérarchique ascendante des clusters. Nous fusionnons pour cela deux à deux les clusters de manière à détériorer le moins possible le critère optimisé afin d'obtenir le modèle le plus probable suite à une fusion (de clusters source ou cible). Ce processus nous permet de simplifier notre modèle de manière maitrisée. Afin de quantifier la perte en terme d'informativité du modèle, nous introduisons une mesure que nous appelons taux d'informativité.

Définition (Taux d'informativité)**.** *Le taux d'informativité τ quantifie l'information conservée par un modèle $\mathcal{M}$.*

$$\tau(\mathcal{M}_S) = \frac{c(\mathcal{M}_S) - c(\mathcal{M}_S^{\emptyset})}{c(\mathcal{M}_S^*) - c(\mathcal{M}_S^{\emptyset})} \tag{3}$$

où $\mathcal{M}_S$ est le modèle simplifié, $\mathcal{M}_S^*$ est le modèle optimal obtenu par optimisation du critère MODL et $\mathcal{M}_S^{\emptyset}$ est le modèle nul, c'est-à-dire le modèle ne comportant qu'un cluster d'antennes source et d'antennes cible.

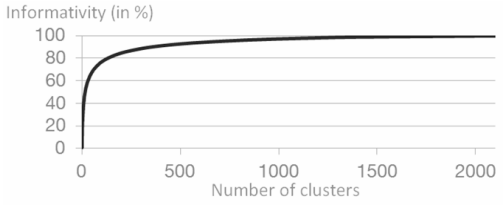


FIG. 1: *Courbe de l'informativité du modèle en fonction du nombre de clusters.*

Cette définition nous permet de construire une courbe du taux d'informativité en fonction du nombre de clusters du modèle simplifié. Cette courbe fait figure de courbe de Pareto du meilleur modèle pour un nombre de clusters donnés. On observe que l'impact des premières fusions sur le modèle est relativement faible. Ainsi, il est possible de réduire le nombre de clusters de plus de 2000 à 85 sur les deux dimensions (antennes source et antenne cible) tout en conservant environ 75% d'informativité du modèle. On utilisera ce niveau de grain pour notre étude à l'échelle nationale, le modèle restant ainsi informatif et suffisemment simple pour être étudié dans sa globalité. Les résultats sont présentés sur la Figure 2. La corrélation entre les clusters d'antennes et leur position géographique est très forte bien que la position des antennes ne soit pas une contrainte dans l'algorithme de coclustering. On en déduit alors que les habitants d'une même zone géographique appellent vers les mêmes endroits. On voit sur cette carte que la France peut être séparée en zones géographiques bien délimitées mais pas nécessairement corrélées avec les frontières des régions administratives ou des départements.

3.2 Une analyse locale

On se propose maintenant d'exploiter la finesse des résultats obtenus. Pour cela, nous utilisons le modèle optimal le plus fin ($\mathcal{M}_S^*$) et nous faisons un zoom sur une métropole française. L'agglomération de Toulouse est divisée en sept principaux clusters, affichés en Figure 3a. Un premier cluster regroupe les antennes du centre-ville (ronds jaune pâle), un autre cluster (ronds vert clair) correspond au quartier de la rive gauche de la ville qui est plus résidentielle que la rive opposée. Le cluster modélisé par des ronds rose pâle correspond au quartier étudiant de Rangueil ainsi qu'à des zones urbaines sensibles, comme Empalot. Les ronds vert pâle couvrent le quartier du Mirail qui possède les mêmes caractéristiques que le précédent cluster. Les ronds orange sont localisés dans des zones périphériques résidentielles

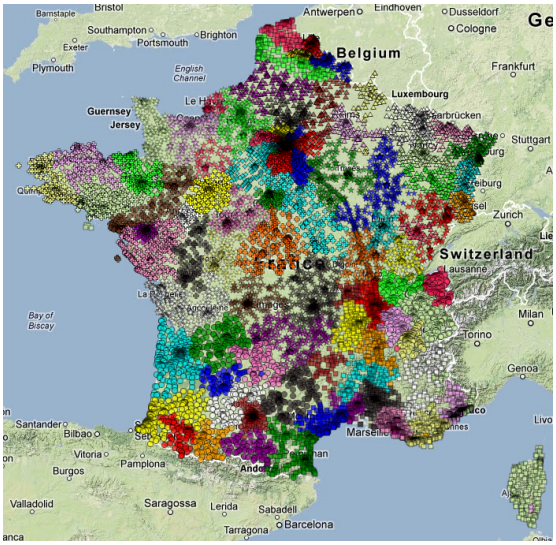


FIG. 2: *Projection des clusters d'antennes source sur une carte de France. Il y a une couleur et une forme pour chaque cluster.*

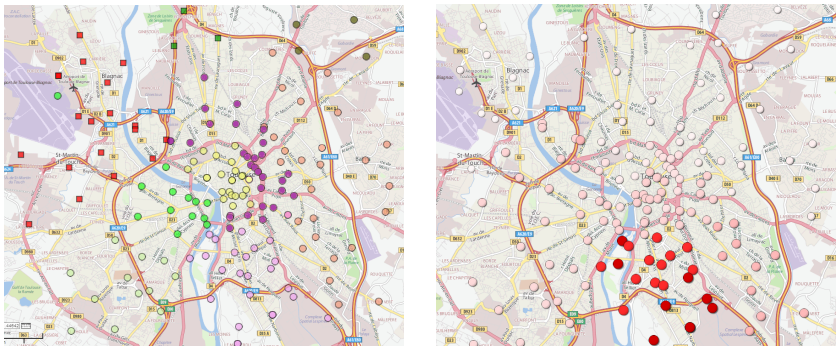
aux caractéristiques socio-économiques relativement diverses, quartiers plus aisés vers le Sud Est et plus populaires vers le Nord-Est. Enfin les carrés rouges couvrent la commune de Blagnac qui regroupe plusieurs zones d'activités de l'agglomération Toulousaine.

Afin de mieux comprendre les raisons de ce regroupement d'antennes source, on s'intéresse à la distribution des appels sortants de ces groupes. L'information mutuelle quantifie la dépendance entre deux variables, ici les antennes source et cible. Cette mesure, notée MI, est définie de la manière suivante (Cover et Thomas, 2006) :

$$MI(C_S, C_T) = \sum_{c_S \in C_S} \sum_{c_T \in C_T} mi(c_S, c_T) = \sum_{c_S \in C_S} \sum_{c_T \in C_T} p(c_S, c_T) \log \frac{p(c_S, c_T)}{p(c_S)p(c_T)} \quad (4)$$

L'information mutuelle est nécessairement positive. Cependant la contribution à l'information mutuelle $mi(c_S, c_T)$ d'un couple de clusters c_S, c_T d'antennes source/cible peut être positive ou négative suivant que la probabilité jointe observée $p(c_S, c_T)$ est supérieure ou inférieure au produit des probabilités marginales des clusters $p(c_S)p(c_T)$, probabilité attendue en cas d'indépendance. L'utilisation d'une telle mesure permet de quantifier l'excès ou le déficit de trafic entre deux groupes d'antennes par rapport à la quantité attendue. Ceci est illustré par la Figure 3b où le trafic depuis le cluster rose pâle de la Figure 3a est étudié. Les antennes en rouge sont les antennes vers lesquelles on observe un excès de trafic ($p(c_S, c_T) > p(c_S)p(c_T)$) alors que les antennes tendant vers le blanc indiquent des antennes vers lesquelles le trafic observé correspond au trafic attendu ($p(c_S, c_T) \approx p(c_S)p(c_T)$ ou $p(c_S, c_T) \approx 0$). Pour ce cluster d'antennes source, il n'y a pas de déficit de trafic ($p(c_S, c_T) < p(c_S)p(c_T)$) vers un groupe d'antennes cible qui apparaitraient en bleu le cas échéant. Il est à noter que les couleurs de cette carte représentent l'information mutuelle, et non la quantité de trafic qui, elle, est représentée par la taille des points : le rayon du point est proportionnel au logarithme du trafic

depuis le cluster de Rangueil (en rose-pâle sur la Figure 3a) vers l'antenne représentée par le point. Ainsi, on observe, pour le cluster étudié, un important excès de trafic vers lui-même et un excès plus léger vers le reste de Toulouse. Pour autant, cela ne signifie pas que le trafic se fait exclusivement au sein du cluster.



(a) *Projection des clusters d'antennes source sur l'agglomération Toulousaine, il y a une couleur et une forme par cluster.*

(b) *Information mutuelle entre le cluster correspondant au quartier étudiant de Rangueil et les clusters tracés.*

FIG. 3: *Étude locale à la ville de Toulouse.*

4 Analyse des corrélations spatio-temporelles

Dans cette seconde étude, nous proposons d'effectuer un coclustering des antennes source et de la variable temporelle. Les données sont donc constituées de 17895 antennes émettrices et de 1,12 milliards d'appels enregistrés sur 5 mois avec une précision à la journée. Le groupement d'antennes source est différent du groupement obtenu dans la Section 3. Ici les antennes groupées entre elles sont similaires car leurs hausses et baisses de trafics sortants se font sur les mêmes périodes temporelles. Nous obtenons 6129 clusters d'antennes source et 117 intervalles de temps. Contrairement à l'analyse du trafic inter-antennes, les clusters sont dispersés sur l'ensemble du territoire Français, ce qui rend une projection sur une carte ininterprétable et ce, pour n'importe quel niveau de grain du modèle. Pour mieux comprendre les phénomènes qui ont mené à une telle structure du coclustering, nous allons étudier l'information mutuelle entre les clusters d'antennes sources et les périodes de temps trouvées. Pour visualiser cette information, nous avons simplifié de modèle de la même manière que dans l'étude précédente et nous avons tracé un calendrier des excès et déficits du trafic sur la Figure 4.

On observe du 13 Mai au 5 Juillet, ainsi que du 1er Septembre au 13 Octobre, une discrétisation régulière et périodique qui correspond au découpage semaine/weekend. On voit qu'en semaine le trafic est en excès pour le cluster du milieu et en déficit pour le cluster du bas, le contraste est d'autant plus fort entre ces clusters qu'on étudie une période éloignée des vacances d'été. Pour les weekend, la tendance s'inverse mais dans une moindre mesure. Le trafic est donc mieux équilibré entre les différents clusters d'antennes. On peut expliquer ces phénomènes par l'activité concentrée sur des zones géographiquement restreintes et généralement urbaines

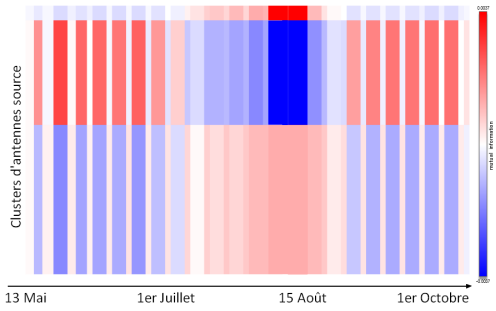


FIG. 4: *Calendrier des excès et déficits d'appels pour* 3 *clusters d'antennes source et* 42 *plages horaires.*

pendant la semaine. Le cluster du haut reste, quant à lui, toujours en déficit en dehors de l'été, bien que ce déficit ait tendance à être plus léger pendant les weekends.

Pendant la période estivale, l'alternance semaine/weekend disparait. On observe alors un déficit d'appels sur toute la période dans le cluster du milieu, alors que le cluster du haut voit un excès significatif d'appels par rapport au trafic habituel et par rapport au trafic de la période. Quant à la dernière zone, on y observe un léger excès d'appels sortant. C'est à cette période de l'année que les contrastes sont les plus forts, c'est pourquoi on va s'y intéresser et tracer une projection géographique des clusters d'antennes sources sur une carte de France (Figure 5).

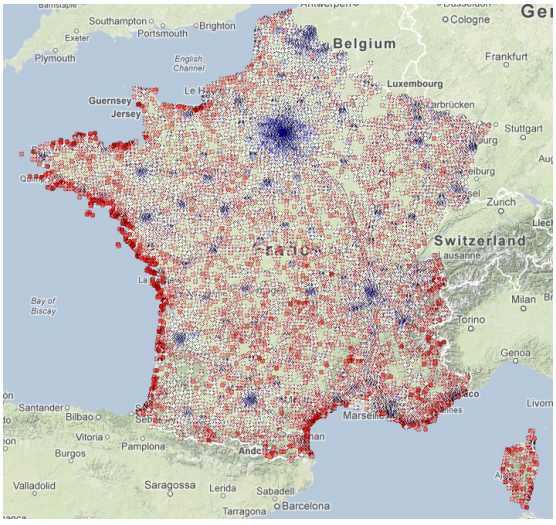


FIG. 5: *Information mutuelle entre les clusters d'antennes et la période du* 5 *au* 15 *Août. En rouge, les excès d'appels, en bleu, les déficit et en blanc, un trafic normal.*

Pendant cette période des vacances scolaires (4 Juillet - 4 Septembre), on observe un excès d'appels au niveau des côtes Atlantiques et du Languedoc-Roussillon principalement. Cela signifie qu'à cette période, un nombre d'appels bien plus important que le trafic habituel a

été émis depuis ces zones. On peut donc qualifier ces zones de saisonniaires car elles sont caractérisées par une répartition des appels très déséquilibrée au cours de l'année, ce qui explique cet excès d'appels à cette période de l'année. C'est pour cette même raison que la Côte d'Azur ne connait pas un excès d'appels aussi fort que la Vendée par exemple. Le trafic y est mieux réparti sur l'année, bien qu'elle soit la première destination estivale des français. Les grandes agglomérations, et notamment Paris, sont colorées en bleu. Ceci peu s'expliquer par la baisse d'activité dans les villes à cette période. On peut donc supposer un déplacement des populations en été, depuis les centres urbains vers les côtes et les zones touristiques. Remarquons tout de même que la couleur tracée sur cette carte n'indique pas la fréquence des appels mais bien l'information mutuelle, les antennes parisiennes demeurant les antennes les plus émettrice, même à cette période de l'année.

5 Conclusion

Nous avons proposé une étude d'un compte rendu d'appels enregistré pendant cinq mois entre les 17895 antennes téléphoniques françaises, ce qui représente un total de $1{,}12$ milliards d'appels. Après avoir présenté des études similaires, ainsi que des méthodes adaptées à ce type d'études, nous avons justifié le choix et détaillé la méthode utilisée : l'approche MODL. Nous avons pu mener deux études de nature différente en utilisant une unique méthode, possédant les propriétés de généricité et de passage à l'échelle suffisantes pour réaliser une analyse complète de nos données. On observe dans cette étude que les antennes groupées dans un même cluster, de part les distributions similaire des appels sortant (resp. rentrant), sont géographiquement très proches et dessinent des frontières précises, aussi bien au niveau national que local. D'autre part, lors d'une étude temporelle, nous avons pu dresser un calendrier de la période temporelle étudiée et déterminer des zones où les appels sortants sont distribués identiquement dans chaque intervalle de temps. Les zones obtenues ont perdu la proximité géographique observée dans la première partie de l'étude mais sont caractéristiques des zones qu'ils décrivent : urbaines, rurales ou touristiques. Les périodes quant à elles, montrent une différence de comportement entre les vacances d'été et les périodes scolaires où on observe une périodicité semaine/weekend où les excès et déficits de trafic s'inversent suivant la nature de la zone. Ainsi, par exemple, les excès de trafic se concentrent dans les zones touristiques en Août et les déficits dans les zones urbaines. Dans des prochains travaux, il serait intéressant de mener une étude où plusieurs dimensions temporelles sont embarquées simultanément dans l'algorithme de coclustering, afin de voir comment se caractérisent les comportements dans les zones géographiques en fonction du jour de la semaine et de l'heure de la journée.

Références

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. 2008*(10), P10008+.

Blondel, V. D., G. Krings, et I. Thomas (2010). Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *the e-journal for academic research on Brussels 42*, 1–12.

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In *Hands-On Pattern Recognition : Challenges in Machine Learning, vol. 1*, pp. 99–130. Microtome.

Boullé, M. (2011). Estimation de la densité d'arcs dans les graphes de grande taille : une alternative à la détection de clusters. In *EGC*, pp. 353–364.

Boullé, M. (2012). Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition 45*(12), 4389–4401.

Cover, T. M. et J. A. Thomas (2006). *Elements of information theory (2. ed.)*. Wiley.

Doreian, P., V. Batagelj, et A. Ferligoj (2004). Generalized blockmodeling of two-mode network data.

Grünwald, P. (2007). *The Minimum Description Length Principle*. Mit Press.

Guigourès, R. et M. Boullé (2011). Segmentation of towns using call detail records. NetMob Workshop at IEEE SocialCom 2011.

Jaynes, E. (2003). *Probability Theory : The Logic of Science*. Cambridge Univ. Press.

Kemp, C. et J. Tenenbaum (2006). Learning systems of concepts with an infinite relational model. In *In Proceedings of the 21st National Conference on Artificial Intelligence*.

Nadel, S. F. (1957). *The Theory of Social Structure*. London : Cohen & West.

Nadif, M. et G. Govaert (2010). Model-based co-clustering for continuous data. In *ICMLA*, pp. 175–180.

Newman, M. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A 103*(23), 8577–8582.

Nowicki, K. et T. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*, 1077–1087.

Reichardt, J. et D. R. White (2007). Role models for complex networks. *The European Physical Journal B 60*, 217–224.

White, H., S. Boorman, et R. Breiger (1976). Social structure from multiple networks : I. blockmodels of roles and positions. *Am. J. of Sociology 81*(4), 730–80.

Summary

For the last few years, the amount of data has significativly increased in the companies. It is the reason why data analysis methods have to evolve to meet new demands. In this article, we introduce a practical analysis of a big database from a telecommunication operator. The problem is to segment a territory and characterize the retrieved areas owing to its unhabitant behaviour in terms of mobile telephony. We have call detail records built during five months in France. We propose a two stages analysis. The first one aims at grouping source antennas which originating calls are similarly distributed on target antennas and vice-versa. A geographic projection of the data is used to display the results on a french map. The second stage discretizes the time into periods between which we note changes in distributions of calls emerging from the clusters of source antennas. This enables an analysis of temporal changes of unhabitants behaviour in every area of the country.