Exploration de données géographiques et externes en MRH

Alexandre Gerbeaux, Guillaume Beraud Sudreau

22 août 2014

***Abstract*— Claims analysis of housing insurance contract leads for a high geographical persistance. Using open data and geocoded historical claims, we will try to capture the geographical risk of housing insurance in a local level. This paper presents the data mining part of our model, including features creation.**

I INTRODUCTION

Cette étude vise à la création de variables capables de capturer le risque géographique propre au client du produit Multirisques-Habitation d'AXA France (hors variable contrat).

Notre base de données client est celle du produit MRH de l'année 2010, uniquement sur les appartements. Cette base est géocodée.

Afin de récupérer des données externes précises et fiables capables d'aboutir à des variables explicatives de bonne qualité, nous avons basé notre étude sur Paris et la petite couronne.

Environ 130 variables ont été créées dans cette analyse et seront susceptibles d'être utilisé dans notre modèle :

- Famille Densité : ces variables capturent le niveau de densité des clients de notre base de données, il correspond à la distance entre le client AXA et le X^{eme} voisin le plus proche.
- Famille POI : ces variables sont basées sur le référencement de différents points d'intérêts [1] issu d'Open Street Map [2]. Il correspond à la distance entre le client d'AXA et le POI le plus proche.
- Famille Bâtiment : Famille de variables qui ont été construites à partir des caractéristiques de la base de données Immeuble 2010 d'AXA France normalisé par la densité au 1000^{eme} client.
- Famille KNN [3] : ces variables sont directement construites à partir de l'historique de la sinistralité des clients. Deux sous famille de variables de fréquences et de coût-moyen ont été construites.

Les cartes en figure 1 et 2 illustrent les persistances géographiques au niveau des sinistres.

La qualité des variables construites a été discutée suivant deux critères :

- La corrélation classique
- La corrélation de spearman (ou corrélation des rangs)

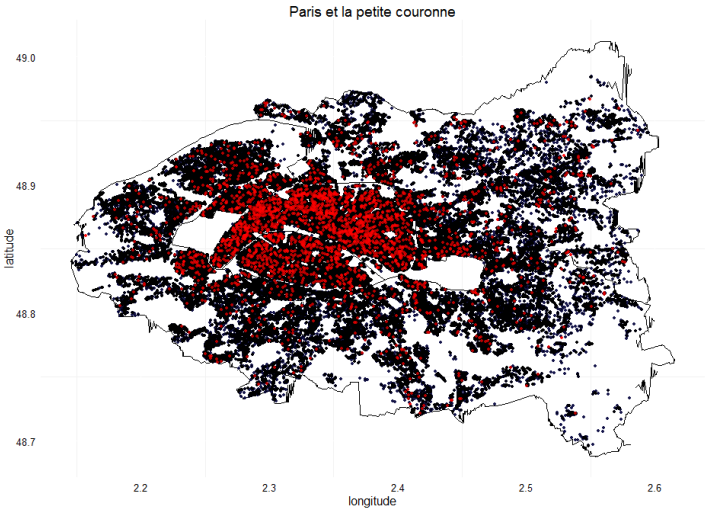

FIGURE 1 – Clients volés chez AXA sur le produit MRH appartement pendant l'année 2010

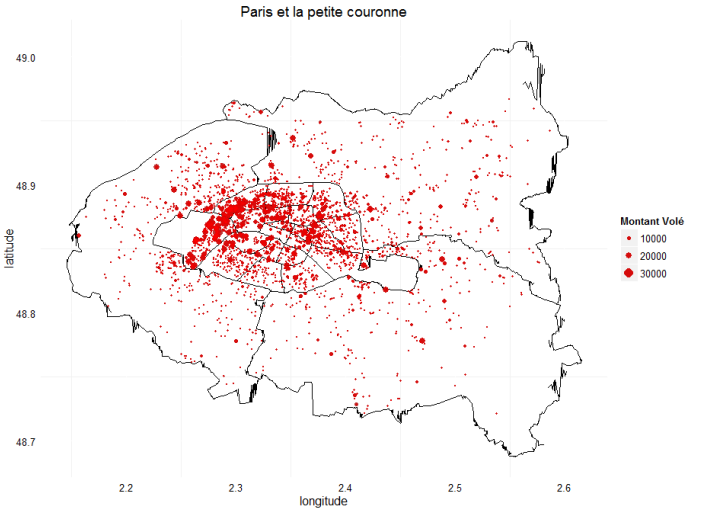

FIGURE 2 – Montant Volé chez AXA sur le produit MRH appartement pendant l'année 2010

1. Exemples : Restaurant, commissariat, bureau de votes...
2. http ://www.openstreetmap.org/
3. k-nearest neighbors

II LES DONNEES

Nous avons privilégié l'approche Fréquence/Coût [4] sur notre base de données clients et pour la création des variables. Nous définissions la fréquence (frequency) et coût moyen (severity) par :

$$frequence_i = \frac{nombre\ d'incidents\ i}{annee\ d'exposition\ i} \text{ pour le } i^{me} \text{ client.}$$

$$Cout\ Moyen_i = \frac{prime\ pure\ observee\ i}{nombre\ d'incidents\ i} \text{ pour le } i^{me} \text{ client [5].}$$

A Densité

Les variables de la famille Densité ont été construites dans le but de capturer l'information sur le voisinage du client.

Nous avons construit :

- 6 variables mesurant la distance entre le client et le i^{eme} voisin le plus proche (i varie de 1 à 1000),
- 8 variables comptant le nombre de voisins pour X mètres (X varie de 50 à 7500).

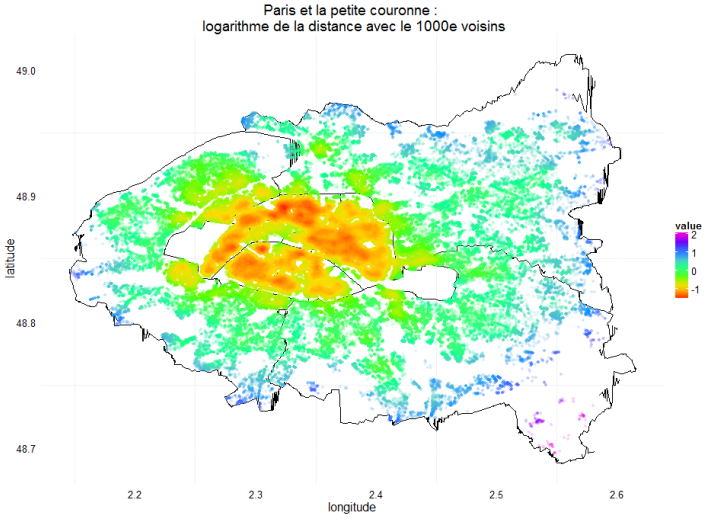

FIGURE 3 – Logarithme de la distance entre le client et son 1000e plus proche voisin

B Bâtiment

Les variables de la famille de Bâtiment exploitent les caractéristiques des contrats sur l'année 2010 de la base bâtiment d'AXA France.

Nous avons construit :

- 4 variables correspondant à la moyenne des dates de constructions des N Bâtiments les plus proches du client (N variant de 1 à 10).
- 4 variables correspondant à la moyenne du nombre de niveaux des N Bâtiments les plus proches du client (N variant de 1 à 10)
- 9 variables correspondant à la moyenne du taux de commerce [6] des N Bâtiments les plus proches du client (N variant de 1 à 100)

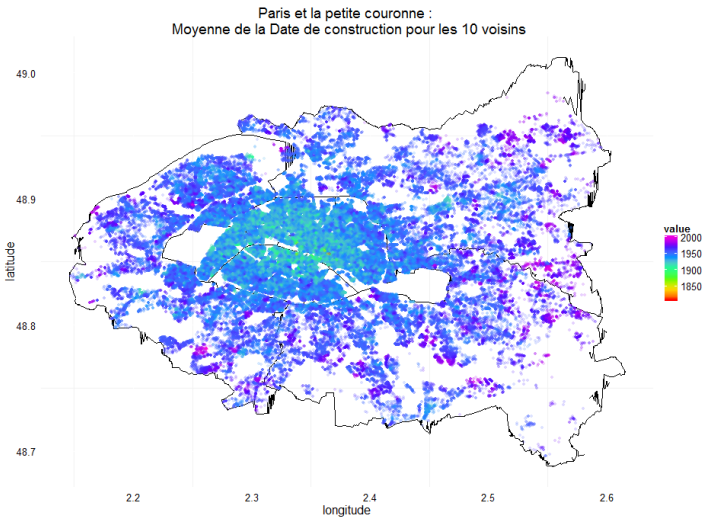

FIGURE 4 – Moyenne des dates de construction de la base Bâtiment 2010 pour les dix voisins les plus proches

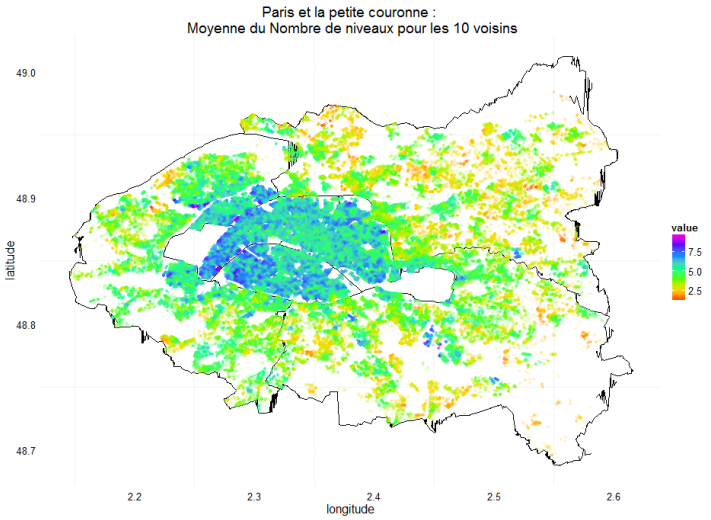

FIGURE 5 – Moyenne des nombres de niveaux de la base Bâtiment 2010 pour les dix voisins les plus proches

4. Méthode qui consiste à voir la prime pure comme le produit d'une variable fréquence et d'une variable coût moyen

5. Attention, dans notre base de donnée, la prime pure observée globale a été divisé par l'exposition client, ce qui n'est pas le cas de la prime pure observée par garantie

6. $taux\ de\ commerce = \frac{nombre\ de\ m^2\ de\ commerce}{nombre\ de\ m^2\ total}$

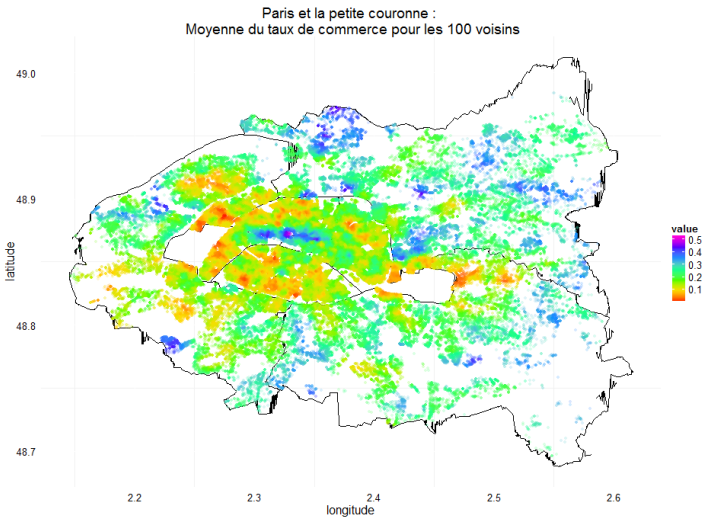

FIGURE 6 – Moyenne des taux de commerce de la base Bâtiment 2010 pour les cents voisins les plus proches

FIGURE 7 – Corrélation entre les features de la famille Bâtiment et les différentes target

FIGURE 8 – Corrélation de Spearman entre les features de la famille Bâtiment et les différentes targets

C KNN

Les variables KNN ont été découpées en fréquence et Coût Moyen. Pour la fréquence, nous avons calculé la moyenne entre le client d'AXA et la fréquence des X voisins les plus proche. Pour le Coût Moyen, nous avons calculé la moyenne entre le client d'AXA et le Coût moyen des X voisins volés les plus proches.

Ces variables prennent en compte la densité de clients AXA puisque ces variables peuvent être vues comme une moyenne de l'information contenue dans un cercle autour du client, et dont le rayon du cercle dépend de l'information disponible dans le voisinage.

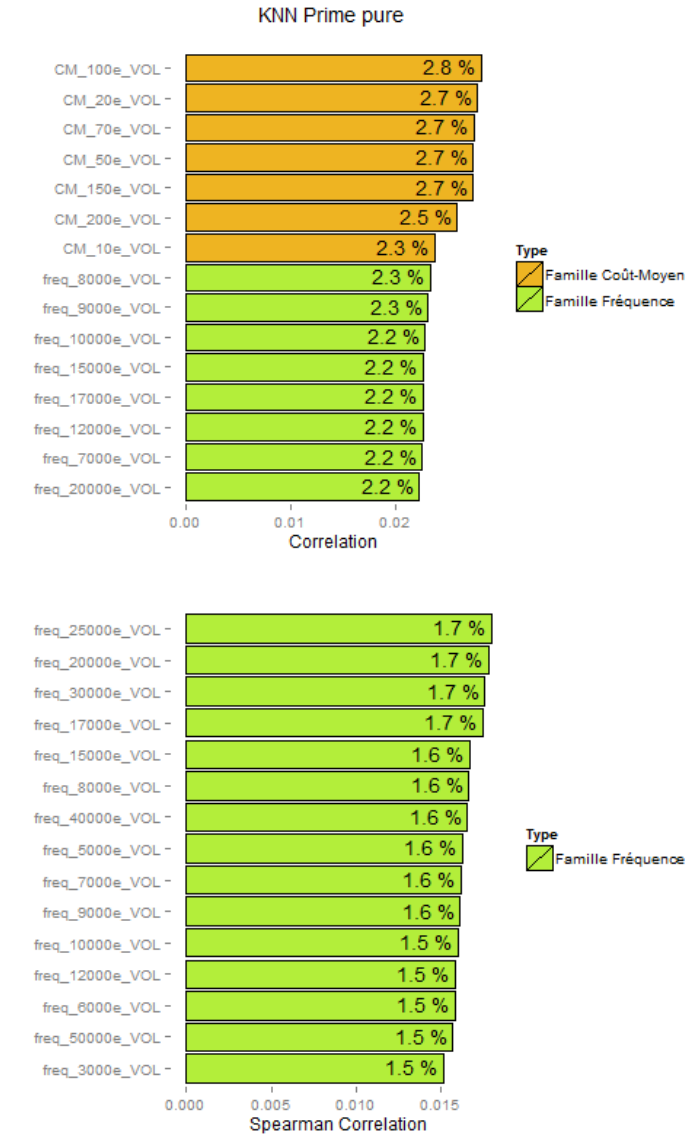

FIGURE 9 – Corrélation et corrélation de Spearman des variables de la famille des Fréquences et Coût Moyen avec la prime pure.

Nous avons confronté ces deux familles de variables face à la corrélation classique et la corrélation de spearman.

C.1 Fréquence

Nous avons balayé de manière logarithmique 28 valeurs de voisinage afin de trouver les optimaux. Il n'est pas étonnant de constater que les valeurs extrêmes (1^e et 20000^e) ne sont pas efficaces. Dans le premier cas nous sommes en présence d'un signal bruité ne contenant pas beaucoup d'information, dans le deuxième cas la variable lisse de manière trop brutale les données (l'information de deux tiers de la base de données est contenue dans un client).

Les variables construites autour d'une plage de valeurs entre un voisinage de 150000 et 300000 semblent efficace du point de vue des deux métriques.

Un léger maximum local autour de 8000 suggère vraisemblablement qu'une autre partie de l'information peut-être capturée avec ces variables, dans un voisinage plus local.

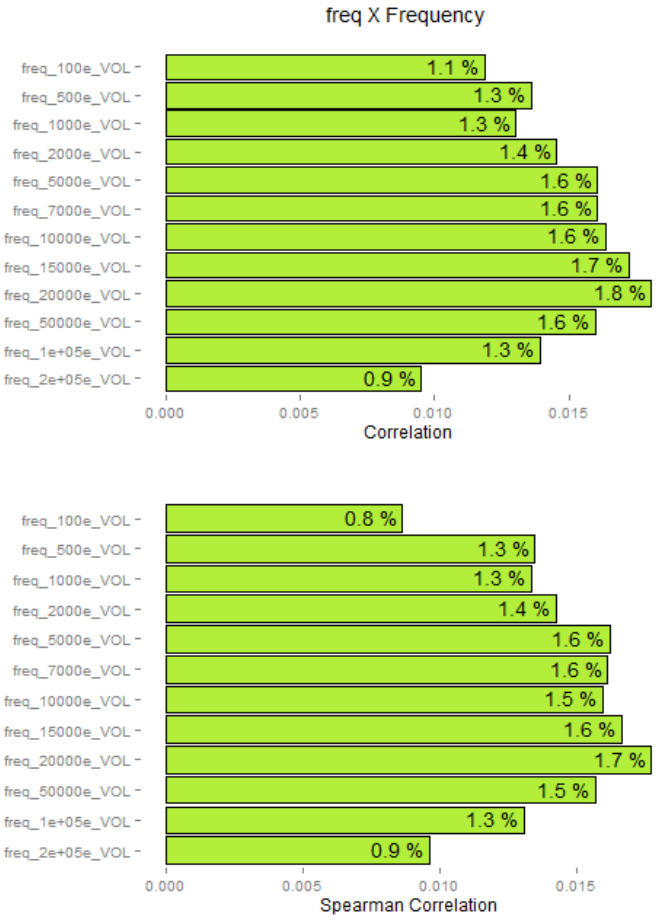

FIGURE 10 – Corrélation et corrélation de Spearman des variables de la famille des Fréquences Moyennes avec la Fréquence.

En annexe, nous trouverons une carte de chaleur des moyennes des fréquences pour les 200000 plus proches voisins. Nous observons un trend nord/sud avec un pic dans le 16^{me} arrondissement.

C.2 Coût Moyen

Comme pour les variables fréquences, nous avons balayé de manière logarithmique 12 variables afin d'observer celle qui ressortent le mieux (figure 3). La variable CM_1e_VOL est un signal bruité, et CM_1000e_VOL lisse de manière trop brutale les données clients. Les variables qui lissent le voisinage entre les 20 et les 150 voisins les plus proches semblent pertinents du point de vue des deux métriques.

Nous avons tracé pour exemple en annexe une carte de chaleur des Moyennes des Coûts Moyens pour les 50 plus proches voisins volés. Nous observons des regroupements globaux notamment dans le 16^{me} et 17^{me} arrondissements.

D POI

Les variables POI ont été construites grâce aux données (géolocalisation et nom du POI) sur open street map. Nous avons retenu les POI qui ont été référencés au moins 50 fois dans Paris et la petite couronne (42 POI). L'idée est de quantifier le quartier du client et de

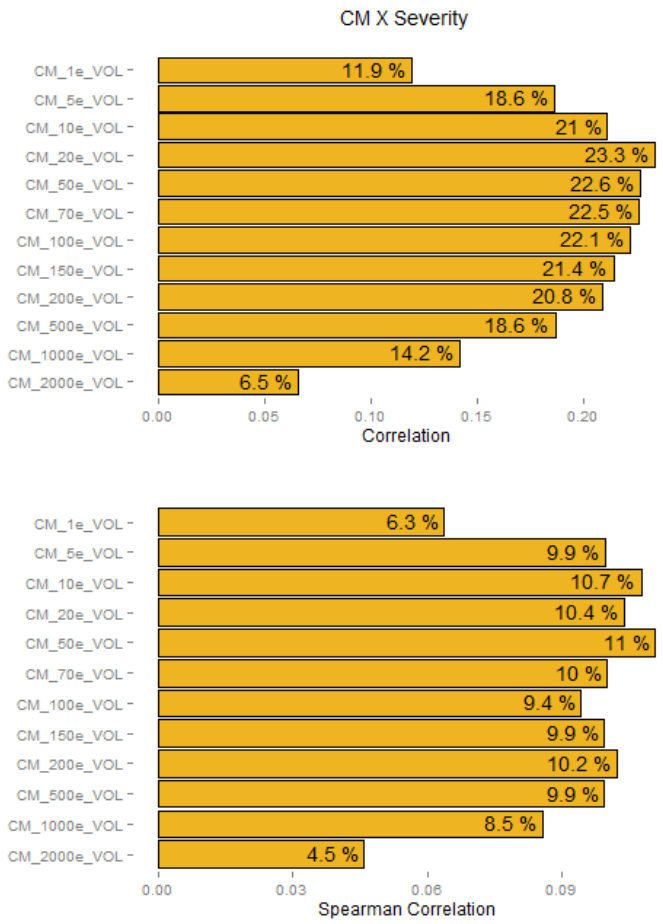

FIGURE 11 – Corrélation et corrélation de Spearman des variables de la famille des Coût Moyens avec le Coût Moyen.

voir si cette information est capable d'expliquer au moins en partie la fréquence et le coût moyen.

Open Street Map est un projet qui vise à référencer de manière libre (basé sur le volontariat) les données géographiques du monde. La fiabilité tant en terme de précision qu'en terme d'exhaustivité peut donc être discutée.

Nous avons créé comme variable la distance entre le client et le POI le plus proche, néanmoins cette distance ne prend pas en compte la densité client. En effet, plus un client est dans un endroit peuplé, et moins l'information éloignée sera susceptible d'avoir une influence sur lui. Pour prendre en compte cet effet, nous normalisons les variables POI par la distance entre le client et son 1000^e voisin le plus proche [7].

En figure 6,7 et 8 nous confrontons les variables POI normalisée face à nos différentes variables cibles selon les deux critères. Les variables $veterinary$,$townhall$ et bar ressortent le plus.

Nous avons mis en annexe les cartes (à gauche les localisations des POI, à droite les cartes de chaleurs associées à la variable construite) des POI les plus efficaces dans notre modèle selon les corrélations.

III CONCLUSION

En résumé nous avons construit :

– 42 variables POI
– 14 variables densité

7. Dans Paris, un nombre non négligeable de clients ont une distance nulle avec leur plus proche voisin, nous choisissons donc un voisinage élevé pour ne pas diviser par zéro.

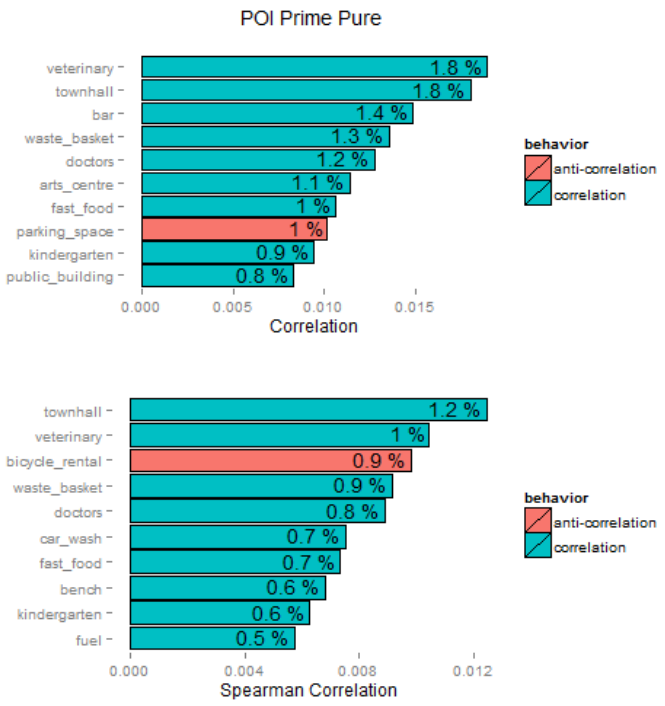

FIGURE 12 – Corrélation et corrélation de Spearman des variables de la famille des POI avec la Prime Pure.

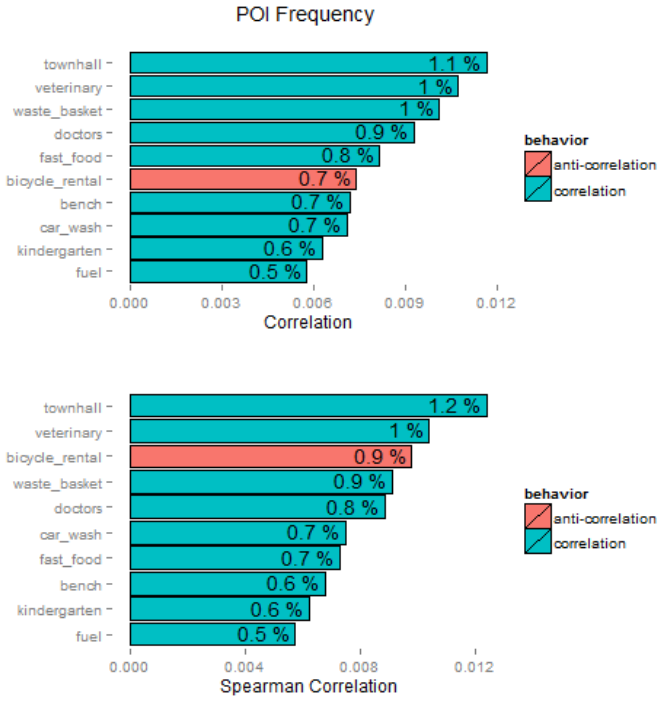

FIGURE 13 – Corrélation et corrélation de Spearman des variables de la famille des POI avec la Fréquence.

– 17 variables bâtiment
– 40 variables KNN (12 variables Coût Moyen et 28 variables fréquences)

Les variables KNN semblent efficaces pour capturer les effets globaux du risque. Les variables issues de la famille POI et bâtiment semblent quand à elles être plus efficaces pour capturer les effets locaux.

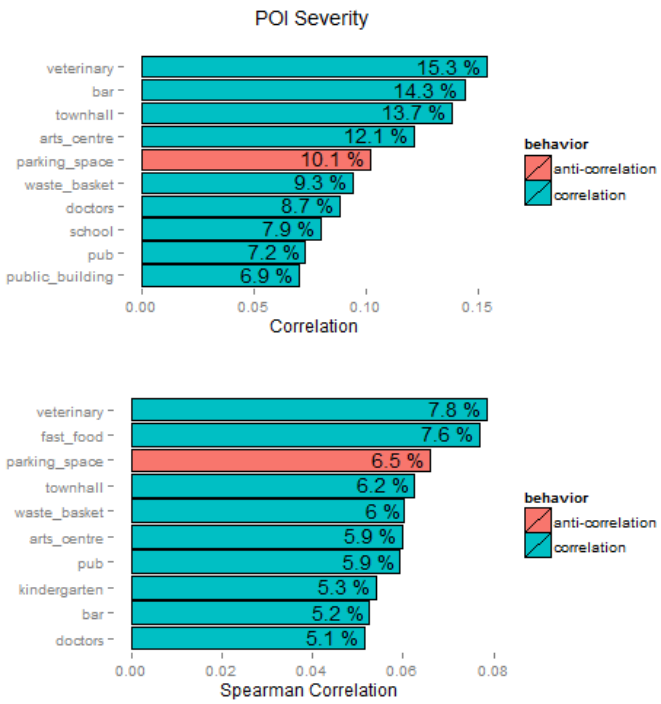

FIGURE 14 – Corrélation et corrélation de Spearman des variables de la famille des POI avec le Coût Moyen.

IV ANNEXE

KNN

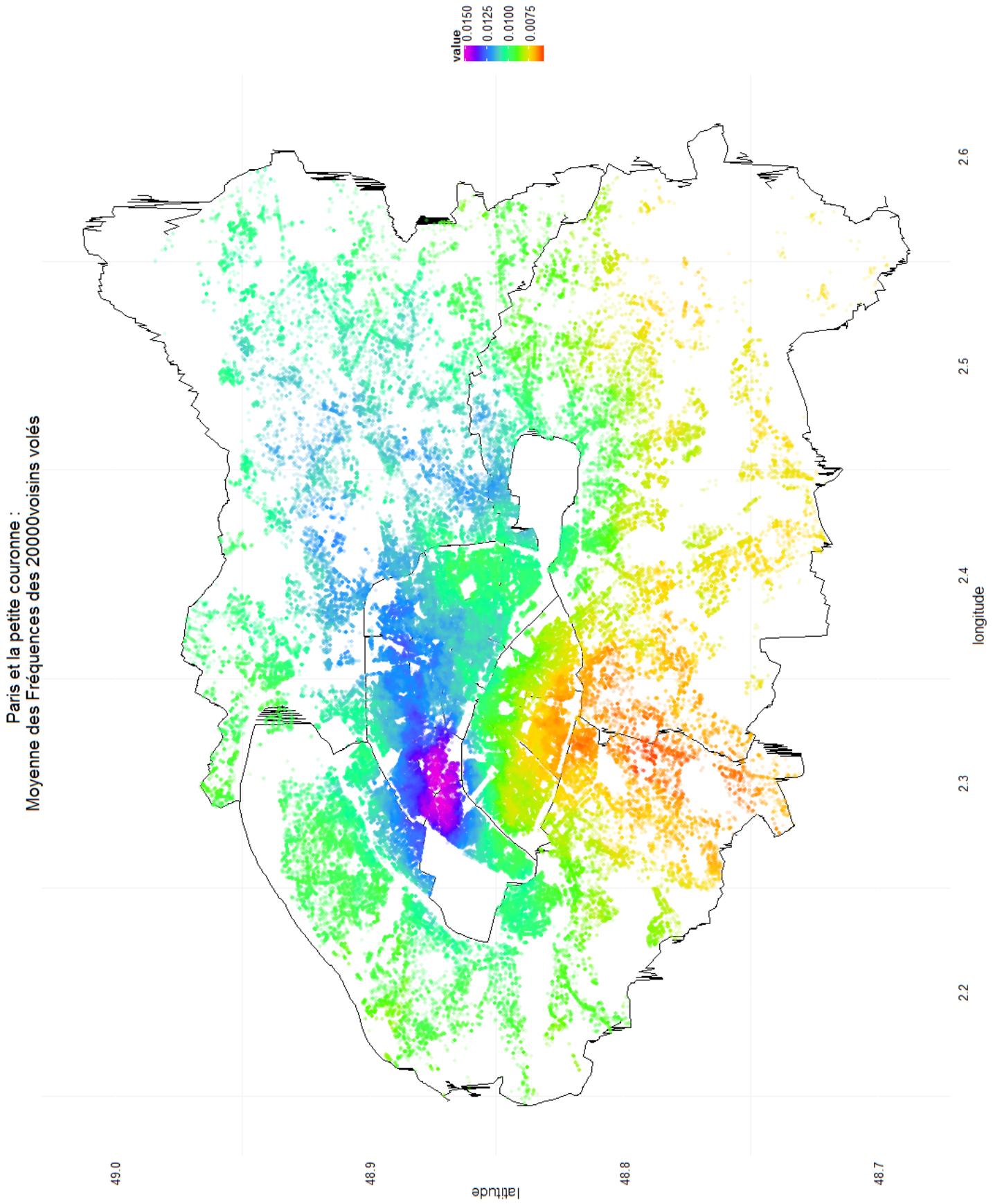

FIGURE 15 – Carte de chaleur des moyennes des fréquences des 20000 voisins.

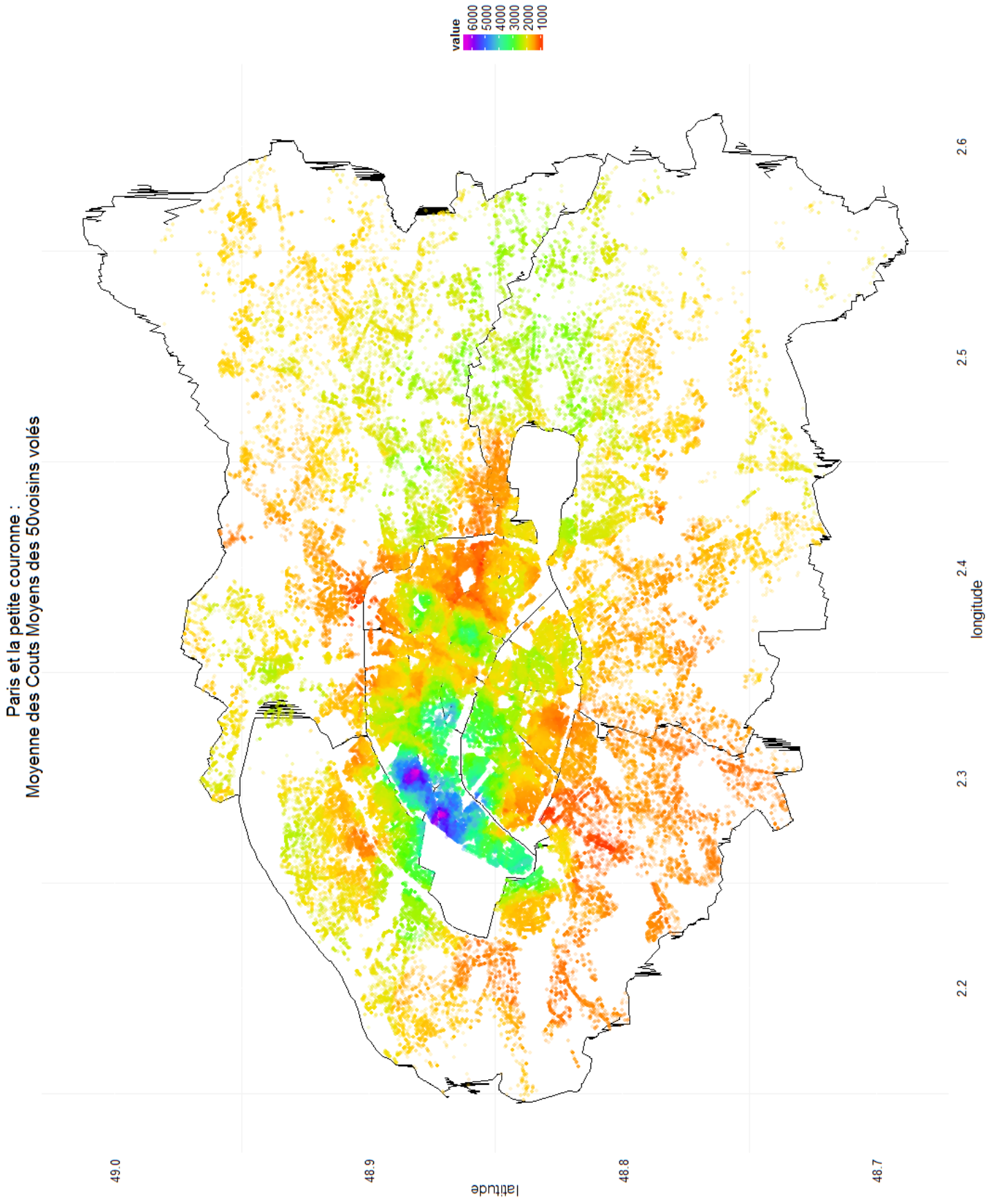

FIGURE 16 – Carte de chaleur des moyennes des Coût Moyens des 50 voisins volés.

POI

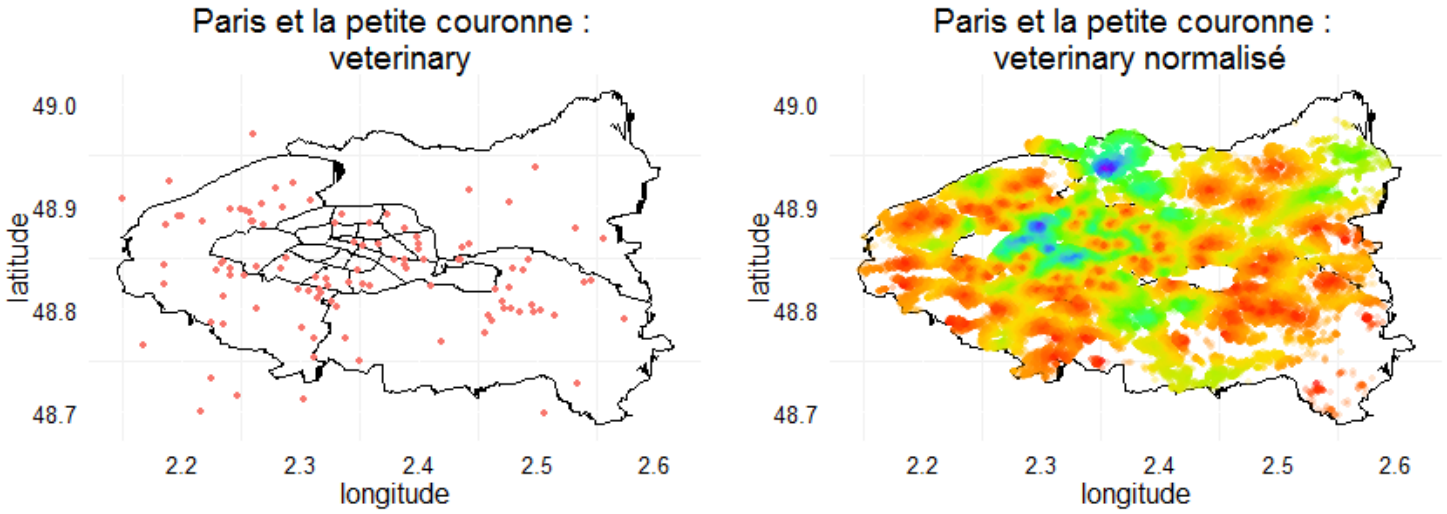

FIGURE 17 – Vétérinaires

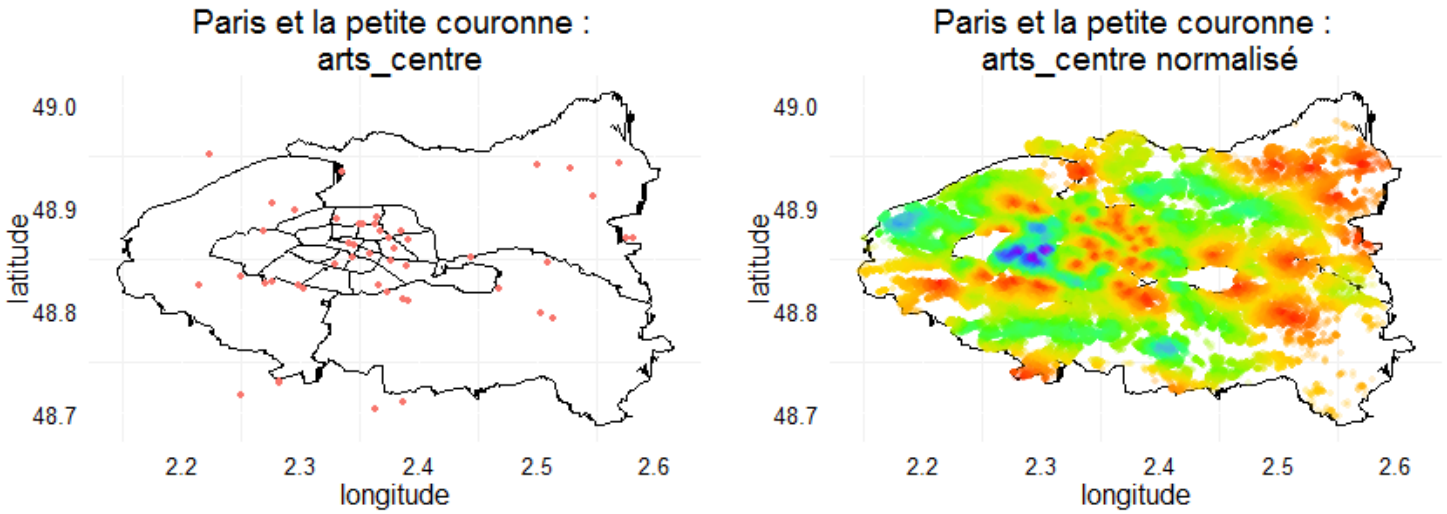

FIGURE 18 – Centres des arts

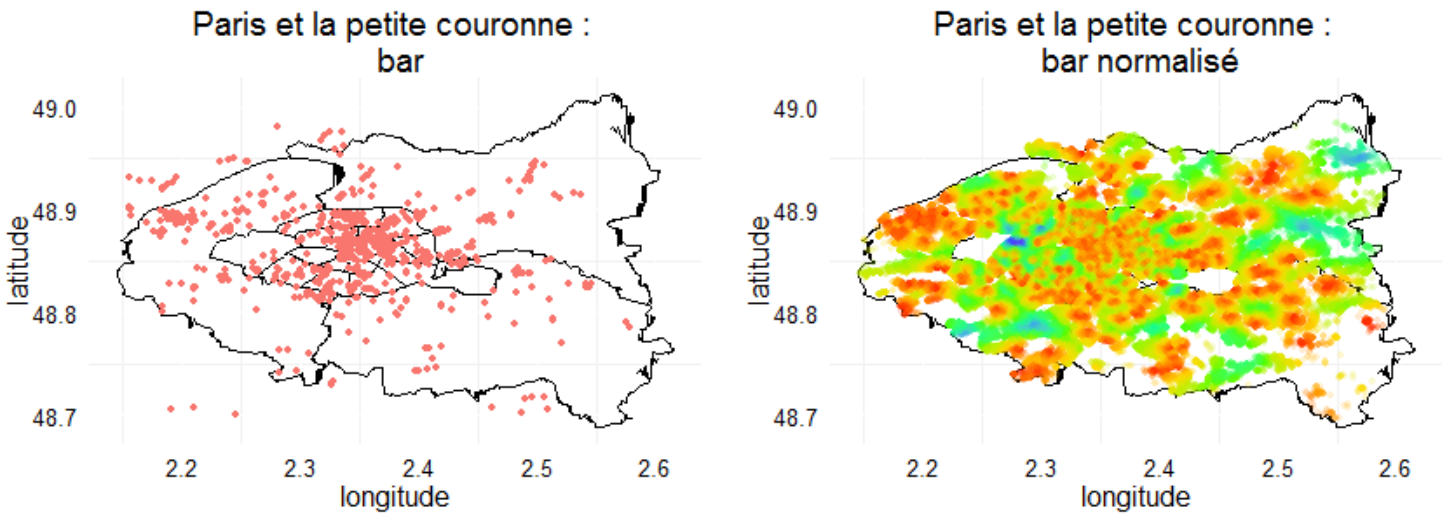

FIGURE 19 – Bars

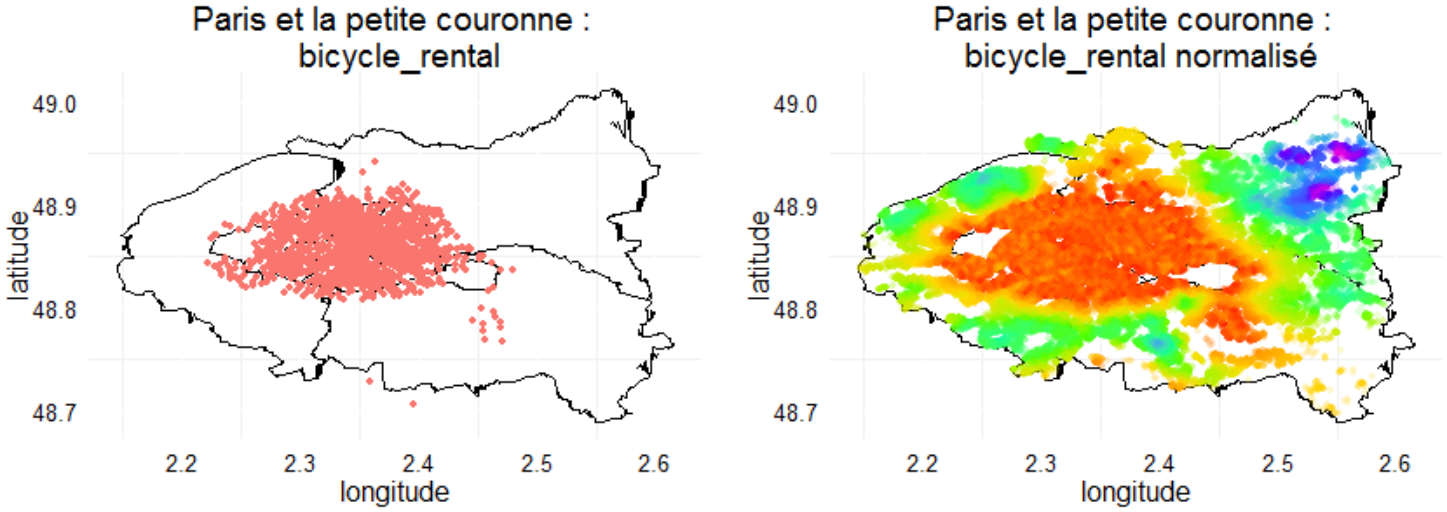

FIGURE 20 – Bornes de location de vélos

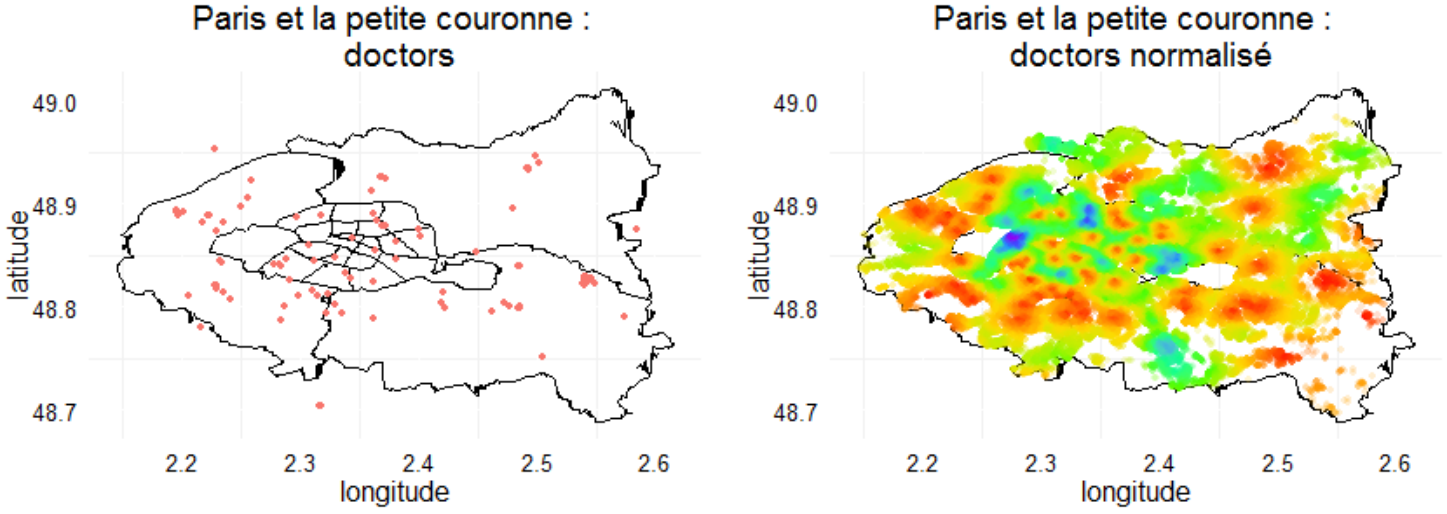

FIGURE 21 – Cabinets de médecin

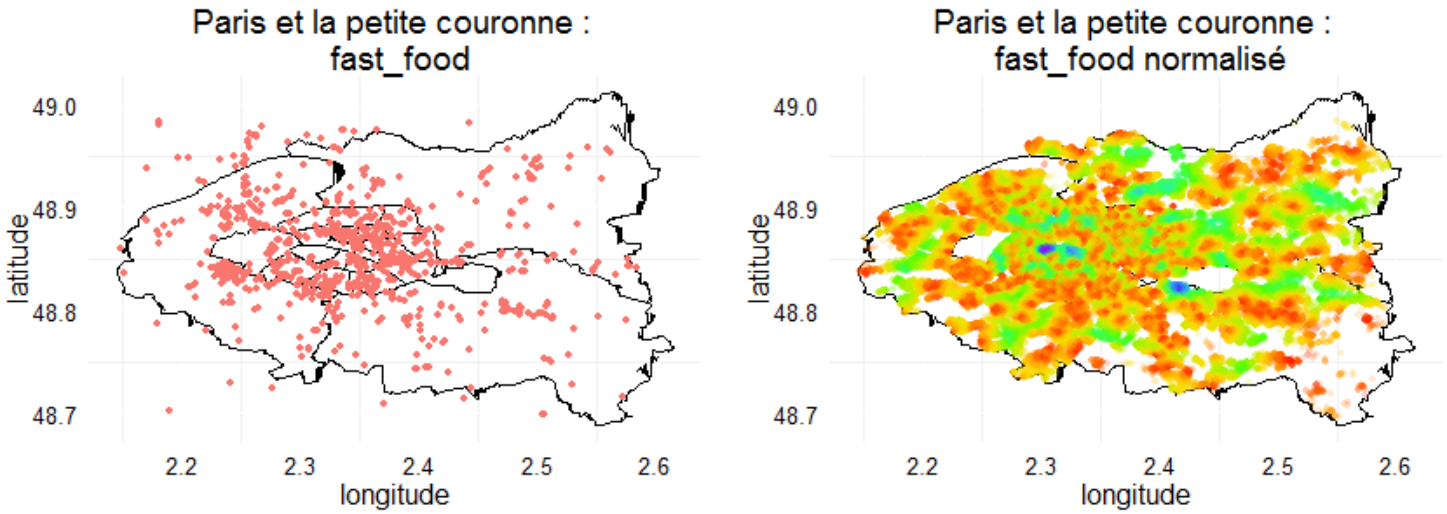

FIGURE 22 – Fast food

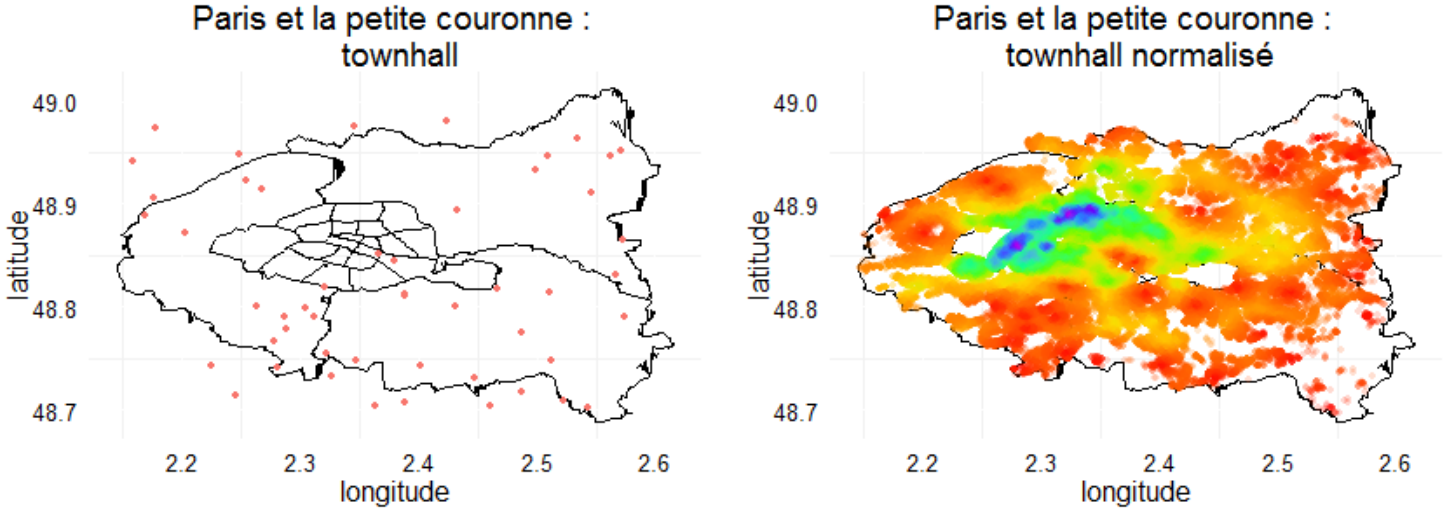

FIGURE 23 – Mairies

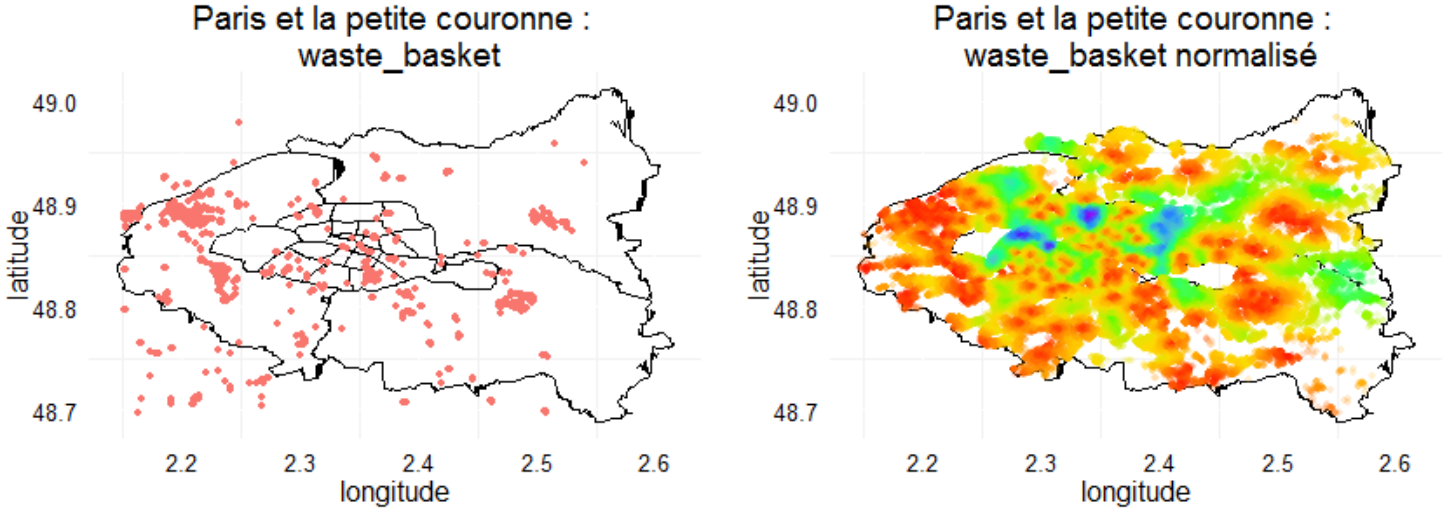

FIGURE 24 – Poubelles

Matrice de Corrélation

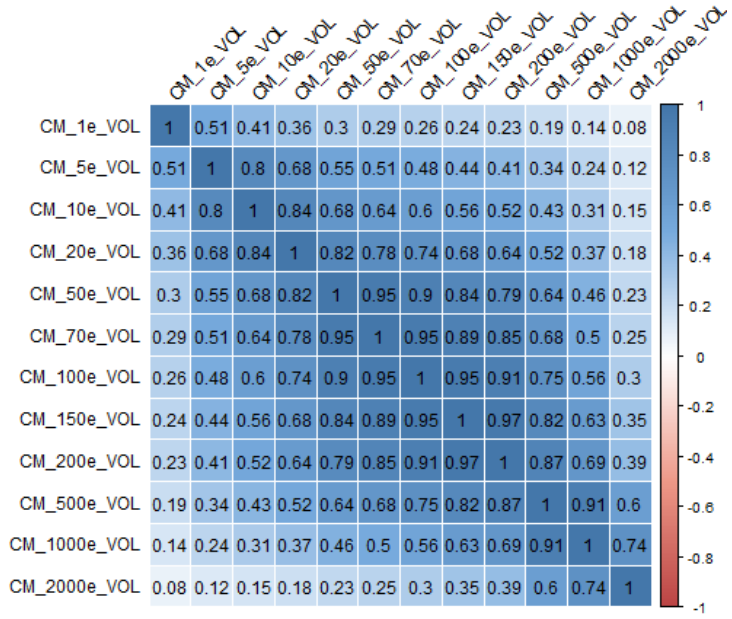

FIGURE 25 – Matrice de Corrélation des features KNN sur le coût Moyen

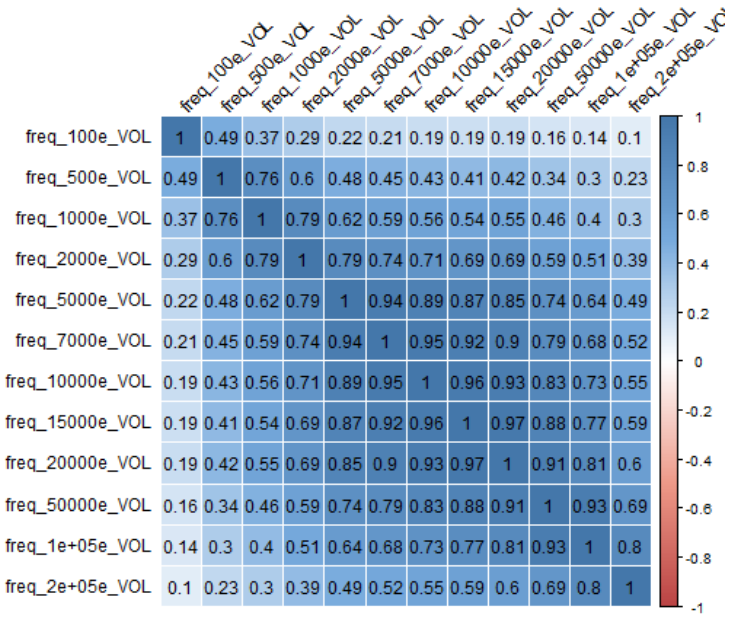

FIGURE 26 – Matrice de Corrélation des features KNN sur le fréquence

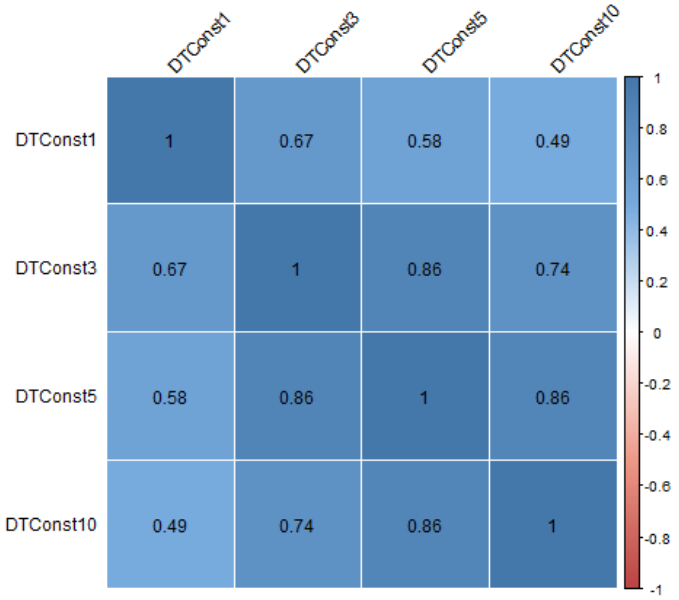

FIGURE 27 – Matrice de Corrélation des features Bâtiment sur les dates de construction

FIGURE 28 – Matrice de Corrélation des features Bâtiment sur le taux de commerce

FIGURE 30 – Matrice de Corrélation des features densité

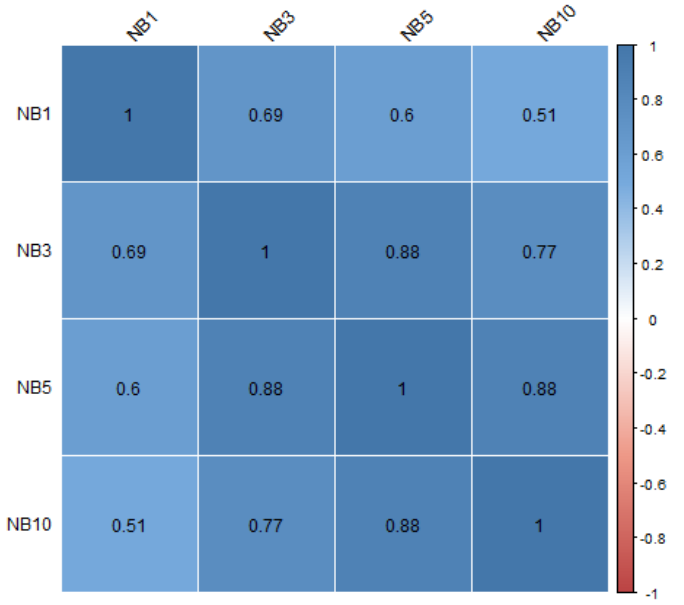

FIGURE 29 – Matrice de Corrélation des features Bâtiment sur le nombre de niveaux

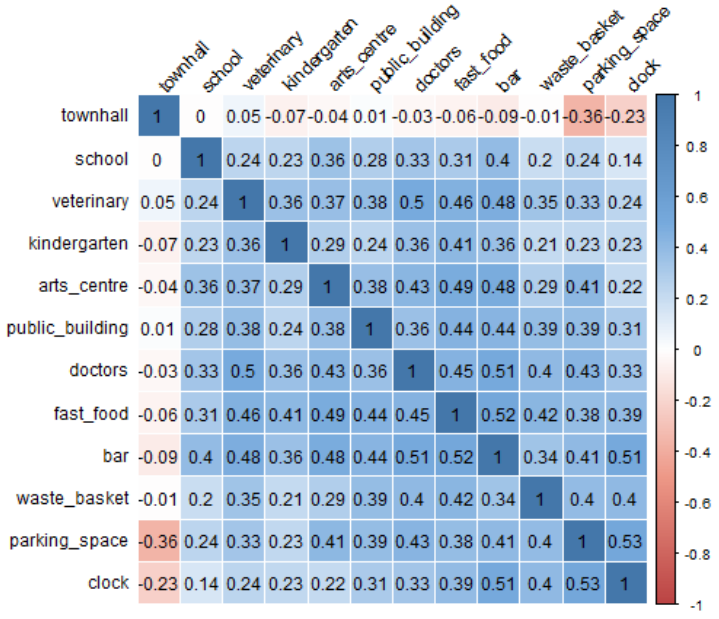

FIGURE 31 – Matrice de Corrélation des features POI des meilleurs POI

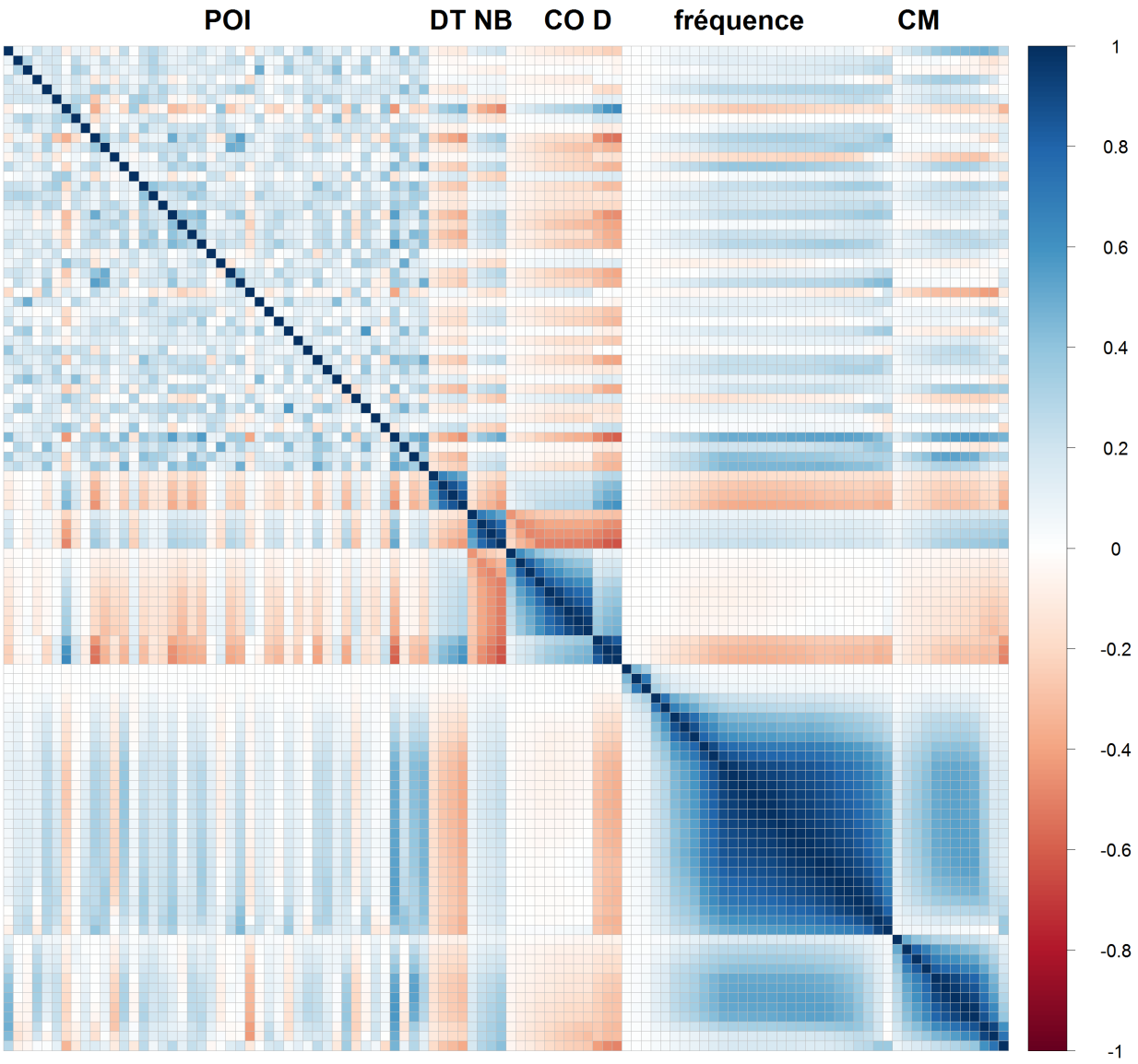

FIGURE 32 – Matrice de Corrélation de toutes les features