AGENIS-NEVERS Marc
Publication : 30/09/2017
Version : 1.1
Repository Gitlab et scripts : gitlab.com/users/agenis/projects

Le paiement du stationnement à Paris, étude d'un jeu de données en *open data*

Contexte

Depuis quelques années, la mairie de Paris a mis en ligne une plateforme de publication d'open data (http://opendata.paris.fr/). Plusieurs centaines de jeux de données y sont présentés sur des thématiques aussi diverses que la fiscalité, les espaces verts en passant par le logement ou les scores électoraux ; certains jeux disposant d'une géolocalisation de leurs mesures. La présente note s'emploie à analyser l'un de ces jeux de données, assez méconnu et pas encore exploité à notre connaissance : les données de paiements des horodateurs parisiens (Mairie de Paris/Direction de la Voirie et des Déplacements 2015) (données sous licence ODbL).

Comme souvent dans l'*open data*, l'analyse de données permet au statisticien d'analyser certains phénomènes et comportements humains au moyen d'informations indirectes et/ou recoupées. A travers le paiement du stationnement de leurs véhicules, les parisiens nous renseignent sur leurs habitudes et horaires de travail, le type de véhicule, mais également l'existence éventuelle de stratégies de non-paiement (obéissance à la règle) éventuellement liées à un positionnement géographique. Les données sont recoupées avec une source externe, la météo, qui sera testée comme facteur influençant le paiement. Cette dernière source est fournie par la start-up METIGATE (http://metigate.com/) qui propose des services dans la modélisation statistique de la météo.

Méthodologie

Fonctionnement du stationnement à Paris

Il est impossible de comprendre un tel jeu de données si on ne se plonge pas dans le fonctionnement particulier du stationnement à Paris. Il faut distinguer en premier lieu le type de stationnement : résidentiel (un tarif préférentiel est appliqué aux détenteurs d'une carte de stationnement et justifiant d'une immatriculation en 75) ou rotatif, ce dernier étant accessible aux autres usagers pour des durées ; il est évidemment plus onéreux (coût horaire 12 fois plus élevé).

Mode rotatif

Le stationnement est payant de 9h à 19h du lundi au vendredi, ainsi le samedi dans certaines rues. Sont exemptés les dimanches, fériés et mois d'août. A noter qu'après 2015 les samedis et le mois d'août sont devenus payants. La durée maximale payable est 2 heures. Il est possible d'anticiper le paiement (avant 9h ou après 19h).

Mode résidentiel

Ce type de stationnement n'est plus limité à 2 heures mais peut être payé pour toute une journée, voire une semaine entière (comptée comme 50 heures). Du fait du faible coût et de l'horizon temporel, on s'attend à ce que l'heure du paiement soit répartie de façon assez indépendante des horaires de travail.

Types de paiement

Le paiement peut se faire par carte bleue à la borne, ou par carte prépayée. Le paiement en espèce n'est plus possible sur le parc d'horodateurs moderne en place en 2014 (qui permet justement la remontée complète des données). L'heure exacte du moment du paiement (jusqu'à la seconde) est enregistrée.

Zones tarifaires et secteurs

La ville de Paris est découpée en zones tarifaires et en secteurs résidentiels. Les arrondissements du cœur de Paris voient leur tarif rotatif doublé (4€/heure au lieu de 2.40). Le découpage en secteurs résidentiels définit 160 zones où un résident aura droit au tarif résidentiel ; s'il se gare en dehors de sa zone attribuée, il devra s'acquitter du tarif rotatif. Certains emplacements (13% des parcmètres) n'autorisent pas le paiement résidentiel, ce sont pas exemple des axes très commerçants ou touristiques (Figure 1).

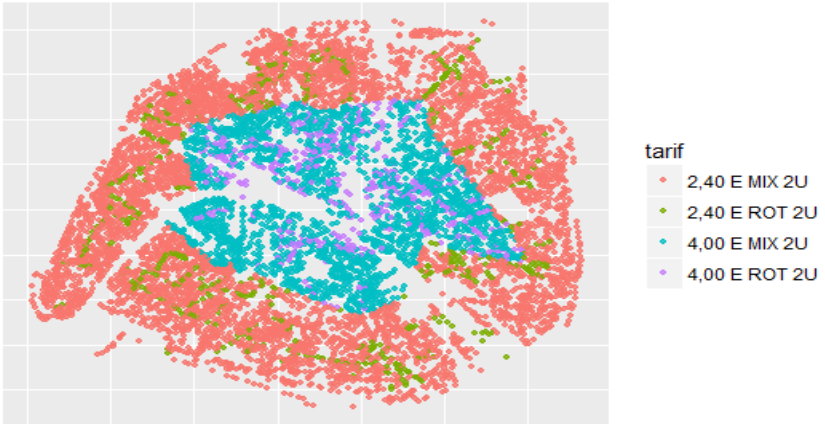


Figure 1: zones tarifaires de parcmètres

Caractéristiques des horodateurs

Dans le jeu de données sur les mobiliers, on trouve des informations internes au horodateurs comme le modèle, le type d'alimentation (pile ou solaire), les coordonnées physiques.

Autres particularités

Pour des raisons évidentes d'anonymisation, les paiements ou les véhicules ne sont pas identifiables et il est donc impossible d'attribuer des paiements à un même utilisateur.

Choix de la source de données météo

Les historiques détaillés de séries climatiques en un point donné sont en accès restreint ; on peut les acheter sur des sites comme Météo France, ou utiliser des données libres comme celles des aéroports fournies par Weather Underground via une interface d'accès (API). Il n'y a toutefois aucune garantie de fiabilité sur cette dernière option. Une solution pour récupérer les données météorologiques de Paris à pas horaire consiste à les approximer par ses aéroports (Orly (14km au

Sud), Le Bourget (14km au Nord-Nord-Est), Roissy CDG (23km au Nord-Est)) ; cette approximation pouvant s'avérer risquée dans la mesure où la pluie est un phénomène local. Par la suite, des données détaillées pour la ville de Paris sont avantageusement fournies par la société METIGATE afin d'illustrer un exemple d'usage intéressant d'historiques météo simulés.

Comment définir la précipitation ?

La définition de la précipitation n'est pas aussi simple qu'il n'y parait. Elle peut être considérée comme variable numérique (hauteur tombée en mm journalier) ou catégorielle (pluie/sec à chaque pas horaire). Si on cherche à approximer la précipitation à partir de données d'aéroports, on pourra utiliser respectivement la moyenne de pluie tombée entre Orly et le Bourget, ou la présence de pluie détectée simultanément sur les deux aéroports. Il faut également se demander si on inclut les évènements neigeux ou de grêle dans les précipitations, c'est le cas ici.

A une échelle journalière, des relevés de cumul de précipitation manuels sont transmis 2 fois par jour (variable *precip12*), fournissant une donnée irréfutable mais sur un découpage horaire un peu décalé. La variable horaire *precip* transmet, elle, un cumul sur chaque heure passée. Aussi, lorsqu'on fait correspondre les données de pluie avec les données d'horodateurs, il faut garder à l'esprit que la première est un cumul de l'heure passée, alors que la seconde est mesurée en continu.

Statistiques descriptives

Nettoyage des données

Le jeu de données ne présente aucune valeur aberrante et aucune valeur manquante, ce qui indique qu'il a probablement déjà subi un prétraitement, puisque des données remontées de capteur présentent toujours quelques erreurs. Aussi, seuls les noms de colonnes sont modifiés.

Dans 5% des cas les durées des transactions d'usagers résidentielles présentent des fractions aléatoires d'heures ; elles correspondent probablement à des semaines ou journées déjà entamées et décomptées partiellement.

Nombre de parcmètres

Les informations de paiement permettent de recenser 7922 parcmètres ; toutefois certains ont été installés au cours de l'année 2014. Seuls 7805 ont une correspondance dans le jeu de données géolocalisé disponible également sur open data Paris (Mairie de Paris/Direction de la Voirie et des Déplacements 2017), donc 112 non représentables sur une carte après fusion des deux jeux de données.

Chiffres globaux

Les montants globaux collectés sur l'année 2014 s'élèvent à **62.4 Millions d'euros**, répartis de la façon suivante entre les mois (Figure 2) montrant un pic notable en octobre, même après correction par le nombre de jours payants. L'année totalise **24,4 millions de transactions**.

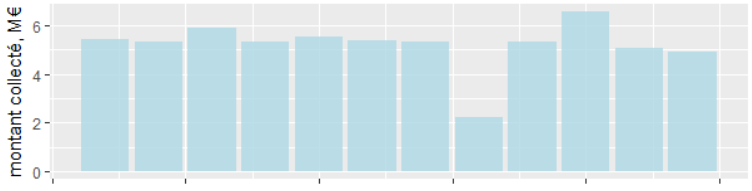


Figure 2: répartition mensuelle (corrigée) des sommes payées

La répartition des recettes selon le type d'usager, ainsi que le nombre de transactions faites, est présentée également (Figure 3). La contribution du stationnement résidentiel n'est que de 18.5% pour 22% des transactions totales.

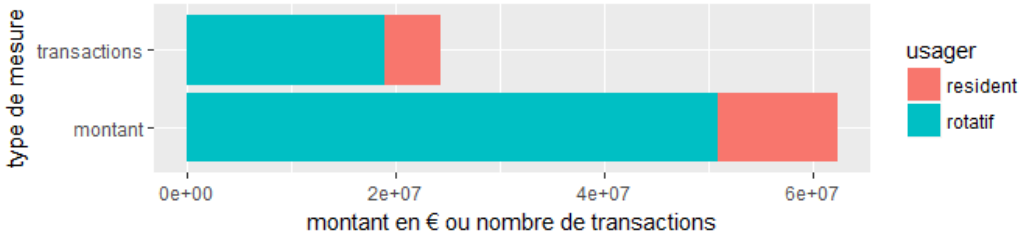


Figure 3: répartition globale des montants et transactions

On peut également ventiler les transactions par arrondissement, montrant dans le trio de tête le 16ème, 15ème et 13ème qui sont également parmi les plus peuplés. Ramené à la population de chaque arrondissement (source), les plus fortes intensités de paiement sont situés dans les 8ème (130€/habitant/an), 7ème et 1er. De l'autre côté du palmarès le 19ème arrondissement contribue à hauteur de 11€/habitant/an à peine. Ces écarts allant de 1 à 10 pourraient s'expliquer par plusieurs facteurs comme le nombre de voitures, le nombre de parkings, l'activité économique du quartier, et bien sûr l'intensité des contrôles (les chiffres d'effectifs de police par arrondissement ne sont pas disponibles).

Au sein d'une journée type, l'évolution globale du nombre de transactions selon le type d'usage (résident ou rotatif) nous renseigne sur les habitudes des parisiens.

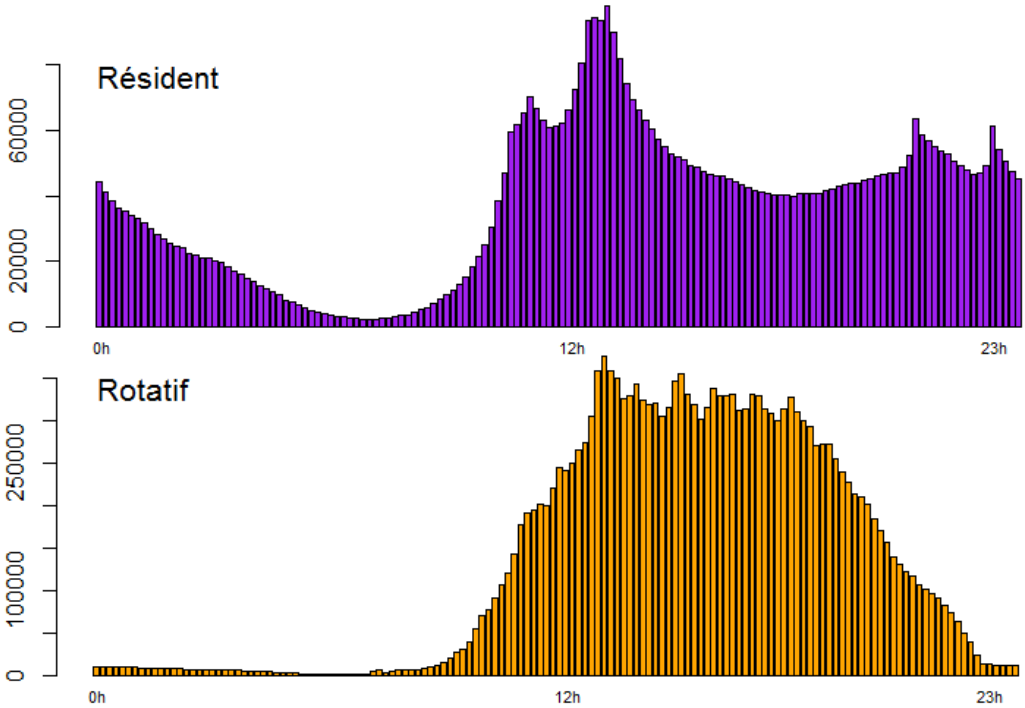


Figure 4: Evolution horaire des paiements, selon l'usager (jour de semaine)

Les deux courbes présentent une forme bien différente : les résidents sont assez régulier au cours de la journée, et même également la nuit avec plusieurs pics marqués : en milieu de matinée (11h), au moment où ils sortent déjeuner (13h), ou bien en sortant (pic de 20h) ou en rentrant d'un diner dehors (pic de 23h). La structure des usagers rotatifs montre une forme en cloche sans écarts majeurs (un pic autour de 12h-13h pour un déjeuner dans une zone hors secteur résidentiel.

Equipements notables, palmarès

Les horodateurs peuvent se classer du moins utilisé au plus « rentable ». A ce titre, les 3 ou 4 de chaque palmarès ont été étudiés à part. Ces résultats (Tableau 1) appellent quelques explications fournies par les images StreetView (Figure 5).

Tableau 1: tops/flops des horodateurs…

Adresse	*arrond^{t}*	*Transactions/an*	*Revenu €/an**	*Première transaction*
FLOP 3				
0 RUE ROBERT ESNAULT PELTERIE	75007	10	9	03/07
63 RUE SAINT CHARLES	75015	17	30	26/12
119 BOULEVARD MACDONALD	75019	28	66	18/2
TOP 3				
10 terre-plein RUE DU DEPART	75015	31 509	89 278	01/01
44 terre-plein BOULEVARD DE VAUGIRARD	75015	28 227	78 267	01/01
terre-plein bis BOULEVARD DE CHARONNE	75020	28 146	67 113	02/01

Figure 5: 2 flops (haut) et 2 tops (bas) en StreetView

Un parcmètre dans la zone de bus/taxi de l'aérogare des Invalides, sans aucune place de parking autour ; un autre sur un boulevard périphérique de Paris qui borde une usine. Un troisième n'a été installé qu'en fin d'année et n'a pas pu fonctionner longtemps ; d'autres ont même été enlevés pour faire place à des places moto. A l'inverse, du côté des équipements très utilisés, on trouve les caractéristiques communes suivantes:

- Le terre-plein : qui permet le stationnement d'un grand nombre de voitures
- Le classement en rotatif pur, ce qui interdit aux résidents de se garer et favorise la rotation rapide des véhicules
- La présence proche d'un centre commercial ou d'un supermarché (Montparnasse).

Deux photos du top 3 montrent une personne en train de payer.

Analyse des comportements

La courbe horaire des montants moyens payés à chaque transaction (pour les usagers rotatifs) est particulièrement riche d'enseignements. On observe sur la Figure 6 une baisse forte du montant

unitaire payé à deux moments bien particuliers : le matin entre 7h10 et 7h50, et le soir entre 22h et 23h. Deux hypothèses sont formulées pour expliquer chacun de ces creux : « voiture mal garée » et « diner en ville ». Une hypothèse pour le creux matinal, alors que le nombre de transactions est très faible (0.1% du total sur ce créneau), serait le déplacement des voitures mal garées ou garées sur des livraisons (les conducteurs vont garer leur voiture sur une place autorisée (à partir de 7h elle peut en effet être enlevée) et payent une petite somme avant leur départ au travail. Cela peut être le signe que les contrôles sont très fréquents le matin entre 9 et 10, et les usagers payent une petite somme pour éviter une amende sur ce créneau très contrôlé. Le creux du soir, qui apparaît tous les jours sauf le dimanche, est peut-être de nature différente : une explication serait les déplacements de diner à l'extérieur, typiquement sur le créneau 20h-23h. Lorsque les convives rentrent chez eux, en général dans une zone plus périphérique de Paris et donc au tarif moitié plus faible, ils payent alors leur stationnement du lendemain et cela influence le montant moyen à la baisse. Ce quota d'usagers qui étaient « de sortie » et a donc attendu d'être rentré chez eux pour payer se répercute de façon symétrique dans le pic juste avant, de 20h-21h. Ces phénomènes ne sont pas observés chez les résidents.

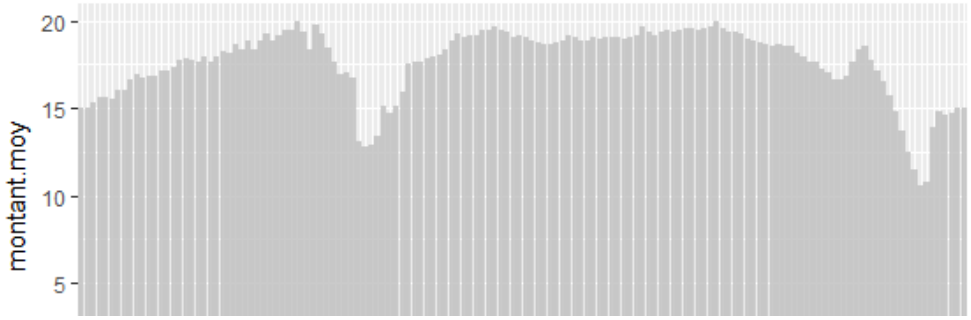


Figure 6: montants rotatifs moyens payés au long de la journée (en haut)

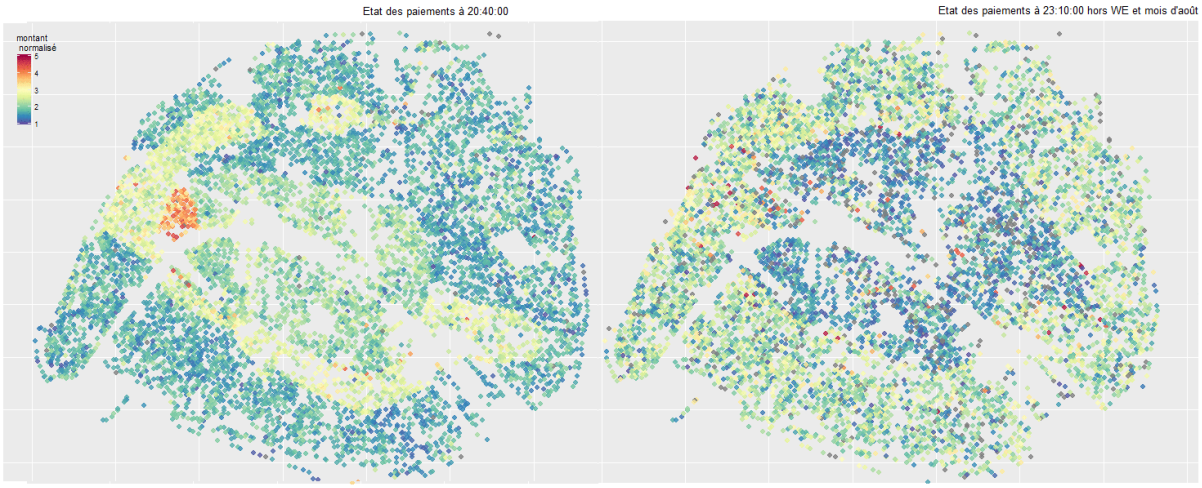


Figure 7: Inversion des flux avant/après diner entre la périphérie et le centre. Note : la zone très rouge est « suspecte » et pourrait refléter un changement de tarif *au cours* de l'année ?

Concentration des paiements

Une mesure de la concentration des paiements est l'indice de Gini, souvent utilisé en économétrie. Un pic indique une concentration des transactions sur un nombre plus restreint d'horodateurs, signifiant par exemple que les automobilistes vont globalement se déplacer dans les mêmes endroits. Sur la Figure 8, le pic le plus fort se situe entre 6 et 8, et globalement le paiement est beaucoup plus diffus pour les résidents que pour le rotatif.

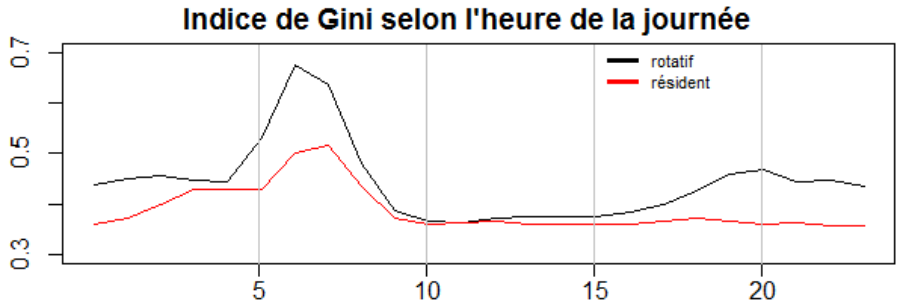


Figure 8: évolution dans la journée de l'indice de Gini du nombre de transactions.

Données météorologiques

Dans un premier temps les données de Weather Underground provenant des 2 aéroports les plus proches (Orly, Le Bourget) sont étudiées. Leurs données de précipitation diffèrent en partie ; si la hauteur d'eau annuelle est proche (+5% de hauteur d'eau à Orly), paradoxalement il y a deux fois plus d'épisodes pluvieux à Orly (1130 heures contre 606 au Bourget). Il n'y a aucune journée comportant des chutes de neige. Quasiment une journée sur deux dans l'année subit un épisode de pluie (les journées étant définies de 0h à 23h). Toutefois, il apparait que les cumuls de précipitation sont largement faux puisque la précipitation totale réelle en 2014 est de l'ordre de 680mm+/-70 contre 480 affiché par Weather Underground. Il y a donc un biais important sur la quantité de pluie reçue pour cette source de données.

La concordance des épisodes pluvieux entre les deux sites est importante pour notre étude ; une matrice dite de « confusion » permet de l'évaluer : 142 journées pluvieuses sont communes aux 2 sites, 69 voient de la pluie sur un seul des deux. A l'échelle horaire, 5.5% des heures annuelles sont pluvieuses aux deux localisations en même temps, près de 9% le sont sur un seul site ; la pluie est donc étalée différemment sur les mêmes journées. Etant donné que la pluie n'est pas distribuée de façon homogène sur les heures de la journée (moins de pluie la nuit), il faudra faire attention à éviter une possible confusion avec d'autres effets.

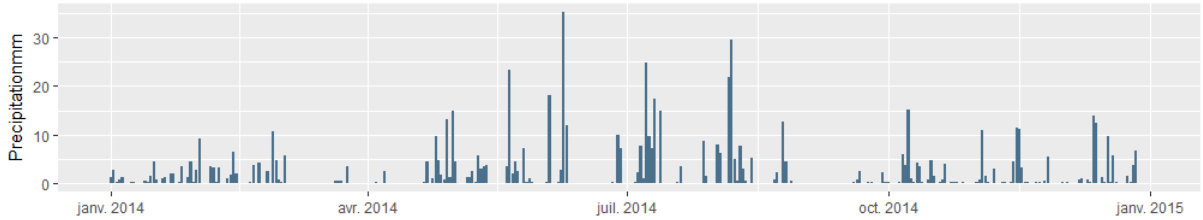


Figure 9: Evolution des précipitations à Montsouris, source : Metigate

Par suite une comparaison est réalisée avec des données simulées fournies par la start-up METIGATE en plusieurs points de la région parisienne dont Paris (réservoir de Montsouris) ou Orly. Ces données proposent une pluviométrie quantitative à pas horaire, ce qui est intéressant ; le cumul sur l'année 2014 est de 693mm à Paris, un chiffre très proche de la valeur fournie par le site InfoClimat (700mm). 148 journées pluvieuses sont communes aux 2 sites Montsouris et Orly, 63 voient de la pluie sur un seul des deux. A l'échelle horaire, 5.8% des heures annuelles sont pluvieuses aux deux localisations en même temps, près de 9% le sont sur un seul site. La comparaison entre Montsouris et un point moyen entre Orly et Le Bourget donne une proximité similaire, voire une légèrement meilleure correspondance entre les heures de pluie.

On vérifie également par des tests de décalage des données qu'il n'y a pas d'erreur de zone temporelle (erreur de transcription des données) ou d'effet de retard (ex : les épisodes de pluie pourraient être systématiquement décalés dans le temps entre les deux sites) ; ce n'est pas le cas.

Les deux sites ci-dessus sont étudiés pour leur précipitation annuelle (Tableau 2) :

Tableau 2: précipitation annuelle à Orly et Montsouris, moyennes sur 11 ans (InfoClimat)

Précipitation annuelle, mm	*Min*	*Moyenne*	*Max*	*Ecart-type*	*Valeur 2014 (Metigate)*	*Valeur 2014 (InfoClimat)*
Paris[1]	499	**603**	703	75	**693**	**700**
Orly	513	**625**	720	72	**705**	**720**

Note : Pour Orly, la localisation des deux sources de données n'est pas tout à fait identique, ce peut expliquer l'écart
[1] *Réservoir de Montsouris, 14ème arrondissement,* (Infoclimat 2015)

Modélisation et Analyse spatiale

Justification

On se propose de modéliser à deux échelles différentes le nombre de transactions reçu par les horodateurs :

A. **à l'échelle horaire, afin d'étudier l'impact de la météo, sur l'ensemble du parc en même temps.**
B. **à l'échelle annuelle, afin d'étudier le revenu moyen de chaque équipement localisé géographiquement.**

Des **modèles séparés** seront construits pour expliquer les paiements de t**ype résidents et de type rotatifs**, puisqu'il a été montré que leur structure était très différente. Par ailleurs, l'analyse de données ayant un caractère géographique, dans la modélisation « **B** », peut s'avérer dangereux dans le cas où une corrélation existe entre des points voisins ; les effets des variables explicatives sont alors surestimés si on n'intègre pas cette dimension spatiale. Les méthodes de la statistique spatiale s'apparentent à celles utilisées dans le traitement des séries temporelles, à la différence qu'au lieu d'avoir une seule dimension exogène (le temps), on en analyse deux (x et y), voire trois. En plus d'obtenir un modèle valide, les modèles spatiaux renseignent sur la répartition et la force des corrélations spatiales, pouvant apporter des éclairages intéressants. La démarche se décompose en 4 phases :

1. *Définir une mesure de distance spatiale.*
2. *Appliquer un modèle linéaire classique.*
3. *Test de l'existence d'une corrélation spatiale sur les résidus du modèle.*
4. *Si l'effet est avéré, mise en œuvre d'un modèle spatial à la place.*

A. Modèle horaire

Les données agrégées Paris à pas horaire sont jointes aux données météorologiques de Montsouris (75014) et un modèle linéaire est réalisé sur le nombre de transactions.

La première étape de la modélisation est de définir les variables à modéliser et leur éventuelle transformation. On peut s'intéresser au montant total ou moyen payé, ou au nombre de transactions sur une unité spatiale ou temporelle. Du fait des écarts importants (on l'a vu avec l'indice de Gini), on se doute que la variable brute peut poser problème. La transformation de *Box-Cox* préconise de prendre la racine carrée si on étudie le nombre de transactions des résidents ; et une puissance plus forte (0.15) pour le montant total des rotatifs. On parvient alors à normaliser de façon satisfaisante les données (Figure 10).

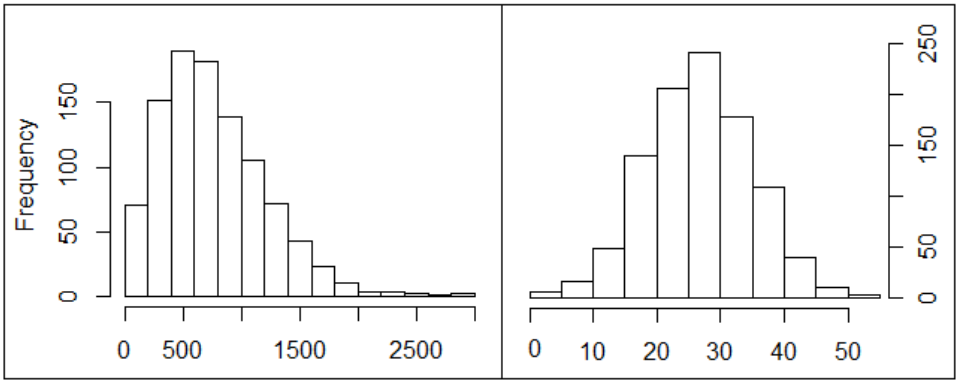


Figure 10: Histogramme des nombres de transactions annuels par horodateur, bruts (gauche) ou en racine carrée (droite). Un test de normalité de Shapiro sur un échantillon de 1000 points valide la normalité pour le cas de droite.

Le modèle proposé pour le stationnement résidentiel est construit à partir des informations et des profils horaires et hebdomadaires, et en utilisant une procédure de sélection de variable *stepwise* ; elle aboutit au modèle [M1] suivant (le $[*]$ indique une interaction) exclusivement calendaire:

$$\textbf{[M1]} \quad \boldsymbol{transactions}_{\boldsymbol{resid}}^{\boldsymbol{0.5}} \sim . \boldsymbol{heure} * \boldsymbol{is.aout} * (\mathbf{1} + \boldsymbol{is.sam} + \boldsymbol{is.dim}) + . \boldsymbol{joursem}_{\boldsymbol{num}}$$

Il intègre donc des coefficients individuels pour chaque heure de la journée, et distincts entre semaine/samedi/dimanche et différents sur le mois d'août qui est très particulier, ainsi qu'un terme décroissant sur le numéro du jour de la semaine (les résidents règlent leur stationnement tôt dans la semaine). Le R^2 obtenu est de 0.61, ce qui est correct mais mériterait un approfondissement (l'effet des jours fériés et vacances pourrait surement l'améliorer). On peut représenter l'ajustement sur quelques jours pris au hasard en février (Figure 11): le modèle suit bien la tendance mais explique encore mal le pic du lundi matin par exemple.

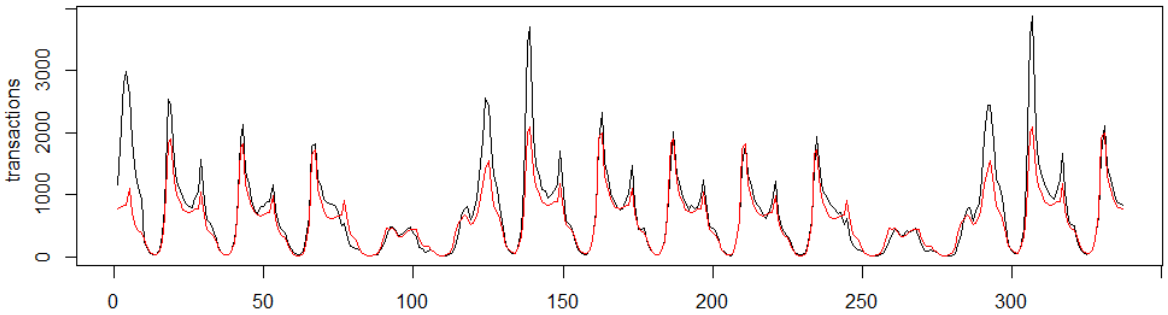


Figure 11: prédictions [M1] du nombre de transactions horaire (rouge) et valeurs réelles (noir) – tarif résidentiel

Transformations de la variable pluie

On souhaite ajouter au modèle l'information climatique. Toutefois, la réponse à la variable pluie peut être plus complexe qu'un simple effet proportionnel à la hauteur d'eau tombée chaque heure. On peut imaginer de décliner et tester cette variable après diverses transformations :

- une transformation de type puissance de la hauteur d'eau tombée
- un effet binaire (l'effet mesure la présence ou absence de pluie)
- un cumul toutes les 12h (pour coller à un pas de temps journalier), variable déjà disponible
- un seuillage quantitatif (l'effet n'existe que pour une pluie suffisamment forte)
- un seuillage temporel (l'effet n'existe que pour une pluie suffisamment longue)
- un décalage (l'effet est retardé ou avancé par rapport à l'épisode de pluie)
- un étalement temporel (l'effet se prolonge avant et après l'épisode de pluie)
- *ou diverses combinaisons des précédentes transformations…*

Le choix de la transformation la plus adaptée se fait par le critère **BIC** (*Bayesian Information Criterion*) qui est connu pour être un critère parcimonieux et efficace en prédiction, avec des modèles fittés non pas sur la variable « nombre de transactions », mais sur les résidus d'un premier modèle utilisant des variables calendaires, sinon le nombre de variables total serait trop grand pour la procédure de sélection. Le résultat de cette procédure est présenté dans le tableau et le choix est fait a priori de ne conserver qu'une seule variable dans le modèle, ici la combinaison $seuil + lissage$ qui donne le meilleur fit et améliore très significativement le modèle principal (test de Fisher $p.value = 6 \cdot 10^{-7}$).

	BIC Numérique	*BIC Binaire*
Precip	**64511**	**64507**
Precip, cumulée 12H	64516	64513
Precip, retardée 1H	64514	64508
Precip, seuil temporel >1h	**64502**	64511
Precip, seuil hauteur >0.4	64511	64508
Precip, lissage moy.mob. +/-2H	64507	**64490**
Combinaisons des précédentes		
Seuil temporel + lissage	**64489**	64510

On interprète donc ces résultats en concluant que **les précipitations ont un effet certain à la baisse, mais de faible ampleur sur le stationnement résidentiel:** le coefficient des précipitations indique que **chaque mm de pluie abaisse le nombre de transactions.** Toutefois, la transformation de la variable Y (racine carrée) et le fait de lisser la hauteur de pluie rend problématique l'interprétation du coefficient, et on préfère estimer le coefficient de la variable binaire correspondante (oui/non) qui donne entre **-10 et -50 transactions/heure en cas de pluie dans les 3 heures**.

L'effet n'est **pas détecté pour des épisodes de pluie trop bref (<1h) et est visible jusqu'à 2h avant et 2h après l'épisode de pluie.** Ce n'est pas parce que les parisiens sont *devins* mais simplement dû au fait que la donnée météo est mesurée au Sud de Paris alors que la pluie se décale souvent d'Ouest en Est, donc a déjà débuté ailleurs dans Paris au moment où elle est détectée par la station Montsouris.

La même méthodologie appliquée au **stationnement rotatif** nécessite un traitement légèrement différent ; en premier lieu il est nécessaire d'appliquer une transformation logarithmique et non racine au nombre de transactions, dû à leur forte étalement vers les grandes valeurs, ainsi qu'un décalage de $+1$ pour éviter de se retrouver avec $\log(0)$. Le modèle retenu [M2] est simplifié par rapport à [M1], fusionne le samedi et dimanche et élimine une bonne partie des interactions avec le mois d'août sans les supprimer totalement ; les jours de la semaine sont considérés identiques.

[M2] $$\boldsymbol{log(transactions_{rotat}) \sim (heure + is.aout) * (1 + is.we)}$$

L'étude des résidus s'avère décevante et on suspecte qu'une variable manque. On intègre alors dans le modèle la donnée des jours fériés 2014 (au nombre de 11), en interaction avec l'heure, aboutissant à une légère amélioration [M2.1] mais sans toutefois résoudre le problème des creux sur certains jours de semaine, régulièrement. Ainsi, on voit des creux importants en mars, ainsi que les 8 et 9 juillet qui sont des jours de milieu de semaine sans pont de vacance, alors que le samedi 12 montre un écart légitime car c'est le pont du 14 juillet. Le stationnement résidentiel montre, lui aussi, des revenus quasi nuls sur la journée du 8 juillet. Une recherche Google de ces cas particuliers permet en fait de matcher avec les décrets de **gratuité de stationnement résidentiel pris lors des pics de pollution** aux particules, et améliore le modèle ([M2.3]). Certaines journées restent encore inexpliquées et pourraient correspondre à des **pannes ou des maintenances** d'une partie importante du parc (hypothèse non étudiée).

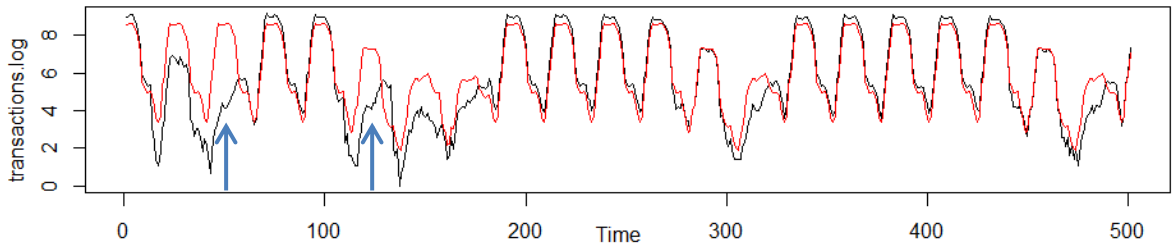


Figure 12 : prédictions [M2.1] du log du nombre de transactions horaire (rouge) et valeurs réelles (noir) – tarif rotatif – les flèches correspondent aux journées du 9 juillet (un mercredi) et au 12 juillet (samedi).

La sélection de la variable précipitations donne le même résultat que pour les paiements résidentiels (lissage +/-2H et seuillage des épisodes pluvieux <1h). On obtient le modèle final **[M2.3]** :

[M2.3] $\boldsymbol{log(transactions_{rotat}) \sim (heure + is.aout) * (1 + is.we) + pics + ferie * heure}$

B. Modèle spatial annuel, définition d'une mesure de distance

La définition d'une distance entre les individus a pour objectif d'aboutir à une matrice de « voisinage », carrée et de dimension égale au nombre d'individus, l'élément {i, j} de la matrice donnant la valeur de proximité entre les observations i et j. Plusieurs types de distance existent :

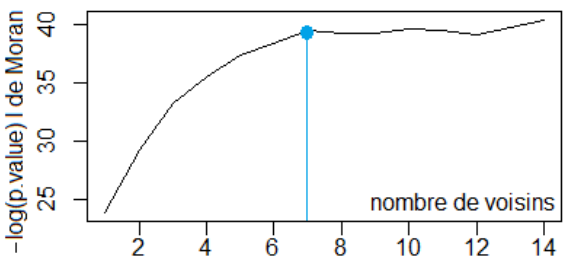
- une distance euclidienne, lorsqu'on dispose de coordonnées géographique, ou lorsque les effets sont a priori inversement proportionnels à la distance ;
- une distance par voisinage si les effets dépendent uniquement d'un individu contigu direct ou de rang 2 ; cette distance est binaire : 1 ou 0, et crée donc une matrice '*sparse*'.
- une distance spécifiquement conçue pour le cas étudié (prenant par exemple en compte la proximité routière des horodateurs. Ce serait complexe à mettre en œuvre.

Ici, l'idéal serait une distance de type Manhattan utilisant le tracé routier, intégrant également une pénalité de sortie de zone de stationnement (distance de type Gower); toutefois une telle démarche serait complexe et à défaut une distance de Manhattan simple sera utilisée pour le calcul de la proximité P entre deux horodateurs i et j avec une distance d'influence limite à 1.5km ; par ailleurs les coordonnées GPS, exprimées en degré de latitude et longitude, ne correspondent pas à des distances homogènes et doivent être transformées en km par les coefficients 73 et 111 km/deg sous peine de fausser l'identifications des voisins :

$$P_{i,j} = \begin{cases} \dfrac{1}{d_{i,j}} \\ 0 \; si \; i = j \; (par \; convention) \\ 0 \; si \; d_{i,j} > 1.5 \end{cases}$$

$$avec \; d_{i,j} = 73 * \left|long_i - long_j\right| + 111 * \left|lat_i - lat_j\right|$$

Dans certains modèles spatiaux, une matrice de proximité par k-plus proches voisins est acceptée en entrée seulement ; le nombre de voisins est choisi à 7, ce nombre minimisant la p-value du test de Moran. La cartographie de voisinage peut être représentée par des liens (Figure 13) ; une image GIF animée montrant le résultat selon le paramètre K est disponible ici.



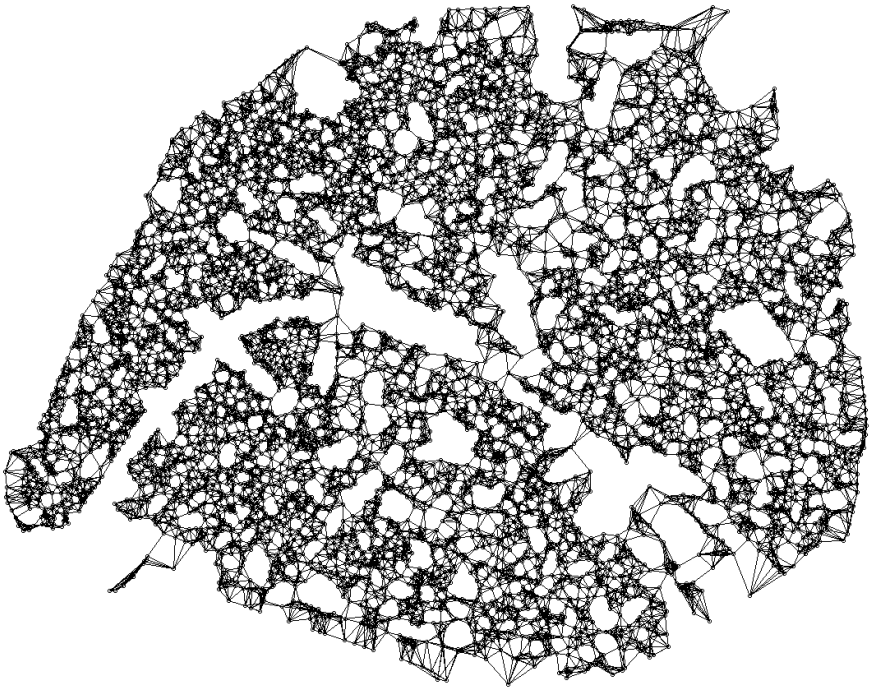


Figure 13: cartographie de voisinage pour K=7

Modélisation spatiale

Des premiers modèles sont réalisés afin de vérifier l'absence d'influence de certains facteurs indésirables comme le modèle de l'horodateur ou le mode de paiement. Il apparaît cependant que :

- le **mode de paiement** affecte significativement le montant payé : la carte bleue est utilisée pour des montants légèrement supérieurs en moyenne (2.60€ contre 1.90€ avec la carte Paris Stationnement). Cela n'est gênant que si la répartition géographique des détenteurs de carte a une structure particulière ; en effet le ratio d'utilisation de la CB varie entre 40 et 50% (écart interquartile) mais les différences entre les arrondissements, par exemple, sont très faibles, entre 1 et 3 points, et on **négligera** ce facteur par la suite.
- le **modèle de parcmètre** est significatif sur le nombre de transactions. Malheureusement, la répartition des modèles est en partie confondue avec les arrondissements (Figure 14). Même si on prend le cas du 18ème arrondissement, où les deux modèles semblent répartis aléatoirement, on note comme partout ailleurs une performance inférieure pour le modèle SAB3 (-15% en transactions et montants) indépendamment d'autres facteurs identifiables (arrondissement, zone tarifaire, heure de la journée ou mois de l'année). La seule différence connue entre ces deux modèles est l'alimentation le modèle SAB3 comportant un **panneau solaire**. Sans plus de détail technique disponible sur les caractéristiques et dates d'implémentation de ces deux modèles, on ne peut fournir d'explication à ce phénomène et cette variable devra être **conservée** dans les modèles ultérieurs.

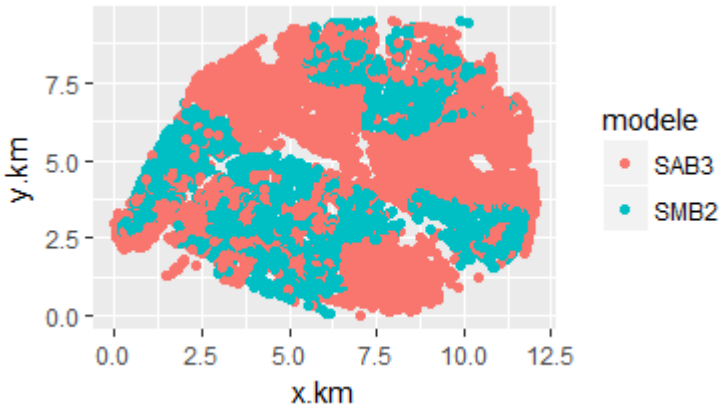


Figure 14: modèles de parcmètres dans Paris

Modèle baseline OLS

On doit d'abord tester un effet spatial sur les résidus d'un modèle linéaire. A l'échelle annuelle, on ne dispose que des facteurs explicatifs suivants pour un modèle linéaire : la zone tarifaire, le modèle (=alimentation) de parcmètre, éventuellement l'arrondissement du fait de la spécificité de chaque arrondissement (habitat, entreprises, etc.) ; la météo n'est pas intégrée car on a supposé qu'elle était globalement identique partout sur l'année entière.

[M3.1] $$\boldsymbol{transactions_{resid}^{0.5} \sim tarif + alimentation + arrondissement}$$

[M3.2] $$\boldsymbol{transactions_{rotat}^{0.5} \sim tarif + alimentation + arrondissement}$$

Le **I de Moran**, dans sa version globale ou locale, est calculé sur les résidus des modèles linéaires précédents et indique l'ampleur de l'écart à une situation fictive de non-corrélation spatiale. Il est identique pour les deux types d'usagers et vaut 0.06 pour la distance euclidienne ou 0.24 pour la distance de voisinage, avec une très forte significativité montrant des corrélations spatiales. La distance par voisinage semble mieux capter les corrélations spatiales. Notons au passage que la variable *arrondissement* « récupère » une bonne part de la corrélation spatiale car sans elle le I de Moran est deux fois plus fort. On peut donc à présent spécifier un modèle à corrélation spatiale ;

Modèle à corrélation spatiale

Les modèles à corrélation spatiale appartiennent à deux grandes familles : les **SAR** (*spatial auto-regressive*, modèles à décalage spatial dans le cas $\lambda = 0$) et les **SEM** (*spatial error models* dans le cas $\rho = 0$), ils sont les homologues des modèles AR et MA pour les séries temporelles (Anselin 2003).

$$Y = \rho WY + X\beta + \varepsilon \text{ , avec } \varepsilon = \lambda V\varepsilon + N(0, \sigma^{2I_n})$$

Dans les premiers la valeur de chaque point influence les points environnants ; dans le second des effets sous-jacents inobservables influencent la variable en certains endroits. La combinaison des deux existe aussi. Le test du multiplicateur de Lagrange permet de décider quelle approche doit être utilisée (Lesage 2008) : ici le modèle à erreur spatiale semble légèrement meilleur que le modèle SAR. Un modèle mixte serait également possible mais étant donné le nombre de points, le calcul des coefficients demande un temps de calcul beaucoup trop long.

Tableau 3: test du multiplicateur de Lagrange sur le nombre de transactions

Forme de modèle	*p-value stationnement rotatif*	*p-value stationnement résidentiel*
RLMerr	< 2.2e-16	1.407e-07
RLMlag	0.0001516	0.5632

En termes d'interprétation, le modèle à erreur spatiale est logique puisque les parcmètres ne sont pas en « compétition » et ne s'influencent pas les uns les autres globalement. On a donc plutôt des spécificités locales (habitat, lieux de travail, centres commerciaux, etc.) qui influencent localement le nombre et le montant des transactions, et créent dans les rues voisines un effet de diffusion, ainsi que pouvait le montrer l'analyse descriptive et les photos.

Le I de Moran sur les résidus du modèle n'est plus significatif ($p.value = 0.276$) et montre qu'on a réussi à bien expliquer la dépendance spatiale avec le modèle **SEM**.

Conclusion

L'analyse des transactions d'horodateurs s'est révélée une tâche délicate par la complexité même des modes de paiement et d'utilisation des horodateurs. Une fois ces mécanismes décortiqués et pris en compte, on peut faire émerger des conclusions intéressantes sur les patterns d'utilisation, les habitudes des parisiens ou les mouvements pendulaires au cours de la journée, ou encore la concentration des quartiers « chauds » ou « calmes » en termes de stationnement.

Une frustration cependant tient à l'impossibilité de coupler les données à des données fiscales d'infraction au stationnement qui auraient permis d'étudier l'influence de la pression du contrôle sur le consentement au paiement.

En revanche, l'ajout de données météorologiques de bonne qualité s'est avéré fructueux avec une meilleure compréhension de l'impact de la pluie sur le paiement. Il reste à ce stade difficile de discriminer si cet effet négatif s'explique par une baisse de la mobilité ou un calcul de risque conscient vis-à-vis d'une éventuelle verbalisation.

Il serait intéressant de continuer ce travail par un modèle à la fois spatial et temporel, avec une résolution plus précise des variables météorologiques. Toutefois, l'évolution temporelle des paiements à pas horaires a pu être modélisée sur l'ensemble de la ville de façon efficace selon le calendrier, avec un impact fort d'évènements particuliers comme les jours fériés et les pics de pollution, rendant possible son éventuelle utilisation par la ville de Paris pour la prise de décision ou l'organisation logistique des contrôles.

L'enseignement le plus marquant de ce travail est sans doute la capacité de l'analyse de données à extraire une quantité singulière d'informations d'un jeu de données comme celui-là, même si certaines interprétations restent sous la forme d'hypothèse à vérifier avec des experts.

Références

Anselin, L. An introduction to spatial regression analysis in R. University of Illinois: Urbana-Champaign, 2003.

Infoclimat. Climatologie de l'année 2015 à Paris - Montsouris. 2015. http://www.infoclimat.fr/climatologie/annee/2015/paris-montsouris/valeurs/07156.html (accès le 08 31, 2017).

Lesage, James. "An Introduction to Spatial Econometrics." Revue d'Economie Industrielle, 2008: 123.

Mairie de Paris/Direction de la Voirie et des Déplacements. Open Data Paris - Horodateurs - Mobiliers. 17 02 2017. http://opendata.paris.fr/explore/dataset/horodateurs-mobiliers/information/ (accès le 08 31, 2017).

—. Open Data Paris - Horodateurs - transactions de paiement. 08 06 2015. http://opendata.paris.fr/explore/dataset/horodateurs-transactions-de-paiement/information/ (accès le 08 31, 2017).