**RAPPORT DEVOIR STATISTIQUE STPATIALE ET STATISTIQUE AVANCEE**

**AUTEUR : HABIB MBOW**

**INTRODUCTION ET PROBLEMATIQUE DE L’ETUDE**

L’analyse spatiale des valeurs immobilières est l’une des clés de observation et de la compréhension des transformations physiques et sociales des centres villes. Au-delà de l’évidente dimension opérationnelle qu’elle recouvre, ne serait-ce que pour identifier les zones de forte pression immobilière ou pour évaluer l’impact des opérations d’aménagement sur la valorisation des biens, la géographie des prix possède une dimension théorique fondamentale. Elle interroge d’une part la formation de la valeur d’un bien et plus précisément, l’influence sur cette valeur des qualités d’un lieu et de son voisinage, elles-mêmes perçues et évaluées à différents échelons (situation dans l’agglomération, caractéristiques du quartier mais aussi de la rue). Elle soulève par ailleurs la question de la juxtaposition locale entre des marchés en plein essor et des quartiers en voie de dépréciation. Plusieurs modèles ont été développés pour explorer cette perspective théorique.  Notre base conserne les transactions immobilières en France métropolien pour l’année 2019 et contient des informations sur le prix, la date, l’adresse, les éléments descriptifs du bien, etc…. Dans un premier temps on cherche à caractériser l’organisation spatiale des variables étudiées (analyse variographique) et à présenter la méthode de krigeage qui permet de prédire la valeur du prix des biens à partir des coordonnées géographiques (longitude et latitude). Dans un second temps de justifier le choix d’outils de statistique spatiale pour faire de l’estimation, modélisation du prix des biens à partir des autres variables.

1. **CONSTRUCTION DE LA BASE DE DONNEES**

Cet ensemble de données dérive de l’ensemble de données de data gouv.fr et conserne les demandes de valeurs foncières de l’année 2019 en France. Il contient des données à caractères personnel et est organisé de la manière suivante:

• identification de chaque mutation par son numéro de disposition;

• affichage d’une ligne par local;

• pour chaque local (chaque ligne), les lots de l’immeuble correspondant (dans la limite de 5) sont affichés ainsi que le nombre total de lots pour cet immeuble;

• la surface réelle est associée au local;

• restitution de la surface carrée associée au lot lorsqu’elle est indiquée.

Le but de cette partie est de transformer l’adresse de chaque transaction immobilières en coordonnées géographiques (GPS). Pour ce faire on a créé une table data frame nommée Adresse contenant les variables suivantes: Code postal, Code départemental, Code commune, Code voie et No voie, puis nous exportons cette table en fichier csv nommée fichier.csv afin de l’importer dans le site adresse.data.gouv.fr/csv pour le géocoder. Après avoir géocodé la base nous obtenons une nouvelle base fichier.geocoded.csv.

1. **STATISTIQUE SPATIALE**

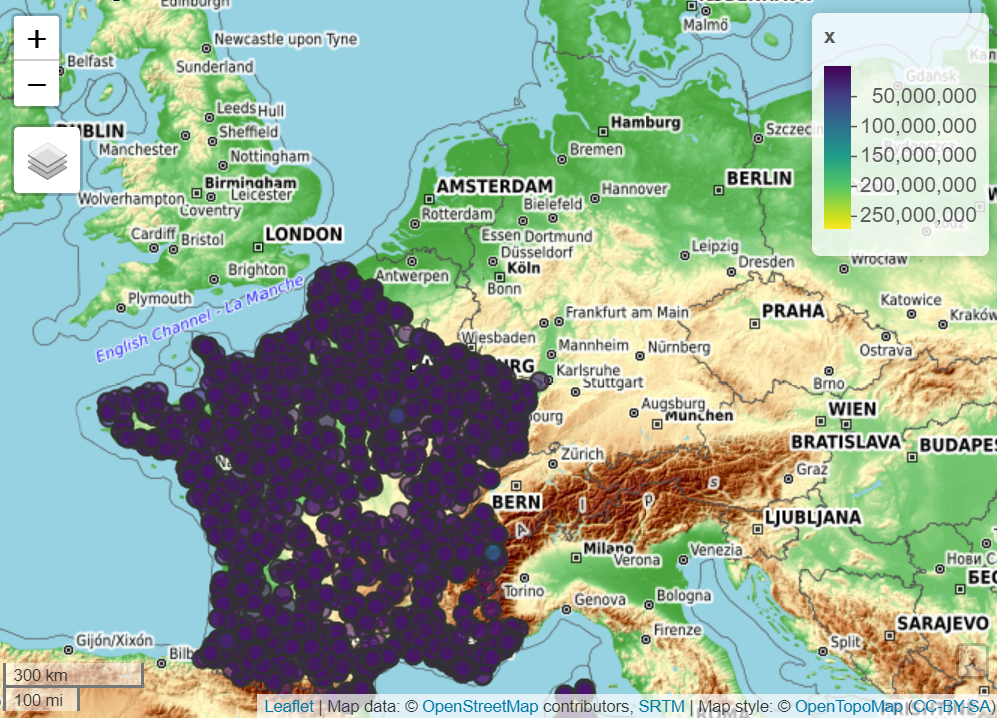
Le but de cette partie vise à prédire le prix des biens de l’année 2019 à l’aide du modèle de krigeage

**1- ANALYSE EXPLORATOIRE PRIX DES BIENS**

L’analyse exploratoire des données spatiales est une étape incontournable dans les processus d’analyse spatiale. On cherche à effectuer une analyse exploratoire sur les prix de biens au cours de l’année 2019 sur le territoire métropolien et les DOM-TOM à l’exception de l’Alsace-Moselle et de Mayotte.

* 1. **Carte du prix des biens en fonctions des coordonnees géographiques**

La carte ci-dessous donne une projection de la répartition du prix des biens sur le territoire métropolien et les DOM-TOM.



* 1. **Etude variographique et identification du variogramme le plus adéquat pour modéliser la dépendance spatiale**

Pour appliquer ce type d’interpolation, il est nécessaire d’explorer la structure spatiale des données afin de vérifier si celles-ci sont bien autocorrélées. L’analyse variographique va nous permettre de mener cette étude à bien. L’outil principal permettant cette analyse est le semi-variogramme qui décrit l’évolution de la semi-variance en fonction de la distance entre les mesures et permet ainsi de d’étudier le lien spatial être les données. Il est défini de la manière suivante

**γ(h)=1/2(V ar(Z(s + h) - Z(s))) ∀s ∈ D**

Le tableau ci-dessous donne les résultats du variogramme calculé par le choix automatique de la fonction autofitvariogramm du package automap de R.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variogramme Expérimental** | | | | | |
| **np** | **dist** | **gamma** | **dir.hor** | **Dir.ver** | **id** |
| 1091956**1** | 0.4212842 | 4.572261e+13 | **0** | **0** | var1 |
| 14683666 | 1.4956999 | 5.523589e+13 | 0 | 0 | var1 |
| 19381329 | 2.4112317 | 6.748102e+13 | 0 | 0 | var1 |
| 30641126 | 3.6452192 | 5.348365e+13 | 0 | 0 | var1 |
| 36758607 | 5.0995166 | 6.104562e+13 | 0 | 0 | var1 |
| 28823838 | 6.4684587 | 6.000785e+13 | 0 | 0 | var1 |
| 17044785 | 8.2525149 | 4.958767e+13 | 0 | 0 | var1 |
| 99277 | 12.9418454 | 7.229217e+12 | 0 | 0 | var1 |

Le variogramme le plus adéquat pour modéliser la dépendance spatiale est le variogramme théorique.

1. **PREDICTION SPATIALE DU PRIX DES BIENS**

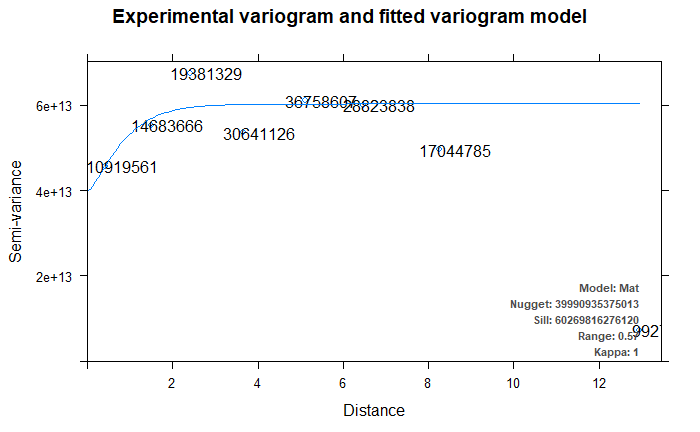
L’idée de la prédiction spatiale du prix des biens ou krigeage est de prévoir le prix des biens en un site non échantillonné s0 par une combinaison linéaire des données ponctuelles adjacentes. On veut une estimation non biaisée et de variance minimale. Ces conditions se résument en quatre contraintes qui vont permettre de construire le système de krigeage.

* 1. **Modèle de prédiction adapté**

Le variogramme (graphe ci-dessous) décrit la continuité spatiale de la variable régionalisée et détermine l’importance des variations aléatoires. Il atteint un plateau le palier.

Le tableau suivant donne les résultats du variogramme estimé par choix automatique

|  |  |  |  |
| --- | --- | --- | --- |
| Modèle Adopté du Variogramme | | | |
| model | **psill** | **range** | **kappa** |
| Nug | 3.999094e+13 | 0 | 0 |
| Mat | 2.027888e+13 | 0.5717283 | 1 |

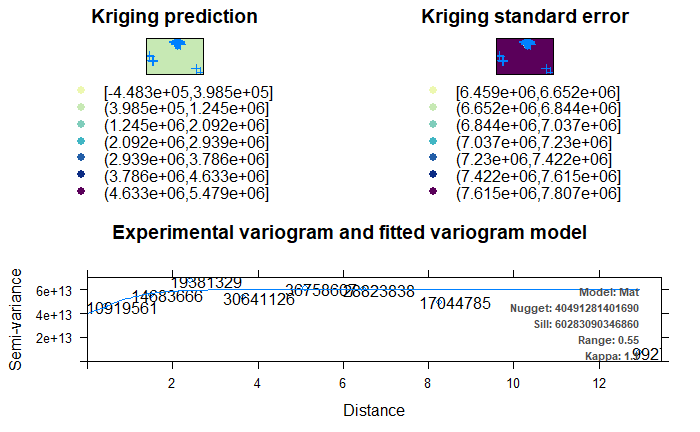


D’après la sélection automatique, le modèle de prédiction le plus adéquat pour modéliser le prix des biens est le modèle de variogramme de Matern (Mat). L’examen de ce modèle fait ressortir une discontinuité du variogramme à l’origine ce qui correspond à un effet de pépite (=3.999094e+13) qui est dû à une microstructure spatiale non détectée par l’échantillonnage ou la présence d’une erreur de mesure. Il s’agit d’un variogramme borné de palier 2.027888e+13 et qui se stabilise à d’une portée de 0.5717283 m ce qui signifie qu’au de la de cette distance la variable du prix des biens n’est plus autocorrélée spatialement.

* 1. **Interpolation du prix des biens à l’aide du krigeage**

Afin de kriger le prix des biens avec R, on a d’abord créé une grille sur laquelle on pourra effectuer le krigeage. Après avoir créée la grille, on a effectué le krigeage et on obtient les résultats suivants :

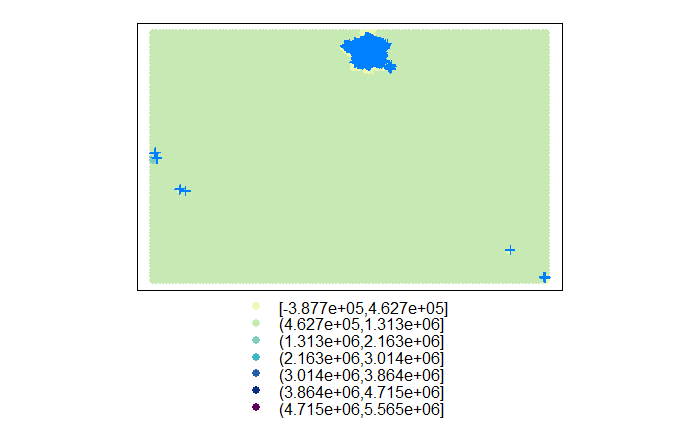
## krige\_output:  
## Object of class SpatialPointsDataFrame  
## Coordinates:  
## min max  
## x -62 56  
## y -22 52  
## Is projected: NA   
## proj4string : [NA]  
## Number of points: 8925  
## Data attributes:  
## var1.pred var1.var var1.stdev   
## Min. :-448328 Min. :4.172e+13 Min. :6459013   
## 1st Qu.: 577133 1st Qu.:6.096e+13 1st Qu.:7807420   
## Median : 577133 Median :6.096e+13 Median :7807420   
## Mean : 576993 Mean :6.077e+13 Mean :7794464   
## 3rd Qu.: 577133 3rd Qu.:6.096e+13 3rd Qu.:7807420   
## Max. :5479367 Max. :6.096e+13 Max. :7807420   
##   
## exp\_var:  
## np dist gamma dir.hor dir.ver id  
## 1 10919561 0.4212842 4.572261e+13 0 0 var1  
## 2 14683666 1.4956999 5.523589e+13 0 0 var1  
## 3 19381329 2.4112317 6.748102e+13 0 0 var1  
## 4 30641126 3.6452192 5.348365e+13 0 0 var1  
## 5 36758607 5.0995166 6.104562e+13 0 0 var1  
## 6 28823838 6.4684587 6.000785e+13 0 0 var1  
## 7 17044785 8.2525149 4.958767e+13 0 0 var1  
## 8 99277 12.9418454 7.229217e+12 0 0 var1  
##   
## var\_model:  
## model psill range kappa  
## 1 Nug 4.049128e+13 0.0000000 0.0  
## 2 Mat 1.979181e+13 0.5522159 1.1  
## Sums of squares betw. var. model and sample var.[1] 3.661897e+32

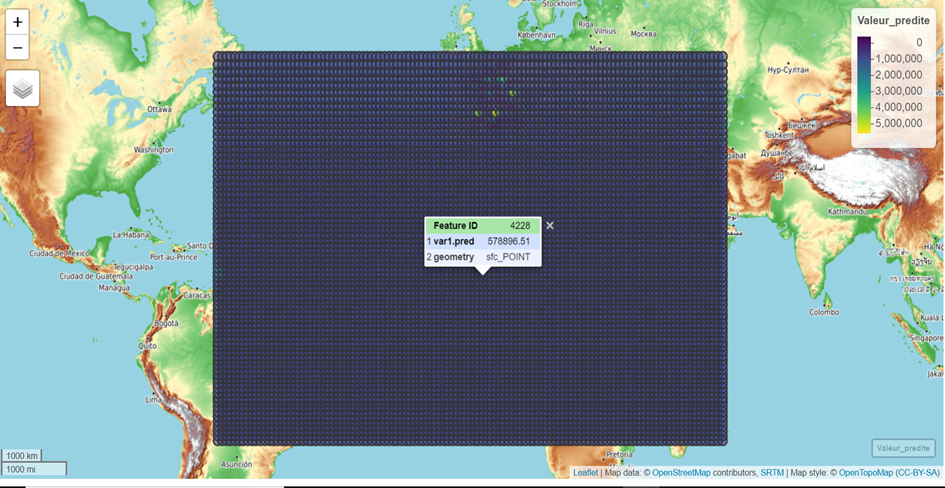


La carte de prédiction du prix (graphe en haut à gauche), La carte de variance du krigeage (graphe en haut à droite)

La figure du bas montre le modèle variogramme de Matern sélectionné automatiquement.

Le graphe ci-dessous donne une interpolation des résultats du krigeage sur la carte :





* 1. **Interprétation de la qualité de prédiction**

La carte de prédiction du prix nous fournit un aperçu de la répartition de cette variable sur le territoire métropolien et les DOM-TOM. Dans l’opération du krigeage les valeurs du prix des biens ont été réparties en classes automatiquement. Nous pouvons remarquer sur cette carte que dans la totalité le prix des biens de 2019 se situe entre 2092000 et 2939000 euros sur le territoire métropolien et les DOM-TOM. Elle est plus élevée dans certaines zones. La position géographique est le facteur conditionnel de cette répartition. La répartition du prix des biens dépend forcément de la position géographique de l’immobilier.

La carte de variance du krigeage de la variable interpolée permet de juger la qualité de la fonction d’interpolation pour le prix des biens. Elle représente la dispersion possible de la variable réelle et inconnue autour de la variable obtenue par krigeage. Plus la variance de krigeage est faible, la valeur interpolée est en moyenne prés de la réalité et donc plus la carte est précise. Une forte variance d’erreur implique une incertitude plus grande et donc une estimation moins précise. La carte de la variance krigeage du prix de des biens immobilier montre que la prédiction de ses valeurs est peu fiable dans les zones non l’échantillonnées. De ce fait nous remarquons que les classes des erreurs de prédictions du prix des biens sont très élevés. Toute fois la variance du variogramme dépend uniquement du modèle de variogramme et de la configuration des données dans le voisinage du krigeage, elle ne dépend pas des valeurs numériques des données à l’intérieur de voisinage. Elle ne dépend pas des valeurs numériques des données à l’intérieur du voisinage.

Les paramétrés du modèle variogramme de Matern obtenue par krigeage sont à peu identiques à ceux du variogramme de Matern dans l’étude variographique.

1. **STATISTIQUE AVANCEE**

Dans cette partie, on cherche à modéliser le prix des biens à l’aide d’un modèle de régression spatiale en fonctions des autres variables.

1. **ETUDE DESCRIPTIVE**

Avant d’opter pour une modélisation spatiale, il est souvent utile d’analyser les données sans formuler d’hypothèses à leur égard. Toute étude sophistiquée d’une base de données doit être précédée d’une étude exploratoire à l’aide de plusieurs outils, certes rudimentaires mais robustes. C’est la meilleure façon de se familiariser avec les données.

La première étape de toute investigation dans les données est l’examen des statistiques univariées des variables aﬁn de détecter d’éventuelles anomalies dans leur distribution (valeurs manquantes, erronées ou atypiques). Dans un second temps, les statistiques bivariées permettent de repérer les incohérences entre les variables. Ces deux étapes permettent l’apurement et le redressement des données (imputation ou non des données manquantes, transformation logarithmique, discrétisation des variables continues, suppression de quelques observations le cas échéant).

* **Etude univarié**

Nous présentons ci-dessous les résultats univariés de notre fichier de données

Notre fichier contient 45 variables et 10017154 observations

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Nbre manquantes** | **Distinctes** | **% manquantes** |
| No disposition | 0 | 415 | 0 |
| Date Mutation | 0 | 179 | 0 |
| Nature Mutation | 0 | 6 | 0 |
| Valeurs foncière | 12785 | 54377 | 1.26 |
| No voie | 410827 | 5055 | 40.39 |
| B/T/Q | 972795 | 39 | 95.64 |
| Type de voie | 424063 | 122 | 41.69 |
| Code voie | 12133 | 13560 | 1.19 |
| Voie | 12193 | 187125 | 1.20 |
| Code postal | 12156 | 5541 | 1.20 |
| Commune | 0 | 24474 | 0 |
| Code département | 0 | 97 | 0 |
| Code commune | 0 | 897 | 0 |
| Préfixe de section | 964778 | 647 | 94.85 |
| Section | 24 | 575 | 0.001 |
| No plan | 0 | 5220 | 0 |
| No volume | 1014839 | 232 | 99.77 |
| 1er lot | 699386 | 5272 | 68.76 |
| 2ème lot | 949848 | 2654 | 93.38 |
| 3ème lot | 1006414 | 802 | 98.94 |
| 4ème lot | 1015324 | 330 | 99.63 |
| 5ème lot | 1013342 | 461 | 99.82 |
| Surface carrée du 1er lot | 926733 | 13603 | 91.11 |
| Surface carrée du 2ème lot | 994621 | 8688 | 97.78 |
| Surface carrée du 3ème lot | 1017154 | 0 | 100 |
| Surface carrée du 4ème lot | 1017154 | 0 | 100 |
| Surface carrée du 5ème lot | 1016949 | 190 | 99.82 |
| Identifiant local | 1017154 | 0 | 100 |
| Nombre de lots | 0 | 65 | 0 |
| Surface réelle bati | 476539 | 2490 | 46.95 |
| Surface de terrain | 321571 | 28499 | 31.61 |
| Nombre pièces principales | 476539 | 31 | 46.95 |
| Type local | 475851 | 4 | 46.78 |
| Code type local | 475851 | 4 | 46.78 |
| Nature culture | 321571 | 27 | 31.61 |
| Nature culture spéciale | 973376 | 106 | 95.70 |
| Longitude | 35817 | 12217 | 3.52 |
| Latitude | 35817 | 12217 | 3.52 |
| Code service CH | 1017154 | 0 | 100 |
| Reference document | 1017154 | 0 | 100 |
| 1 Articles CGI | 1017154 | 0 | 100 |
| 2 Articles CGI | 1017154 | 0 | 100 |
| 3 Articles CGI | 1017154 | 0 | 100 |
| 4 Articles CGI | 1017154 | 0 | 100 |
| 5 Articles CGI | 1017154 | 0 | 100 |

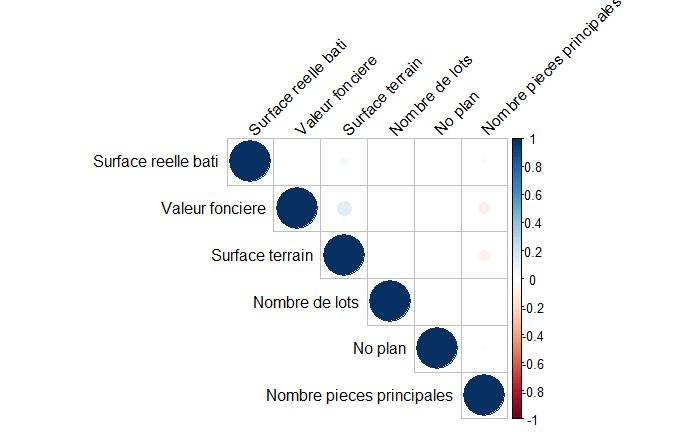
On remarque les variables (Code service CH, Reference document, 1 Articles CGI, 2 Articles CGI, 3 Articles CGI, 4 Articles CGI, 5 Articles CGI, Identifiant local, Surface Carrez du 4eme lot, Surface Carrez du 3eme lot) et (Surface du 1er lot, Surface du 2ème lot, Surface du 5ème lot, Nature de culture spéciale, 1er lot, 2ème lot, 3ème lot, 4ème lot, 5ème lot, No volume, B/T/Q, Préfixe section ) ont respectivement 100% et plus de 80% de valeurs, du coup on a décidé décidés de supprimer ces variables. Ensuite j'ai enlevé les variables qu'on avait utilisés pour le géocodage (Code postal, Code département, Code commune, Code voie, No voie, Commune) et la variable date de mutation. Il nous reste à la fin 13 variables et nous utilisons ces variables pour notre étude.

* **Etude Bivariéé**

Corrélation entre les variables quantitatives

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Valeur foncière | No plan | Nombre de lots | Nombre de pièces principales | Surface réelle bâti | Surface de terrain |
| Valeur foncière | 1 | -0.0011 | -0.0087 | -0.08 | 0.0028 | 0.129 |
| No plan | -0.0011 | 1 | -0.006 | 0.013 | -0.003 | -0.01 |
| Nombres de lots | -0.0087 | -0.006 | 1 | 0.003 | -0.002 | -0.007 |
| Nbre de pièces principales | -0.08 | 0.013 | 0.003 | 1 | -0.022 | -0.074 |
| Surface réelle bâti | 0.0028 | -0.003 | -0.002 | -0.022 | 1 | 0.045 |
| Surface de terrain | 0.129 | -0.01 | -0.007 | -0.074 | 0.045 | 1 |

corrélogramme avec la fonction corrplot de R



On remarque le nombre de pièces principales et surface de terrain sont corrélés avec le prix des biens. On n’utilisera pas ces variables pour expliquer le prix des biens en MCO.

Corrélation entre varia

1. **MODELISATION SPATIALE DU PRIX DES BIENS**

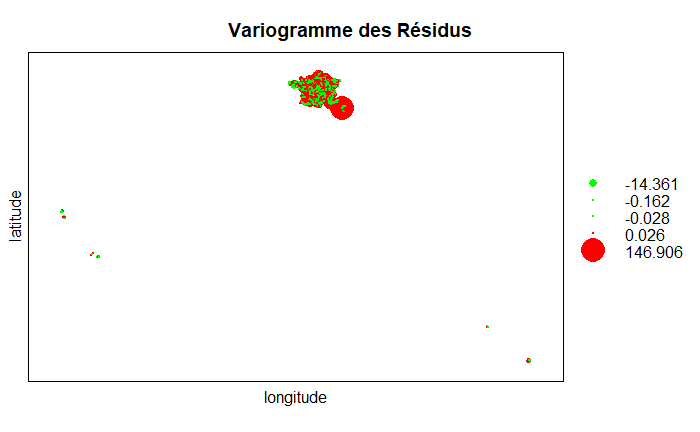
L’estimation du prix des bien réalisée avec les moindres carrés ordinaires fournit les résultats suivants :

|  |  |  |
| --- | --- | --- |
| **Variables** | **Coefficients estimés** | **P(| T|> t)** |
| No plan | 1.382e+07 | < 2e-16 |
| Nature culture (AG) | -2204000 | 7.48e-14 |
| Nature culture (B) | -5175000 | 8.33e-06 |
| Nature culture (BF) | -4550000 | 0.004086 |
| Nature culture (BM) | -4588000 | 0.279900 |
| Nature culture (BP) | -5297000 | 0.000257 |
| Nature culture (BT) | -5319000 | < 2e-16 |
| Nature culture (BR) | -5270000 | 2.30e-11 |
| Nature culture (BS) | -5135000 | 8.06e-09 |
| Nature culture (CA) | -8630000 | 0.007659 |
| Nature culture (CH) | -5934000 | 0.162207 |
| Nature culture (E) | -6264000 | 2.93e-16 |
| Nature culture (J) | -5083000 | < 2e-16 |
| Nature culture (L) | -4416000 | < 2e-16 |
| Nature culture (LB) | -4376000 | 0.715027 |
| Nature culture (P) | -5216000 | < 2e-16 |
| Nature culture (PA) | -4895000 | 2.99e-07 |
| Nature culture (PC) | -5087000 | 0.019063 |
| Nature culture (PH) | -4391000 | 0.008531 |
| Nature culture (PP) | -4406000 | 0.000281 |
| Nature culture (S) | -4194000 | < 2e-16 |
| Nature culture (T) | -5029000 | <2e-10 |
| Nature culture (TP) | -4138000 | 0.729937 |
| Nature culture (VE) | -6058000 | <2e-10 |
| Nature culture (VI) | -5028000 | 3.35e-10 |
| Surface réelle bâti | 77 | 2.85e-05 |
| Nombre de lots | -2980000 | 1.81e-11 |
| Type local (Dépendance) | -4895000 | < 2e-16 |
| Type local (Local industriel. Assimilé | -7367000 | < 2e-16 |
| Type local (Maison) | -9387000 | < 2e-16 |

D’après la sélection ascendante de l’estimation par MCO, on constate tous ces variables estimés ont un effet significatif sur le prix des biens. Le test de multiplicateur de Lagrange donne une statistique de Fisher à 30 et 271346 degrés de liberté avec une probabilité critique < 2.2e-16, on peut en déduire que le modèle MCO est globalement significatif.

Les coordonnées des sites permettent de fournir une vision exploratoire de la spatialisation des

résidus du modèle MCO.



Le modèle MCO ne permet pas d’introduire de structure spatiale sur les résidus. On ne peut

l’introduire que dans un modèle autorégressif.

**EFFECTUONS LES TESTS PAR LA METHODE DU SPECIFIQUE AU GENERAL**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SARMA** | **LM\_lag** | **LM\_err** | **RLM\_lag** | **RLM\_err** |
| **Test** | 21108 | 20946 | 20145 | 963.16 | 162.13 |
| **df** | 2 | 1 | 1 | 1 | 1 |
| **P\_value** | < 2.2e-16 | <2.2e-16 | <2.2e-16 | 0.4377 | <2.2e-16 |

La probabilité critique du test joint est inférieur au seuil de significativité de 10%, on rejette l’hypothèse de nulleté des coefficients autorégressifs ρ et λ. Donc ce modèle est significatif. D’où la nécessicité d’effectuer les tests LM\_lag et LM\_err. On peut remarquer que la statistique de test pour une alternative SEM est inférieure à celle correspondant à une alternative SAR. Pour conclure de façon plus crédible, on regarde les tests robustes à la présence de la spécification alternative de l’autocorrélation spatiale. Il s’agit pour le RLMlag de tester l’absence de terme autorégressif spatial. Lorsque le modèle contient déjà un terme autorégressif spatial dans les erreurs (RLMlag), ou inversement pour RLMerr de tester l’absence de terme autorégressif spatial dans les erreurs lorsque le modèle contient un terme autorégressif spatial. Les versions robustes RLMerr et RLMlag sont fortement significatives et que la statistique du test RLMlag est supérieur a celle du test. Nous estimons donc un modèle à effets fixes avec un processus autorégressif. C’est à dire un modèle SAR.