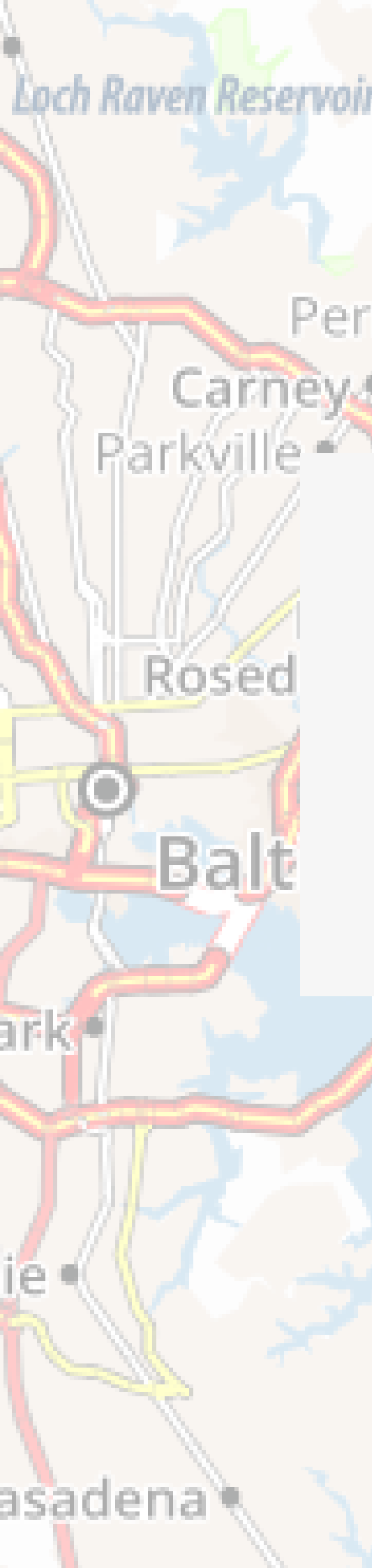


Institut du Risque & de l'Assurance
Le Mans Université

ECONOMÉTRIE SPATIALE

Prix des maisons à Baltimore

PRÉSENTÉ PAR :

Abanga Jacques
Blidi Lamia
Ndoye Alioune

Table des matières

Introduction

Étudier le prix de vente d'une maison revient à définir ses différentes caractéristiques, les facteurs qui l'entourent comme son environnement. En effet, le prix d'une maison varie selon sa localisation à un endroit précis ou encore les infrastructures qui sont proposées.

Bien qu'il soit intuitivement évident que la qualité et l'accessibilité des quartiers devraient affecter les prix des logements, dans la réalité ce n'est pas toujours le cas. Pour cause, la plupart des régressions hédoniques spatiales montrent qu'il y a peu de coefficients significatifs des variables explicatives. Une méthode alternative serait donc de modéliser l'autocorrélation résultante dans le terme d'erreur.

Dans ce projet, nous allons étudier le prix de vente des maisons à Baltimore. Pour cela, nous disposons de deux types de fichiers, un fichier excel qui est notre base de données avec les différentes variables dont nous verrons une présentation plus détaillée et un fichier de forme qui nous servira de représentation géographique des différents points dans l'Etat de Maryland où se trouve la ville de Baltimore.

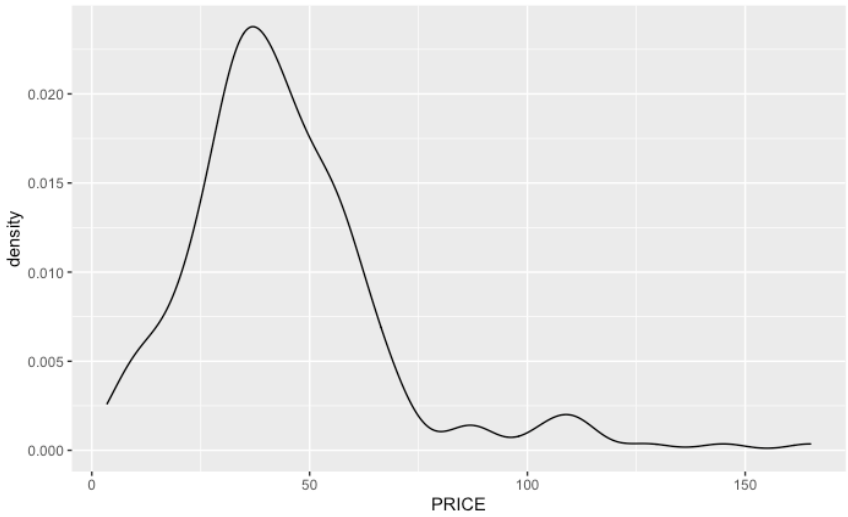
Notre base de données contient 211 observations et 17 variables, dont 6 variables discrètes. Il n'y a aucune valeur manquante dans la base.

- ***Station*** : variable ID.
- ***Price*** : prix de vente de la maison en milliers de dollars.
- ***Nroom*** : nombre de pièces.
- ***Dwell*** : variable indicatrice prenant la valeur 1 si maison individuelle, 0 si mitoyenne.
- ***Nbath*** : nombre de salle de bains.
- ***Patio*** : variable indicatrice prenant la valeur 1 si équipée d'un patio, 0 sinon.
- ***Firepl*** : 1 si cheminée, 0 sinon.
- ***Ac*** : 1 si climatisation, 0 sinon.
- ***Bment*** : 1 si sous-sol, 0 sinon.
- ***Nstor*** : nombre d'étages.
- ***Gar*** : nombre de places de parking dans le garage (0 = pas de garage).
- ***Age*** : âge du logement.
- ***Citcou*** : 1 si logement situé dans le comté de Batimore, 0 sinon.
- ***Lotsz*** : taille du terrain en centaine de pieds carrés.
- ***Sqft*** : taille de la maison en centaine de pieds carrés.
- ***X*** : coordonnée X sur la grille du Maryland.
- ***Y*** : coordonnée Y sur la grille du Maryland.

I - Analyse exploratoire

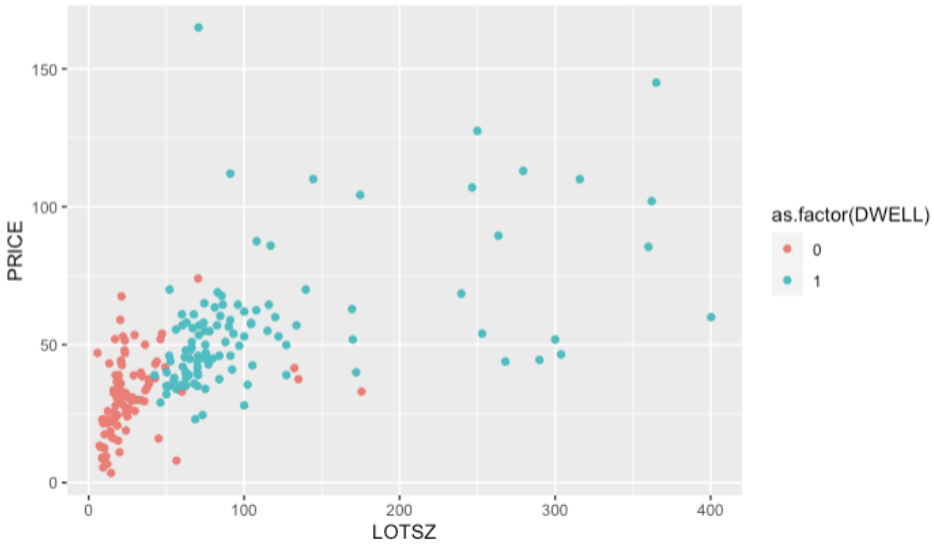
1 - Analyse descriptive

A – La distribution du Prix des maisons



La distribution de la variable « price » nous permet d'avoir une idée du prix des logements dans la ville de Baltimore. On constate que cette variable semble suivre une loi gamma. Le prix moyen des maisons à Baltimore est de 44 310$ et peut aller jusqu'à 165 000$.

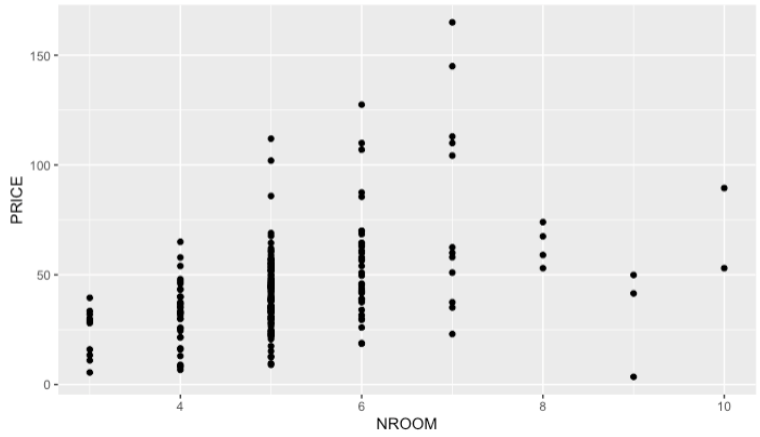
B - La répartition de la taille du terrain en fonction du prix



On observe également que les maisons mitoyennes dont la surface du terrain est inférieure à 5000ft² sont les moins chères, leur prix se situe globalement en dessous des 50 000$ malgré quelques valeurs supérieures à ce seuil. Ces maisons représentent près de la moitié de

l'échantillon. L'autre moitié est constituée majoritairement de maisons individuelles dont la surface du terrain est comprise entre 5000 et 10000ft².

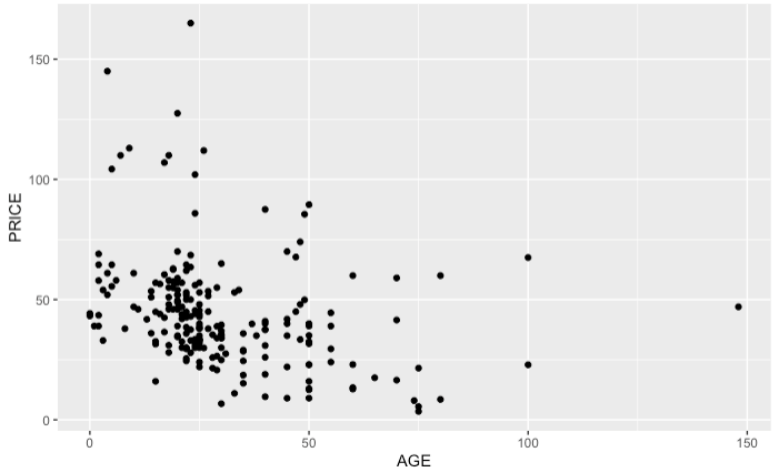
C – La répartition du prix en fonction du nombre de chambre



Concernant la variable « nroom », on constate que plus de la moitié des maisons comportent au moins 5 pièces. De plus, on peut observer un nombre non négligeable des maisons dont le prix est inférieur à 75 000$.

D – La répartition du prix en fonction de l'âge de la maison

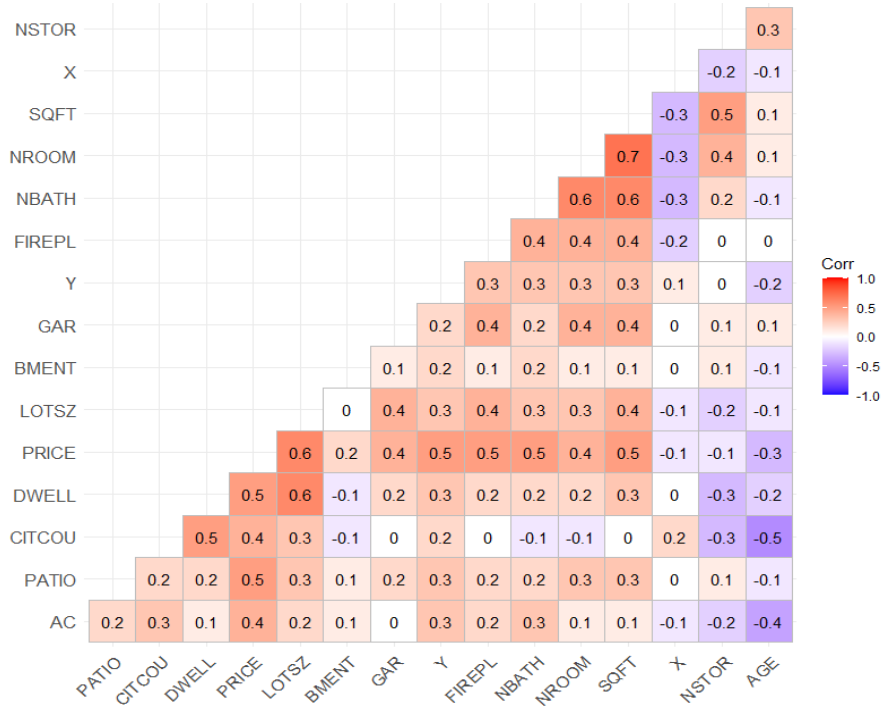
Enfin, il apparaît qu'il n'y a que très peu de logements récents. On observe une forte concentration de maisons de 20 à 25 ans qui représente environ la moitié de l'échantillon.



E – Matrice de corrélation

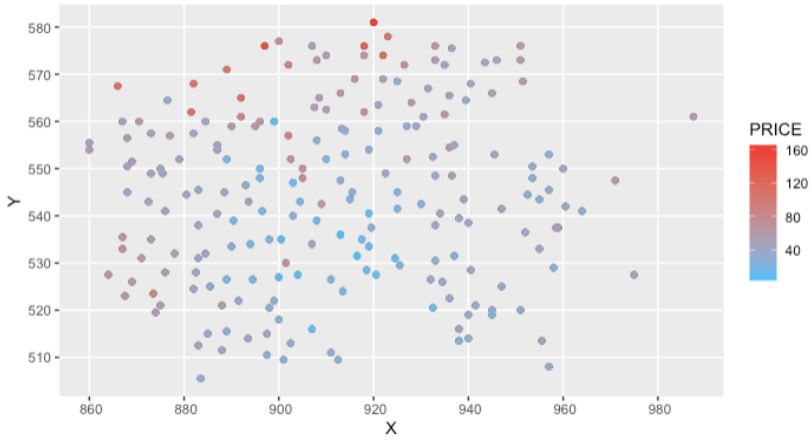
Grâce à la matrice de corrélation, on observe notamment que les prix des maisons à Baltimore sont assez fortement corrélés positivement avec la variable « LOTSZ», ce qui veut dire que plus la taille du terrain sera grande, plus son prix sera élevé. On voit aussi que PRICE est corrélé avec DWELL. Ainsi une maison non mitoyenne sera plus chère qu'une maison mitoyenne dans la région. De plus, la présence d'un patio ou d'air conditionné augmentera également le prix du

bien. Enfin, la maison sera plus chère si elle se situe à Baltimore même plutôt que dans ses alentours. A l'inverse, un nombre d'étage élevé dans la maison aura tendance à faire baisser son prix, il en va de même pour son âge.



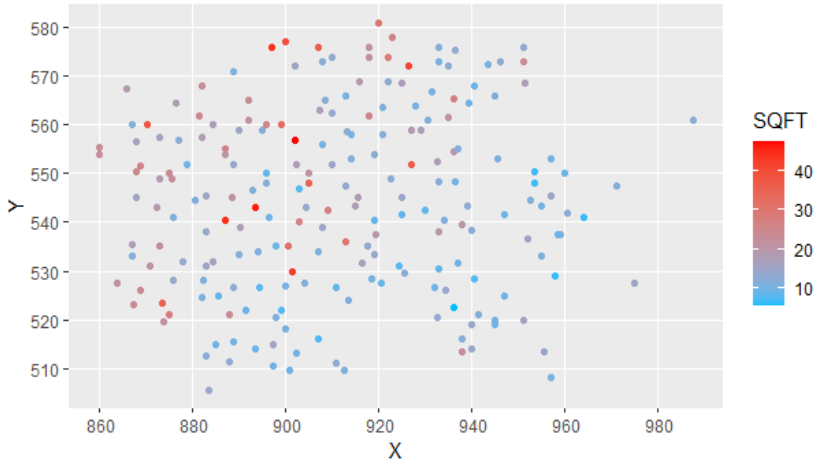
2 - Analyse spatiale exploratoire

A – Répartition spatiale du prix de vente des maisons



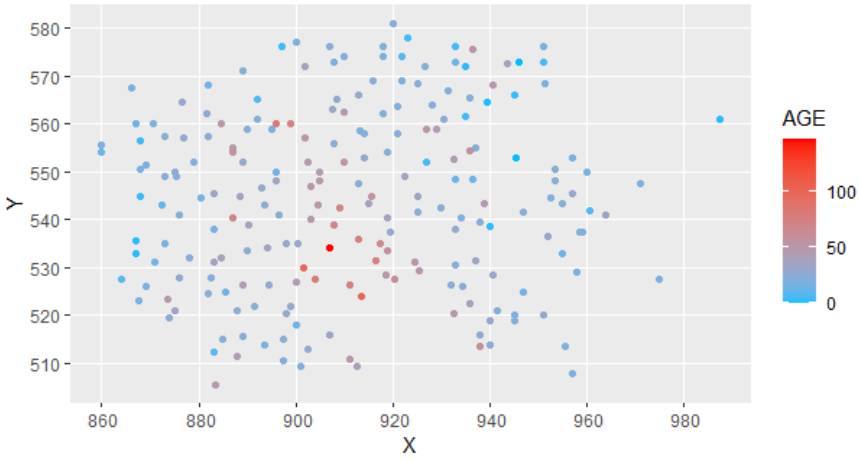
En voyant la répartition du prix dans le plan XY, on peut dire que les maisons dont le prix est le plus élevé se concentrent également dans le Nord-Ouest de la ville. Le centre-ville et le Sud de la ville sont quant à eux composés en grande majorité de logements dont le prix n'excède pas les 80 000$. On remarque ainsi une certaine autocorrélation spatiale pour le prix des maisons. Nous confirmerons cette intuition grâce à l'indice de Moran.

B - Répartition de la taille des maisons



On constate que les maisons avec une valeur de SQFT inférieure à 25 sont réparties de manières aléatoires. Les maisons qui ont les plus grands terrains, c'est-à-dire ceux qui ont une valeur de de SQFT supérieure à 30 ont l'air de plus se trouver à l'ouest de la ville.

C – Répartition spatiale de l'âge des maisons



Nous pouvons dire que les maisons les plus âgées se trouvent principalement au centre. Les plus récentes sont situées aux extrémités. Nous pouvons donc dire Baltimore est une ville qui s'aggrandit au fil des années.

D – L'indice de Moran

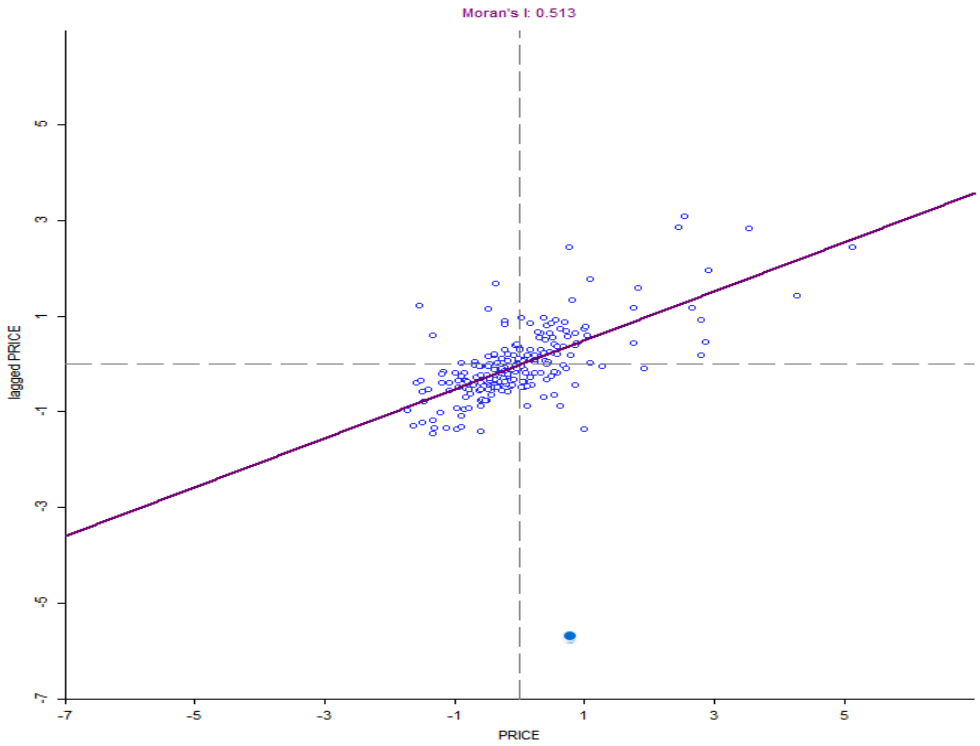
L'indice de Moran permet de mesurer l'autocorrélation spatiale. La valeur de cet indice est comprise entre -1 et 1. Les valeurs négatives de l'indice indiquent une autocorrélation spatiale négative, de même les valeurs positives indiquent une autocorrélation positive. Une valeur nulle est significative d'un modèle spatial parfaitement aléatoire.

Avec i et j les unités spatiales, n le nombre d'unité spatiales, xi : valeur de la variable dans l'unité i et wij la matrice de poids, d'interactions spatiales.

$$I_{Moran} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

E – Le diagramme de Moran

Une interprétation facile de l'indice de Moran peut être faite à l'aide du diagramme de Moran, qui représente sous la forme d'un nuage de points, les couples de valeurs correspondant à la valeur de la variable dans chaque unité spatiale (en abscisse) et la moyenne des valeurs des zones contigües (en ordonnée). Le diagramme de Moran ci-dessous nous a permis de mesurer l'autocorrélation de la variable prix. Nous pouvons constater que la droite de régression a une pente positive, ce qui veut dire qu'on est en présence d'une autocorrélation spatiale positive avec un I de Moran égal à 0.513.

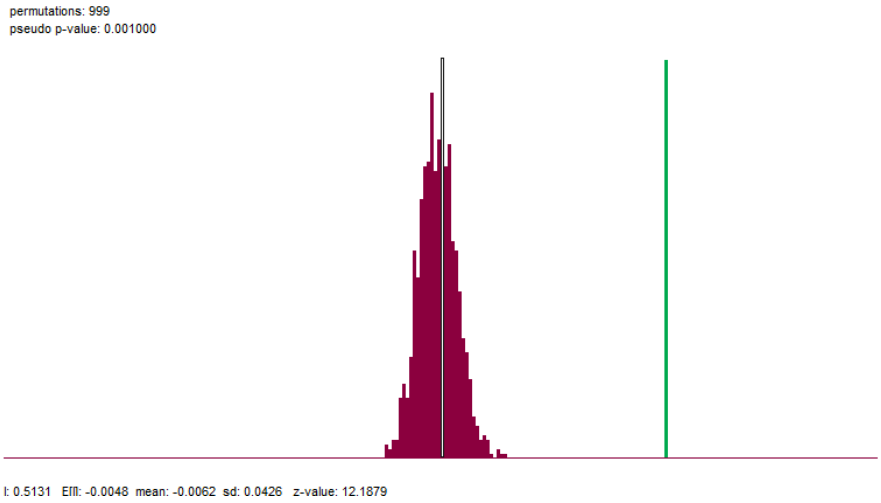


F – Significativité du test de Moran

La figure ci-dessous permet d'observer la distribution du I de Moran de la variable prix en utilisant la méthode de randomisation avec 999 permutations et la matrice de poids de 4 plus proches voisins. Nous pouvons constater que cette distribution suit une loi normale, ce qui nous permet de tester la présence d'autocorrélation spatiale en posant comme hypothèse :

$$\begin{cases} H_0 : absence\ d'autocorrélation\ spatiale \\ H_1 : présence\ d'autocorrélation\ spatiale \end{cases}$$

La p-value à l'issue du test est de 0,001 ce qui nous permet de rejeter l'hypothèse nulle d'absence d'autocorrélation spatiale au seuil de 5%, donc on accepte l'hypothèse d'une présence d'autocorrélation.



Indice de Geary

Contrairement à l'indice I de Moran qui est une mesure de l'autocorrélation spatiale globale, l'indice de Geary quant à lui mesure l'autocorrélation spatiale locale. Les valeurs de l'indice de Geary sont comprises entre 0 et 2. Une autocorrélation est parfaite lorsque la valeur de l'indice se rapproche de 0 alors qu'une valeur proche de 2 signifie qu'il y a une dispersion parfaite. Lorsque la valeur de l'indice est proche de 1, il y a une absence d'autocorrélation spatiale. L'indice de Geary varie en sens inverse de l'indice de Moran.

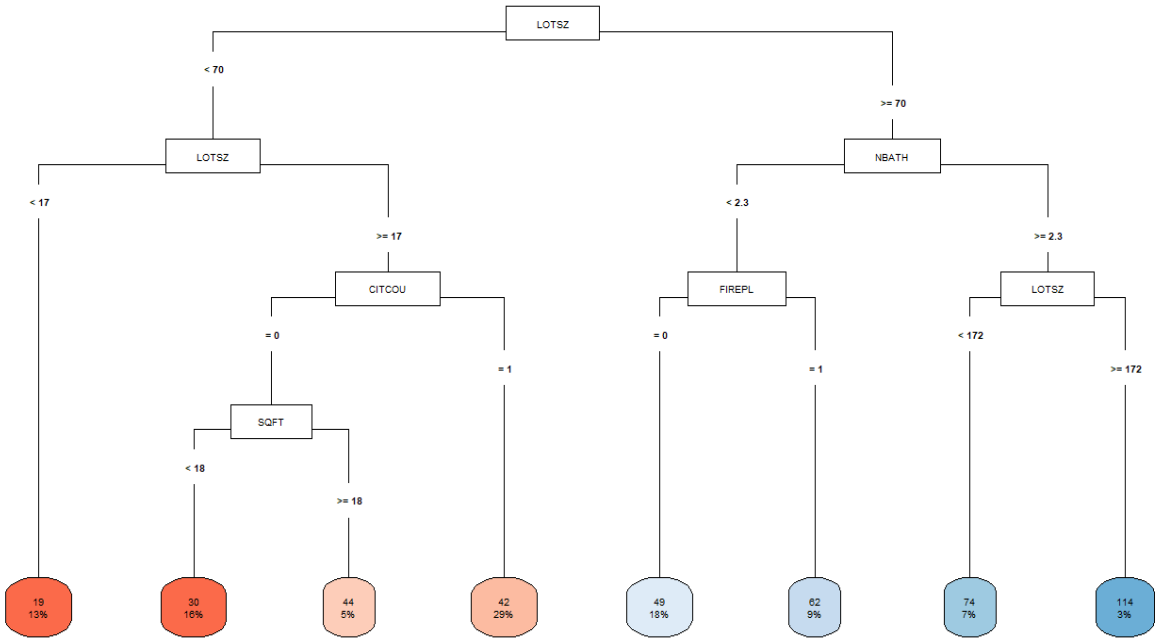
$$C_{Geary} = \frac{(n-1)\sum_{i-1}^{n}\sum_{j-1}^{n} w_{ij}\left(x_i - x_j\right)^2}{2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

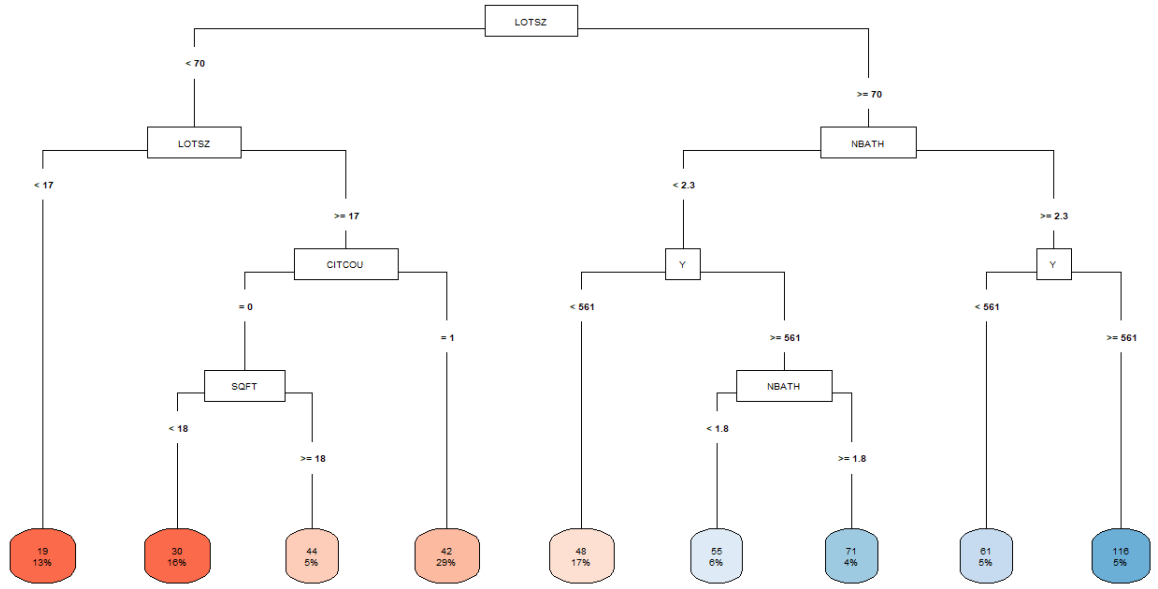
Nous obtenons le résultat suivant en faisant le test de Geary sur la variable PRICE sous R. La statistique de test est de 0.51 ce qui se rapproche de l'indice de Moran. Cela veut donc dire que nous avons de l'autocorrélation spatiale locale pour ce qui est des prix des maisons.

```
        Geary C test under randomisation

data:  baltimore$PRICE
weights: listwknn

Geary C statistic standard deviate = 13.634, p-value < 2.2e-16
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
      0.514215851       1.000000000       0.001269448
```

II - Modélisation

1 – Arbres de régression





Nous avons réalisé deux arbres de régression ; un prenant en compte les données spatiales, et un second les excluant. Il apparaît que les deux arbres ont identifié la variable relative à la surface du terrain comme racine première, c'est donc la variable la plus importante pour déterminer le prix d'une maison à Baltimore. La branche gauche, représentant les maisons avec un terrain dont la surface est inférieure à 7000ft², est identique dans les deux modèles. La

différence apparaît pour les maisons dont la surface du terrain est supérieure ou égale à 7000ft² ; dans l'arbre ne prenant pas en compte les variables spatiales, on retrouve des branches relatives au nombre de salle de bain, à la présence ou non d'une cheminée et une fois encore à la surface du terrain.
L'arbre de régression prenant en compte les données spatiales a quant à lui des branches représentant le nombre de salles de bain et la variable Y de coordonnées géographiques. Les maisons dont le terrain a une surface supérieure à 7000ft², qui dispose de 3 salles de bain et dont la longitude est supérieure à 561 aura par exemple une valeur estimée à 116 000$.

2 - Spécification du modèle économétrique par les MCO

Afin de déterminer les variables significatives du modèle, on fait un premier modèle de régression linéaire avec toutes les variables explicatives. Le R² de ce modèle, qui mesure l'adéquation entre le modèle et les données observées, est de 0,7226. Etant proche de 1, il indique que les valeurs estimées sont proches des valeurs observées.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.26748   39.87287   0.032 0.974674
NROOM        0.29099    1.08510   0.268 0.788855
DWELL1       5.34428    2.57069   2.079 0.038946 *
NBATH        5.49272    1.90592   2.882 0.004400 **
PATIO1       8.40045    2.79250   3.008 0.002978 **
FIREPL1      9.98951    2.44944   4.078 6.63e-05 ***
AC1          6.13333    2.47795   2.475 0.014179 *
BMENT1       1.30136    6.23164   0.209 0.834801
BMENT2       3.56551    2.83195   1.259 0.209540
BMENT3      10.44613    3.11718   3.351 0.000968 ***
NSTOR       -4.90795    2.85579  -1.719 0.087291 .
GAR          6.05872    1.75380   3.455 0.000677 ***
AGE          0.02007    0.05870   0.342 0.732764
CITCOU1     12.88211    2.44470   5.269 3.64e-07 ***
LOTSZ        0.03489    0.01708   2.043 0.042454 *
SQFT         0.28629    0.22047   1.299 0.195637
X           -0.07313    0.03545  -2.063 0.040450 *
Y            0.14479    0.05605   2.583 0.010533 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.43 on 193 degrees of freedom
Multiple R-squared:  0.7451,    Adjusted R-squared:  0.7226
F-statistic: 33.18 on 17 and 193 DF,  p-value: < 2.2e-16
```

De plus, il apparaît que certaines variables telles que « nroom », « bment1 » et « bment2 » ne sont pas significatives. Ces deux dernières n'étant pas significativement différentes de « bment0 », nous pouvons les regrouper dans une même catégorie, créant ainsi la variable « bmentbis ».

On réestime par la suite nous paramètres en prenant en compte la variable créée. Elle est significative dans le nouveau modèle, le coefficient R² ajusté est également plus proche de 1.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.42411   39.57973  -0.036 0.971334
NROOM        0.37039    1.07457   0.345 0.730697
DWELL1       5.14224    2.54912   2.017 0.045032 *
NBATH        5.84862    1.84476   3.170 0.001767 **
PATIO1       8.23152    2.77746   2.964 0.003417 **
FIREPL1     10.14304    2.43619   4.163 4.69e-05 ***
AC1          5.67725    2.37594   2.389 0.017820 *
Bmentbis1    7.43474    2.01246   3.694 0.000286 ***
NSTOR       -4.47000    2.81263  -1.589 0.113613
GAR          6.35242    1.72114   3.691 0.000290 ***
CITCOU1     12.11071    2.16727   5.588 7.61e-08 ***
LOTSZ        0.03744    0.01687   2.220 0.027592 *
SQFT         0.23515    0.21522   1.093 0.275892
X           -0.07372    0.03516  -2.097 0.037296 *
Y            0.15610    0.05481   2.848 0.004868 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.39 on 196 degrees of freedom
Multiple R-squared:  0.7428,    Adjusted R-squared:  0.7245
F-statistic: 40.44 on 14 and 196 DF,  p-value: < 2.2e-16
```

En étudiant la relation entre le prix et l’âge, on constate graphiquement que les deux variables n’ont pas de relation linéaire. On peut s’aider du test de Rainbow pour confirmer cela.
La p_value étant inférieure au seuil de 5%, on confirme l’absence de relation linéaire entre ces deux variables. Pour corriger cela, nous prendront en compte la non-linéarité dans les modèles suivants.

Afin de connaître le degré du polynôme de la variable âge, on crée pour chaque degré une variable correspondante que l’on intègrera à chaque modèle. On sélectionne par la suite le meilleur modèle en minimisant le MSE (mean square error). Il apparaît ainsi que le modèle ayant le meilleur pouvoir de prédiction est le modèle construit sur la base du polynôme de degré 3.

Enfin, la non-significativité de la variable « nroom » peut s’expliquer par le fait qu’il y a un nombre limité de maisons avec un grand nombre de pièces.

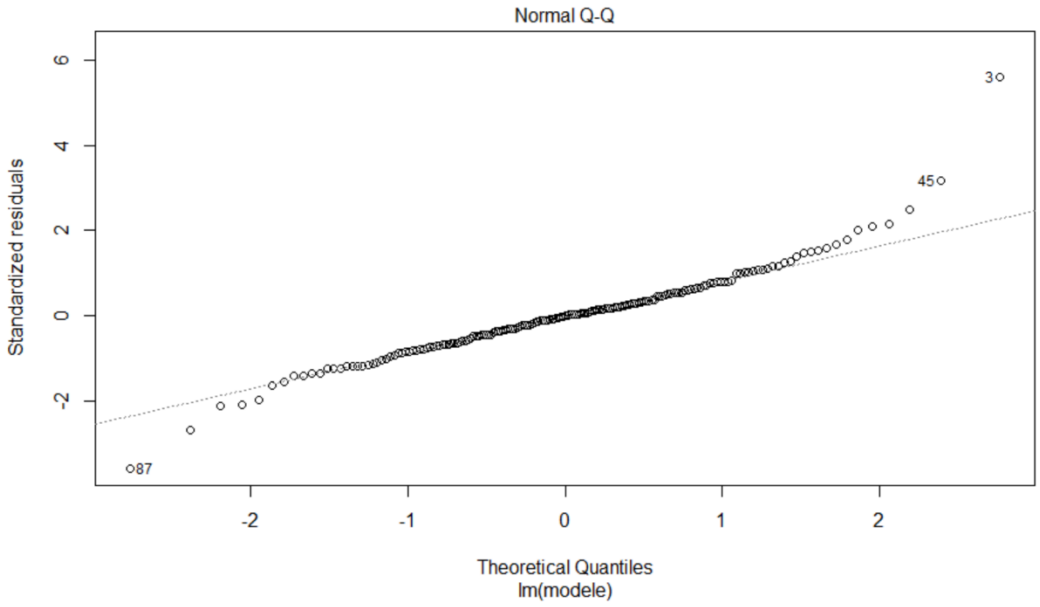
NROOM	3	4	5	6	7	8	9	10
Freq	10	35	106	39	12	4	3	2

On observe dans la répartition de la variable qu’il y a peu de maisons de plus de 7 pièces, nous regroupons donc dans une variable « nroombis » les maisons avec un nombre de pièces habitables supérieur à 7. Cette nouvelle variable apparaît comme étant significative dans le modèle.
La variable « ac1 » n’étant pas significative et ne comprenant que deux catégories (0 et 1), nous la supprimons du modèle car on ne note pas de différence significative entre ces deux catégories.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.348e+00  4.296e+01   0.218 0.828015
AGE         -2.516e-01  8.831e-02  -2.848 0.004963 **
NroomBis     1.516e+00  1.329e+00   1.141 0.255554
NBATH        3.845e+00  1.939e+00   1.983 0.049107 *
PATIO1       1.029e+01  2.914e+00   3.532 0.000538 ***
FIREPL1      8.817e+00  2.616e+00   3.370 0.000939 ***
Bmentbis1    6.439e+00  2.062e+00   3.123 0.002124 **
NSTOR       -1.144e+01  2.783e+00  -4.110 6.28e-05 ***
GAR          6.687e+00  1.774e+00   3.769 0.000229 ***
CITCOU1      1.111e+01  2.481e+00   4.478 1.42e-05 ***
LOTSZ        3.832e-02  1.670e-02   2.295 0.023012 *
SQFT         6.203e-01  2.168e-01   2.861 0.004775 **
X           -1.047e-01  3.877e-02  -2.700 0.007664 **
Y            2.163e-01  6.013e-02   3.597 0.000427 ***
age_poly3    2.277e-05  5.424e-06   4.198 4.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 12.04 on 162 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.7821,    Adjusted R-squared:  0.7633
F-statistic: 41.54 on 14 and 162 DF,  p-value: < 2.2e-16
```

Le modèle estimé est significatif, avec une p_value inférieure au seuil de 0,1% et un fort pouvoir explicatif, le coefficient de détermination R^2 étant de 0,7821. Les termes d'erreurs associés à chaque estimation sont également très faible, le modèle est donc stable.

Par la suite, nous avons réalisé un test d'autocorrélation des résidus de Durbin-Watson. L'interprétation de ce test se base sur la statistique de test DW. Une statistique de test proche de 0 indique la présence d'autocorrélation positive des résidus, elle représente une autocorrélation négative lorsqu'elle tend vers 4 et démontre l'absence d'autocorrélation des résidus lorsqu'elle est proche de 2. Ainsi, il apparaît que ce modèle ne présente pas d'autocorrélation des résidus.



Enfin, le QQ-plot nous permet de tester la normalité des résidus. Les résidus semblent ici suivre une loi normale, il y aurait donc un faible écart entre les valeurs observées et les valeurs prédites, mais on observe cependant un biais aux valeurs extrêmes. Nous avons donc réalisé un test de Shapiro-Wilk afin de tester la normalité des résidus.

La p_value du test est significative, on rejette donc l'hypothèse nulle de normalité des résidus, les estimateurs ne sont donc pas optimaux.
D'autre part, le graphique des résidus nous amène à penser que l'hypothèse d'homoscédasticité risque de ne pas être respectée (car on observe une forme de cône).
Enfin, il existe certainement de la dépendance spatiale concernant les prix des logements. En effet, le diagramme de Moran affiche un indice de 0,513. Il est donc nécessaire de s'aider des modèles d'économétrie spatiale afin de prendre en compte ces phénomènes spatiaux.

3 - Spécification du modèle économétrique spatial

Afin de savoir vers quel modèle spatial s'orienter, nous réalisons des tests du multiplicateur de Lagrange.

```
        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = model, data = baltimore)
weights: listwknn

LMErr = 9.9798, df = 1, p-value = 0.001583


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = model, data = baltimore)
weights: listwknn

LMlag = 15.037, df = 1, p-value = 0.0001054


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = model, data = baltimore)
weights: listwknn

RLMerr = 0.61515, df = 1, p-value = 0.4329


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = model, data = baltimore)
weights: listwknn

RLMlag = 5.6722, df = 1, p-value = 0.01724
```

Le premier a pour hypothèses :

$$\begin{cases} H_0 : absence\ d'autocorrélation\ spatiale\ (\lambda = 0)\ sous\ hypothèse\ \rho = 0 \\ H_1 : présence\ d'autocorrélation\ spatiale\ (\lambda \neq 0) \end{cases}$$

La p_value (0,001583) étant inférieure au seuil de 1%, on rejette l'hypothèse nulle, on conclut alors à la présence d'autocorrélation spatiale, on ne peut à ce stade pas choisir le modèle SEM.
Le second a pour hypothèses :

$$\begin{cases} H_0 : absence\ d'autocorrélation\ spatiale\ (\rho = 0)\ sous\ hypothèse\ \lambda = 0 \\ H_1 : présence\ d'autocorrélation\ spatiale\ (\rho \neq 0) \end{cases}$$

La p_value (0,0001054) étant inférieure au seuil de 1%, on rejette l'hypothèse nulle, on conclut alors à la présence d'autocorrélation spatiale, on ne peut à ce stade pas choisir le modèle SAR.
Les deux hypothèses nulles des tests LM étant rejetées, nous réalisons par la suite les tests robustes du multiplicateur de Lagrange.
Le premier a pour hypothèses :

$$\begin{cases} H_0 : \theta + \rho\beta = 0 \\ H_1 : \theta + \rho\beta \neq 0 \end{cases}$$

La p_value (0,4329) étant supérieure au seuil de 5%, on accepte l'hypothèse nulle. Nous rejetons donc le modèle SEM.
Le second a pour hypothèses :

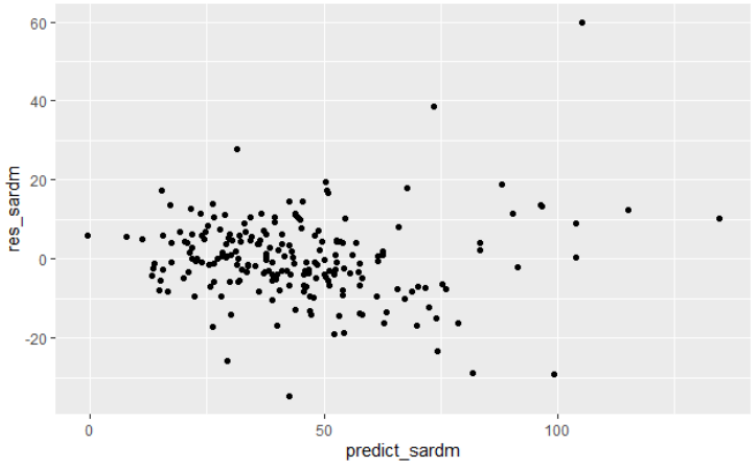
$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$

Enfin, le test robuste du multiplicateur de Lagrange relatif au modèle SAR affiche une p_value inférieure au seuil de 5%, on rejette donc l'hypothèse nulle. Le modèle SAR est donc le modèle spatial le plus adapté à la modélisation du prix des maisons à Baltimore.

L'approche ascendante de choix de modèle économétrique nous a donc fait sélectionner le modèle autorégressif spatial (SAR). Il reprend les caractéristiques de la régression linéaire mais inclut également la dépendance entre la variable à expliquer et sa valeur de voisinage. Il apparaît ainsi que les prix des maisons à Baltimore dépendent des variables explicatives sélectionnées par MCO mais aussi des prix des maisons voisines.
Ce modèle est de la forme :

$$Y = \rho WY + X\beta + \epsilon$$

```
Coefficients: (asymptotic standard errors)
              Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  2.0084e+01  3.6373e+01  0.5522 0.5808270
AGE         -2.5344e-01  7.5170e-02 -3.3716 0.0007474
NroomBis     2.7419e+00  1.0814e+00  2.5355 0.0112279
NBATH        4.7502e+00  1.6495e+00  2.8797 0.0039805
PATIO1       7.2885e+00  2.4857e+00  2.9322 0.0033654
FIREPL1      7.5355e+00  2.1926e+00  3.4367 0.0005888
Bmentbis1    7.1771e+00  1.8365e+00  3.9080 9.307e-05
NSTOR       -9.8396e+00  2.2922e+00 -4.2926 1.766e-05
GAR          6.8182e+00  1.5642e+00  4.3589 1.307e-05
CITCOU1      7.3958e+00  2.3584e+00  3.1359 0.0017132
LOTSZ        3.9803e-02  1.4814e-02  2.6868 0.0072137
SQFT         4.2770e-01  1.7971e-01  2.3800 0.0173132
X           -2.3391e-02  3.4495e-02 -0.6781 0.4977191
Y            2.1927e-02  6.3648e-02  0.3445 0.7304630
age_poly3    2.3945e-05  4.8427e-06  4.9446 7.630e-07

Rho: 0.3663, LR test value: 12.839, p-value: 0.00033953
Asymptotic standard error: 0.093947
    z-value: 3.8989, p-value: 9.6611e-05
Wald statistic: 15.202, p-value: 9.6611e-05

Log likelihood: -808.6076 for lag model
ML residual variance (sigma squared): 123.38, (sigma: 11.108)
Number of observations: 211
Number of parameters estimated: 17
AIC: 1651.2, (AIC for lm: 1662.1)
LM test for residual autocorrelation
test value: 0.21823, p-value: 0.64039
```



Le modèle est significatif, il affiche une p_value inférieure au seuil de 1% et quasiment toutes les variables ont gardé leur effet explicatif sur le modèle. De plus, les erreurs associées à chaque estimation sont plus ou moins proches de 0, le modèle est donc stable et a un bon pouvoir prédictif. On remarque l'âge a un effet négatif sur le prix, une maison perdra ainsi de sa valeur avec le temps. Le nombre d'étages a également cet effet sur la variable prix.
Les résidus n'ont quant à eux pas de structure et sont concentrés dans l'intervalle [-20 ; 20] malgré quelques points aberrants.

```
        Moran I test under randomisation

data:  sar$residuals
weights: listwknn

Moran I statistic standard deviate = 0.53274, p-value = 0.2971
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.0090136577     -0.0047619048      0.0006686289
```

Le test de Moran appliqué au modèle SAR nous affiche une p_value supérieure au seuil de 5%, on rejette ainsi l'hypothèse nulle de présence d'autocorrélation spatiale.

Conclusion

L'autocorrélation spatiale est une situation qui génère une relation de dépendance entre les observations du fait de leur localisation dans l'espace ; en présence d'autocorrélation spatiale il n'est plus possible de réaliser une estimation par la méthode de Moindres Carré Ordinaires (MCO) car les estimateurs son biaisés et inefficients.

Les tests d'autocorrélation spatiale, et notamment le test de Moran, nous ont permis de détecter l'existence d'autocorrélation spatiale dans la répartition du prix de vente des maisons à Baltimore. Pour le choix du modèle d'économétrie spatial adapté, nous avons en amont spécifié un modèle à l'aide des MCO ce qui nous a permis de choisir les variables à intégrer dans le modèle. Suite à cela, nous avons fait le test du multiplicateur de Lagrange afin de savoir vers quel modèle d'économétrie s'orienter. Cela nous a ainsi conduit à choisir le modèle SAR.
On observe ainsi pour ce modèle une grande influence du nombre d'étages qui est une fonction décroissante du prix. Sachant que les logements avec beaucoup d'étages sont souvent des logements sociaux, nous pouvons ainsi comprendre cet effet. La seconde variable la plus influente du modèle nous permet de savoir si le logement se trouve à Baltimore ou bien à proximité. Les logements qui se trouvent donc dans Baltimore seront plus chers que le reste. Les variables comme fireplace, patio et bmentbis influent de la même manière le prix, c'est-à-dire positivement.

Bibliographie

Les indices d'autocorrélation spatiale, https://oasis.irstea.fr/wp-content/uploads/2013/10/10-Autocorrélation.pdf

L'indice de Geary, https://fr.wikipedia.org/wiki/Indice_de_Geary

https://rural-urban.eu/sites/default/files/05_Spatial%20Autocorrelation%20and%20the%20Spatial%20Durbin%20Model_Eilers.pdf

http://iml.univ-mrs.fr/~reboul/SP3.pptx.pdf