Projet de Big Data n°2

Ugo Devoille & Zoé Joubert

Janvier 2021

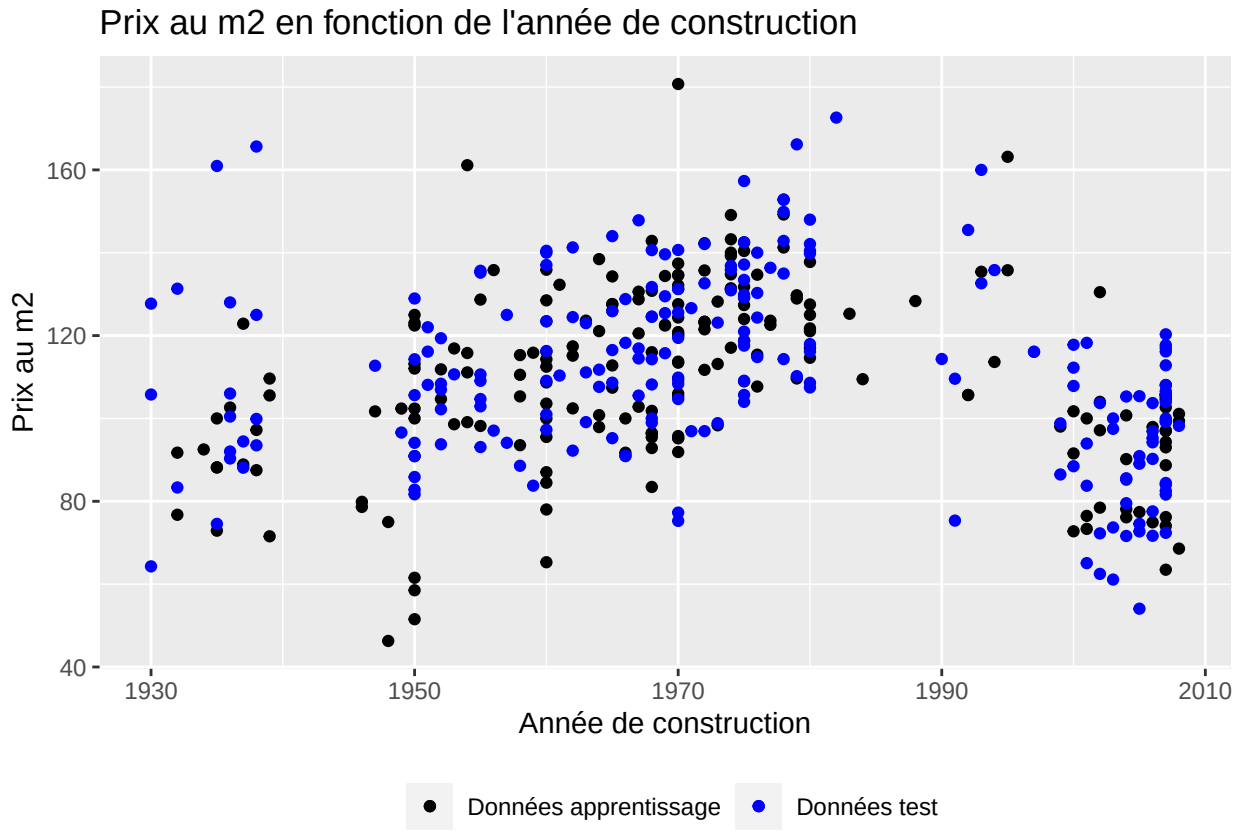
Contents

1 Introduction

Le jeux de données qui va être étudié ici concerne les prix de vente des appartements de Varsovie (Pologne). Celui ci contient des informations telles que la surface en mètres carrés, l'étage, le nombre de chambres, l'année de construction et le district. De plus, on trouve une variable (le prix au mètre carré), que nous chercherons à prédire ici. Pour cela, nous utiliserons dans un premier temps uniquement la variable de l'année de la construction pour faire notre prédiction, puis nous utiliserons toutes les variables explicatives. Pour toutes nos méthodes, nous avons fait le choix de diviser la base données en deux (partie test et partie apprentissage) afin de calculer un critère MSE qui a servi à déterminer la meilleure méthode.

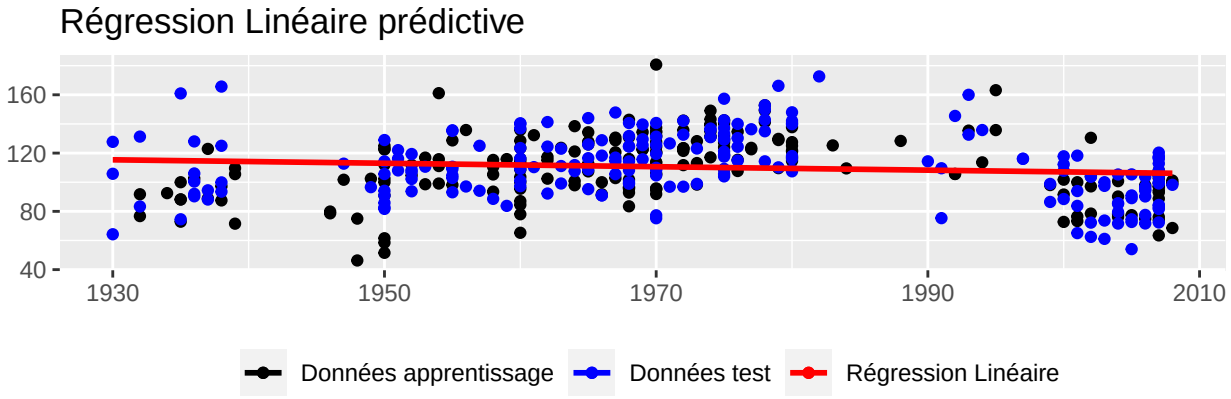
2 Régression avec la date de construction

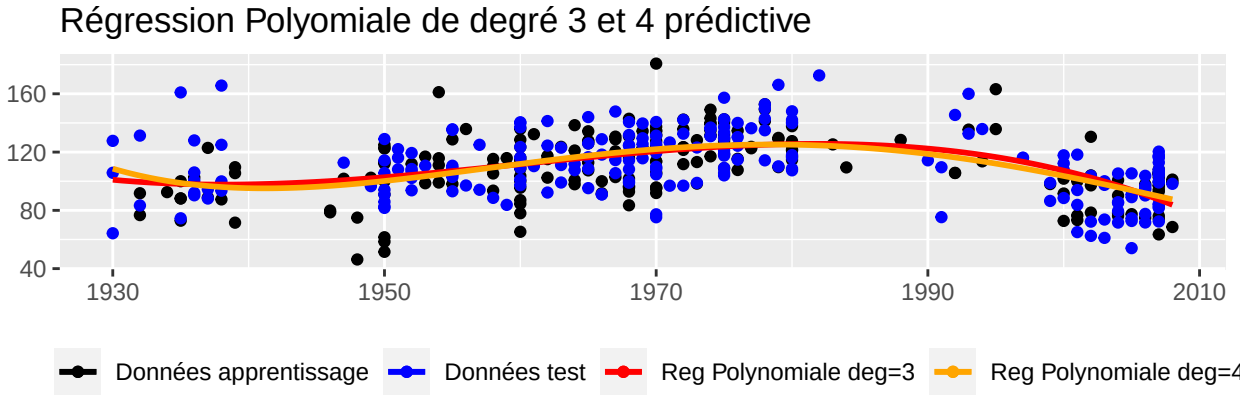
Cette première partie sera très visuelle car nous avons deux variables quantitatives, ainsi tracer un nuage de points avec les différentes régression tentant de prédire sera intéressant. Nous avons fait le choix de prendre cette variable explicative en particulier, car nous souhaitions une variables quantitative, et que la surface en mètre carrée ne peut pas vraiment expliquer le prix du mètre carré (assez logiquement). Cela est affirmé en visualisant le nuage de point, qui apporte trop peu d'informations. On notera que l'année de construction est une variable "qualitative", mais en réalité on peut facilement l'adapter ici pour en faire une variable quantitative, au vue des nombreuses modalitées qui sont de plus une approximation de la réalité (on approxime un instant t par une année). Enfin, c'est le seul nuage de points pour lequel il semble possible d'utiliser des régressions un minimum intéressantes. On retrouve donc le nuage de points de nos deux variables :



2.1 Quelques méthodes non-concluantes

Avant de montrer quelle est la meilleure méthode pour prédire le prix du mètre carré à l'aide de l'année de construction, on peut montrer quelques autre méthodes qui sont non adaptées au jeu de données. On retrouve évidemment la régression linéaire, qui n'a aucun intérêt dans ce cas précis.

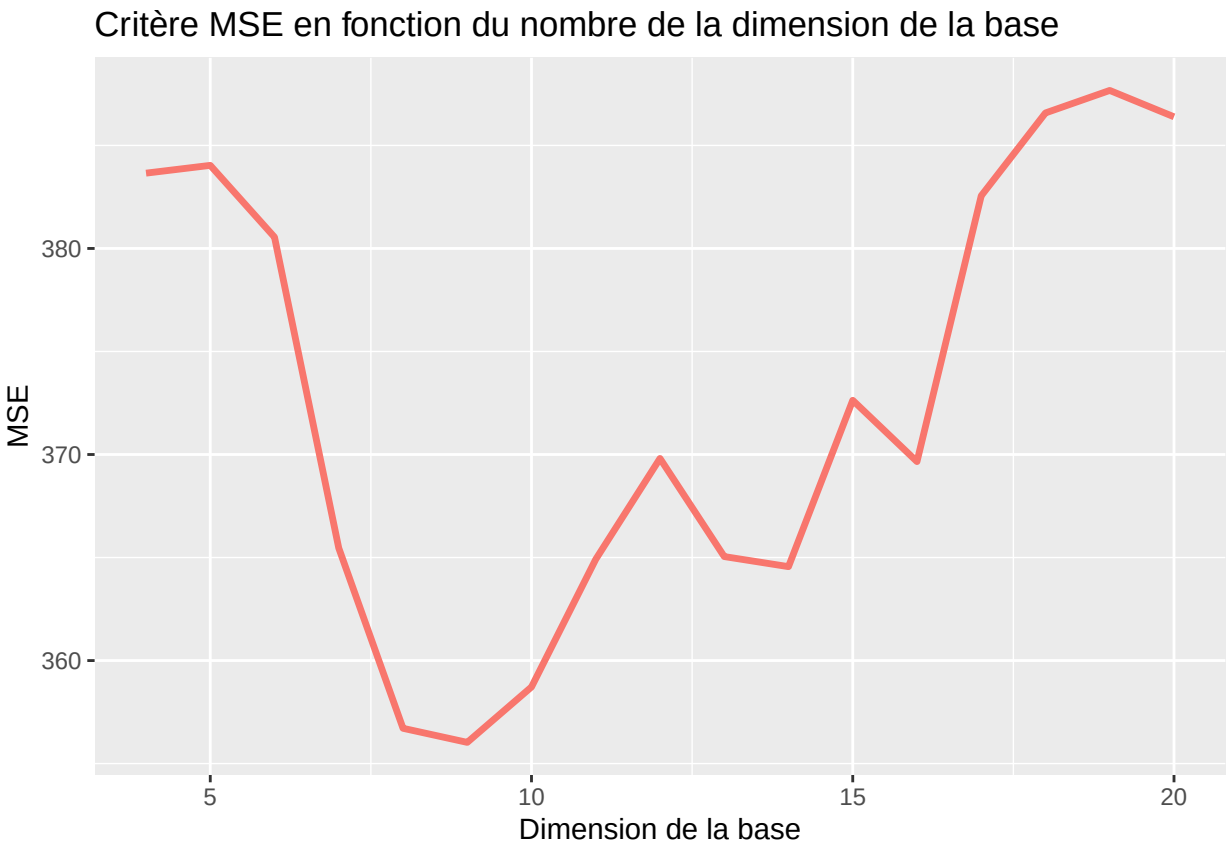




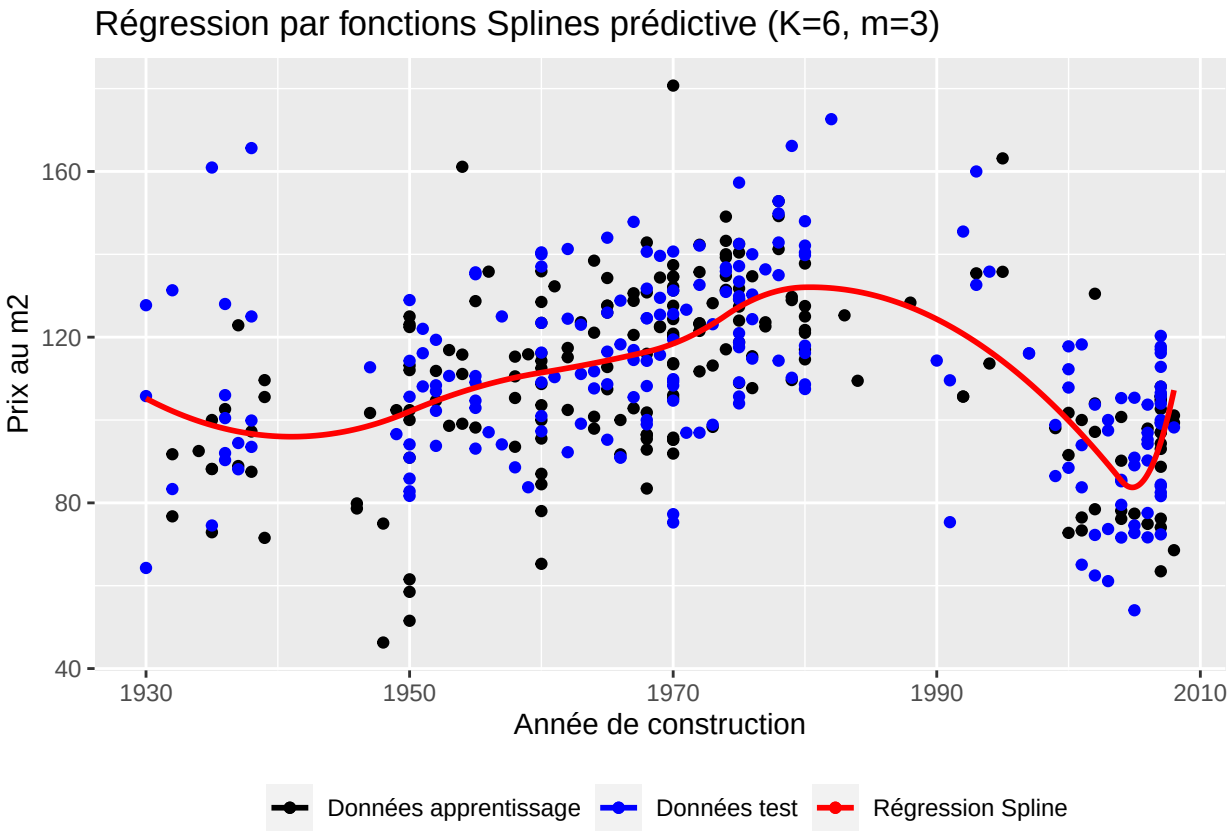
Cette régression affiche un critère MSE de 500.4640641, qui semble vraiment mauvais. On peut ensuite continuer en utilisant cette fois ci les régressions polynomiale. On étudie ici celles de de degré 3 et 4.

On constate que ces deux régressions font une meilleure prédictions des données. Elles semblent meilleure pour ce jeu précisément. Cependant, avec des critère MSE respectifs de 389.330832 et 381.804829, on espère tout de même trouver un peu mieux par la suite.

On peut aussi utiliser la régression avec les fonctions B-Splines. Pour celles ci, il est nécessaire de trouver la dimension de la base optimale qui minimise le critère MSE. Nous avons donc créé un graphe donnant ce critère en fonction de la dimension de la base. On a aussi choisi de fixer m=3 par convention.

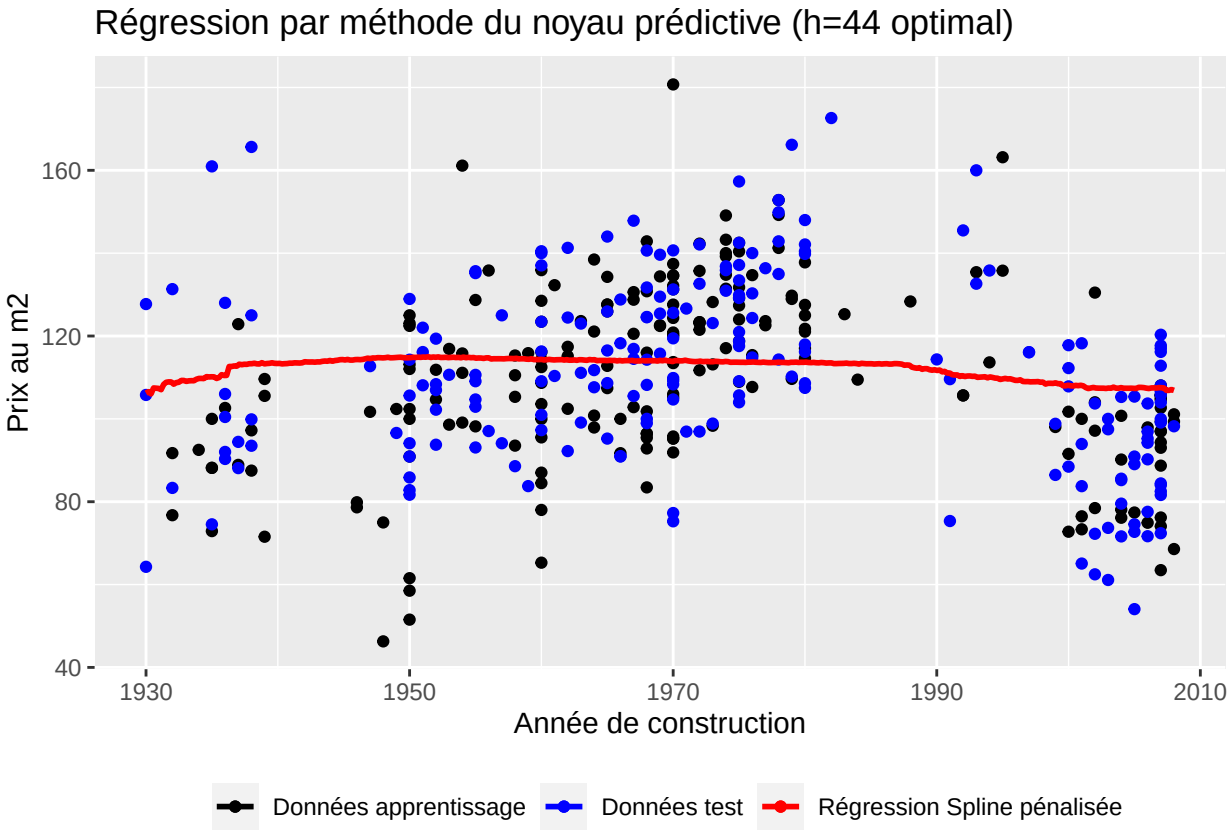


Grâce à ce graphe, on observe que le MSE est minimal pour df=9, ce qui signifie que K=6 (le nombre de noeuds) et m=3 (le degré du polynome).



On obtient par conséquent une régression par fonctions splines encore meilleure que les précédentes, avec un critère MSE de 356.0285889. Cette méthode prédit donc bien les données, mais il semble possible de faire encore un peu mieux.

Dans un dernier temps, il est possible de construire un estimateur à noyau de A à Z (Création de code plutôt qu'utilisation de packages). Nous allons pour cela utiliser le noyau d'Epanechnikov, qui est défini par $K(u) = 3/4(1-u^2)1_{\{|u|\leq 1\}}$. Le choix de ce noyau est subjectif, il est possible d'en prendre d'autres. Cependant, nous avons observés de moins bon résultats avec les noyaux gaussiens, rectangulaires, ou encore triangulaires.

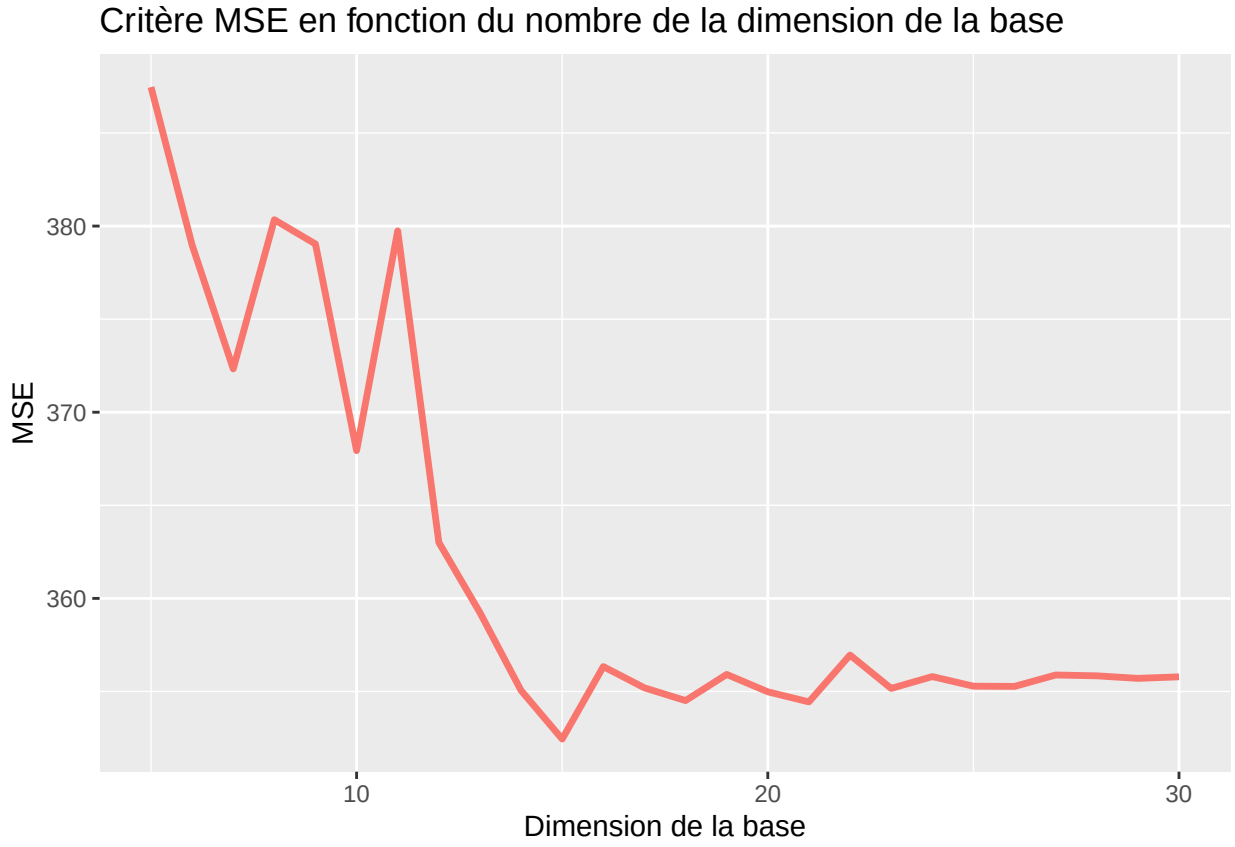


Avec cette méthode, nous obtenons le graphe ci-dessus. On remarque que la courbe varie très peu, malgré l'optimisation du paramètre h qui est ici estimé à 44. Pour cette valeur, on observe alors un critère MSE de `mini`. En fin de compte, cette méthode ne s'avère pas très concluante. On se tourne alors vers la prochaine, qui sera aussi la meilleure.

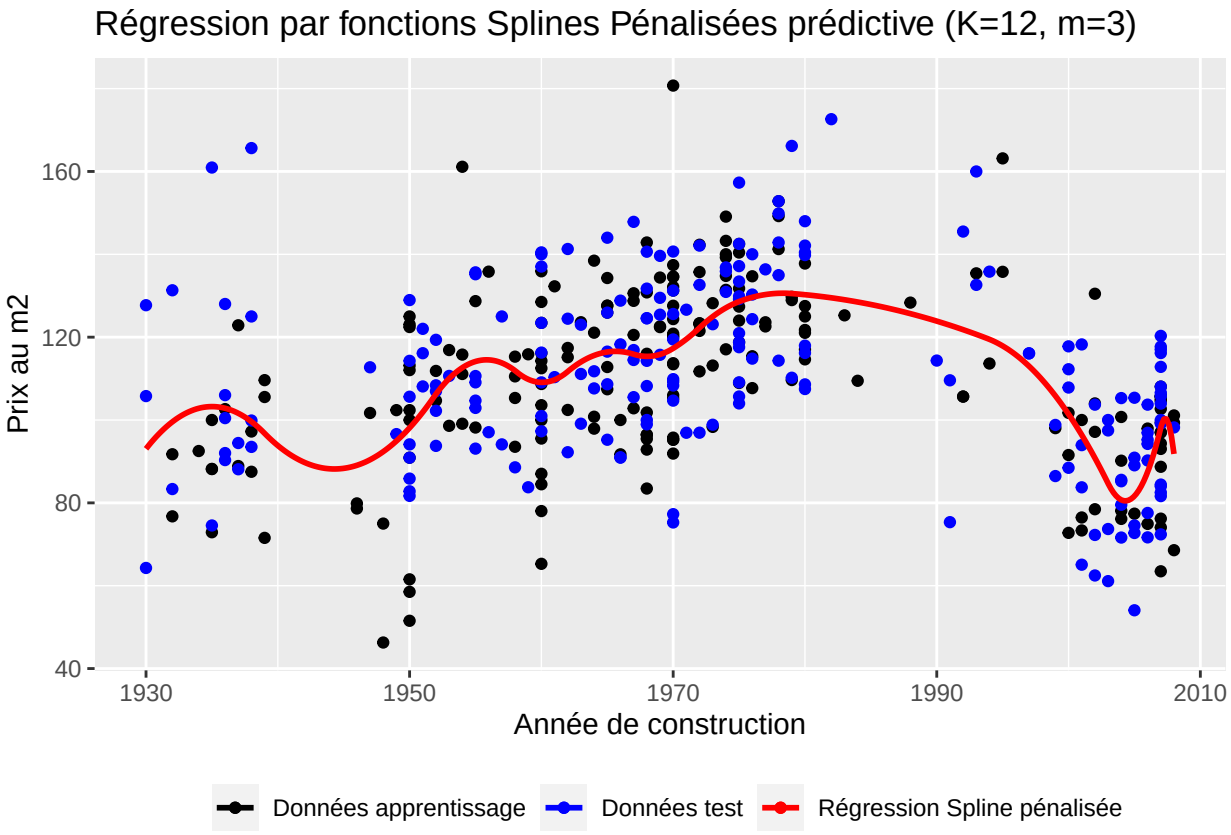
2.2 La meilleure méthode

La méthode des splines pénalisées est une extension de la méthode utilisant les fonctions B-splines, auxquelles on applique une pénalitée sur les dérivées des fonctions.

Ici, nous allons fixer le degré des fonctions B-Splines à 3, et aussi fixer la pénalité uniquement sur les dérivées secondes. Enfin, on choisit alors de faire varier une nouvelle fois la dimension de la base df, et de créer un graphe qui représente le critère MSE en fonction de celle ci, le but étant de minimiser ce critère.



On constate que pour df = 15, ce critère est minimal. Cela correspond à 12 noeuds, et un degré de 3 pour les fonctions Splines.



Avec cette méthode, on obtient un critère MSE de 352.4423532, qui est encore meilleur par rapport aux méthodes précédentes.

3 Analyse avec plusieurs variables explicatives

Dans cette partie, on va à nouveau expliquer le prix au mètre carré des appartements en prenant en compte cette fois-ci plusieurs variables explicatives : la surface, le nombre de chambres, le quartier, l'étage et l'année de construction. On utilise toujours la fonction `gam` du package `mgcv` pour estimer des modèles additifs non paramétriques ou semi-paramétriques.

On remarque la présence d'une variable explicative qualitative : le district. On décide de considérer deux scénarios : - la variable District a un effet linéaire sur le prix au mètre carré - la variable District a un effet non-linéaire sur le prix au mètre carré.

On fait d'abord une analyse non-paramétrique.

On distingue donc deux cas :

- Cas ou la variable District a un effet linéaire :
- Cas ou la variable District a un effet non-linéaire :

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## areaPerMzloty ~ district + s(n.rooms, k = 4, by = district) +
##     s(surface, k = 4, by = district) + s(construction.date, k = 4,
##     by = district) + s(floor, k = 4, by = district)
##
## Parametric coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         113.0377     1.2879  87.768  < 2e-16 ***
## districtSrodmiescie -12.3686     2.2082  -5.601 4.13e-08 ***
## districtWola          2.3395     3.0909   0.757    0.450
## districtZoliborz     -0.8918     3.8636  -0.231    0.818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                           edf Ref.df      F  p-value
## s(n.rooms):districtMokotow              1.409  1.704  2.053  0.23089
## s(n.rooms):districtSrodmiescie          1.354  1.622  2.720  0.04906 *
## s(n.rooms):districtWola                 1.000  1.000  0.363  0.54703
## s(n.rooms):districtZoliborz             1.000  1.000  0.047  0.82872
## s(surface):districtMokotow              2.714  2.931  4.160  0.01046 *
## s(surface):districtSrodmiescie          2.922  2.992  4.574  0.00399 **
## s(surface):districtWola                 1.965  2.222  3.565  0.03004 *
## s(surface):districtZoliborz             1.507  1.798  0.557  0.59074
## s(construction.date):districtMokotow    2.869  2.985 29.422  < 2e-16 ***
## s(construction.date):districtSrodmiescie 2.961  2.998  9.587 4.42e-06 ***
## s(construction.date):districtWola       1.916  2.241  1.782  0.18588
## s(construction.date):districtZoliborz   2.698  2.927  5.146  0.00648 **
## s(floor):districtMokotow                1.000  1.000  1.780  0.18301
## s(floor):districtSrodmiescie            2.825  2.972  3.936  0.00566 **
## s(floor):districtWola                   1.000  1.000  4.242  0.04012 *
## s(floor):districtZoliborz               1.000  1.000  2.927  0.08792 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.457   Deviance explained = 50.1%
## GCV =  290.1  Scale est. = 265.88    n = 409
```

Nous n'avons affiché uniquement le second modèle, puisque l'information qu'il contient n'est pas nécessaire dans le rapport. La même chose sera faite sur les prochains modèles.

On compare les deux modèles emboités via une analyse de variance (un test de Fisher) :

```
## Analysis of Deviance Table
##
## Model 1: areaPerMzloty ~ district + s(n.rooms, k = 4) + s(surface, k = 4) +
##     s(construction.date, k = 4) + s(floor, k = 4)
## Model 2: areaPerMzloty ~ district + s(n.rooms, k = 4, by = district) +
##     s(surface, k = 4, by = district) + s(construction.date, k = 4,
##     by = district) + s(floor, k = 4, by = district)
##   Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
## 1    394.15     112864
## 2    372.61      99669 21.542    13195 2.3038 0.0009414 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette le 1er modèle non-paramétrique. Celui où on considère que l'effet du district sur le prix au mètre carré est non-linéaire est donc meilleur.

On peut ensuite faire un simple modèle linéaire pour le comparer avec le 2ème modèle que l'on a obtenu. Celui ci montres que les districts Wola et Zoliborz sont les seules variables qui n'ont pas d'impact sur le prix au M2.

On compare à nouveau les deux modèles emboités via une analyse de variance :

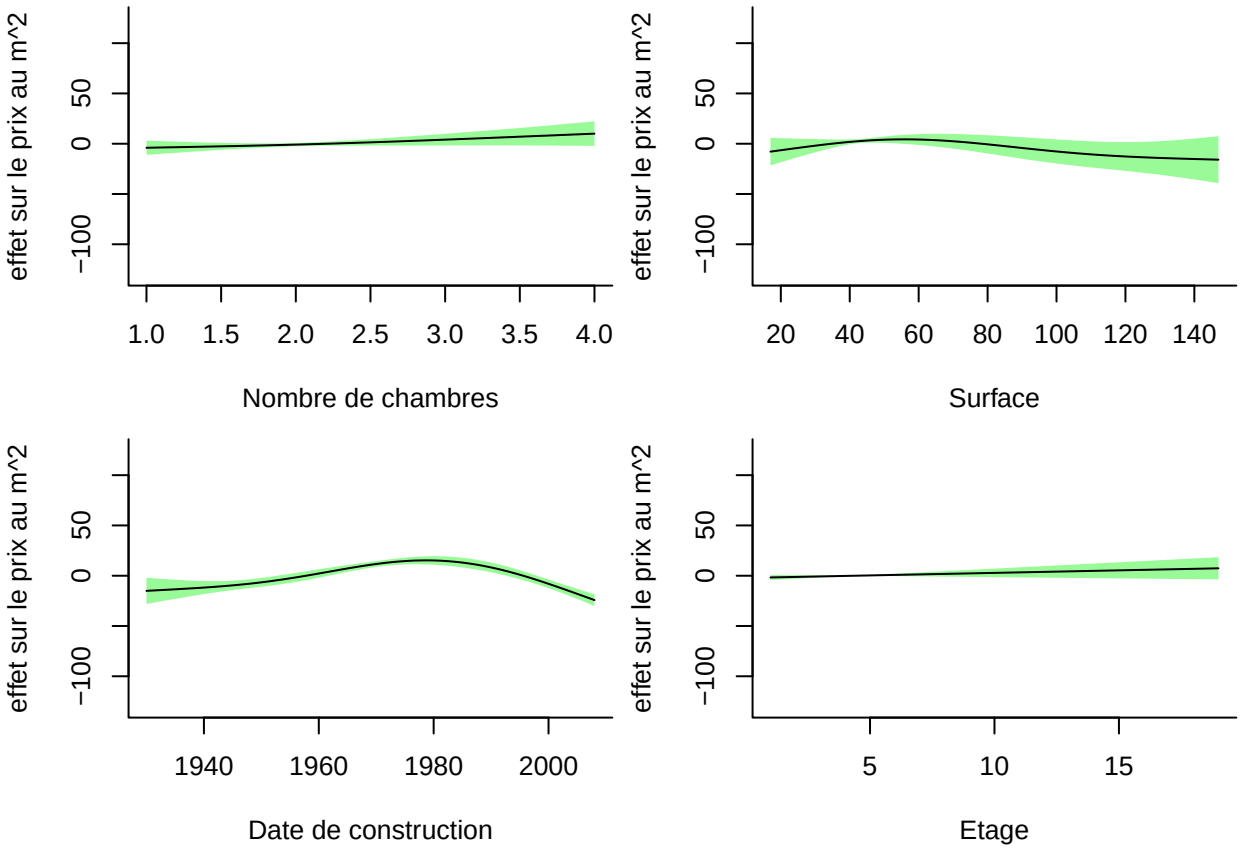
```
## Analysis of Deviance Table
##
## Model 1: areaPerMzloty ~ district + s(n.rooms, k = 4, by = district) +
##     s(surface, k = 4, by = district) + s(construction.date, k = 4,
##     by = district) + s(floor, k = 4, by = district)
## Model 2: areaPerMzloty ~ district + n.rooms + surface + construction.date +
##     floor
##   Resid. Df Resid. Dev      Df Deviance     F    Pr(>F)
## 1    372.61      99669
## 2    401.00     157586 -28.393   -57918 7.672 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value est très faible, on rejette donc l'hypothèse H_0 : le modèle linéaire.

On trace l'effet des différents prédicteurs sur le prix au mètre carré afin de visualiser si les effets de chaque variable sont linéaire ou non.

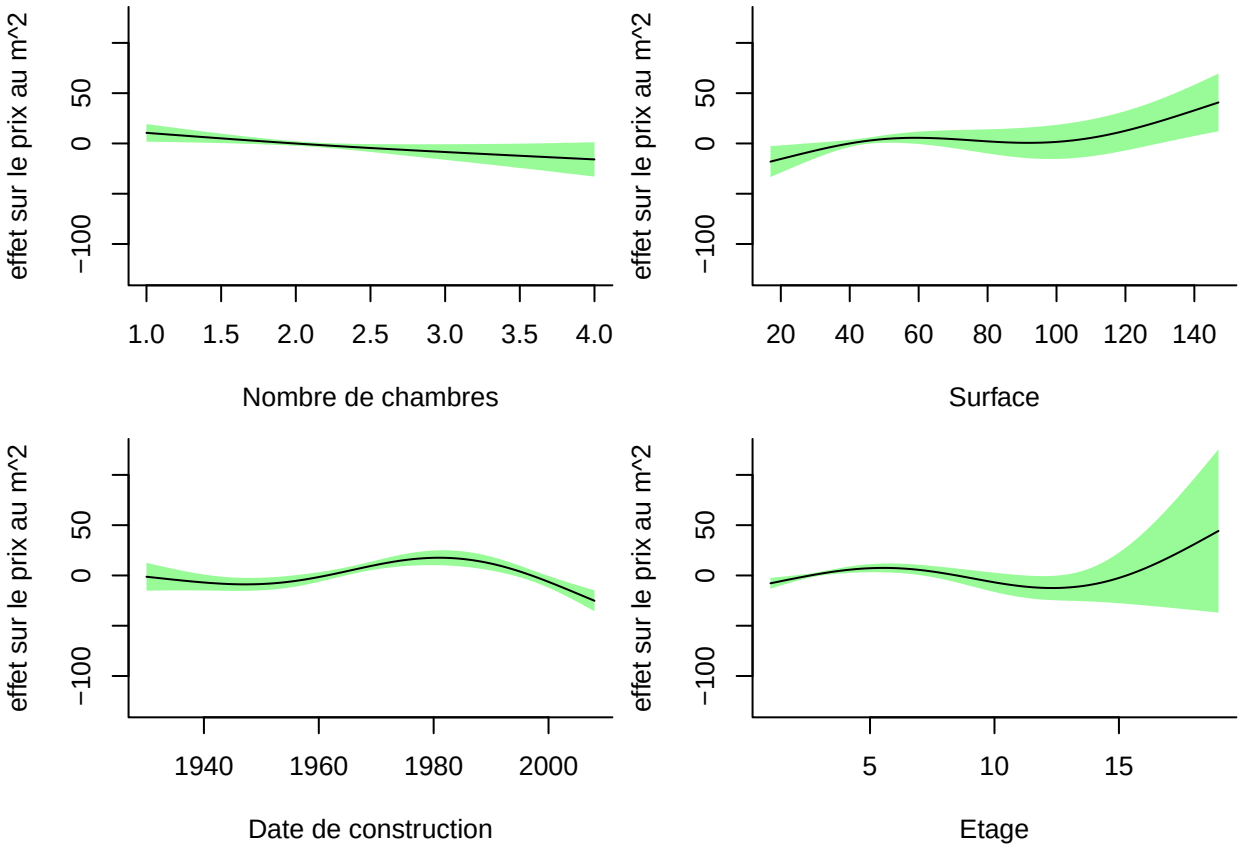
Ici, on a 16 graphes car on en fait 4 pour chaque modalité de la variable District.

- District="Mokotow" :



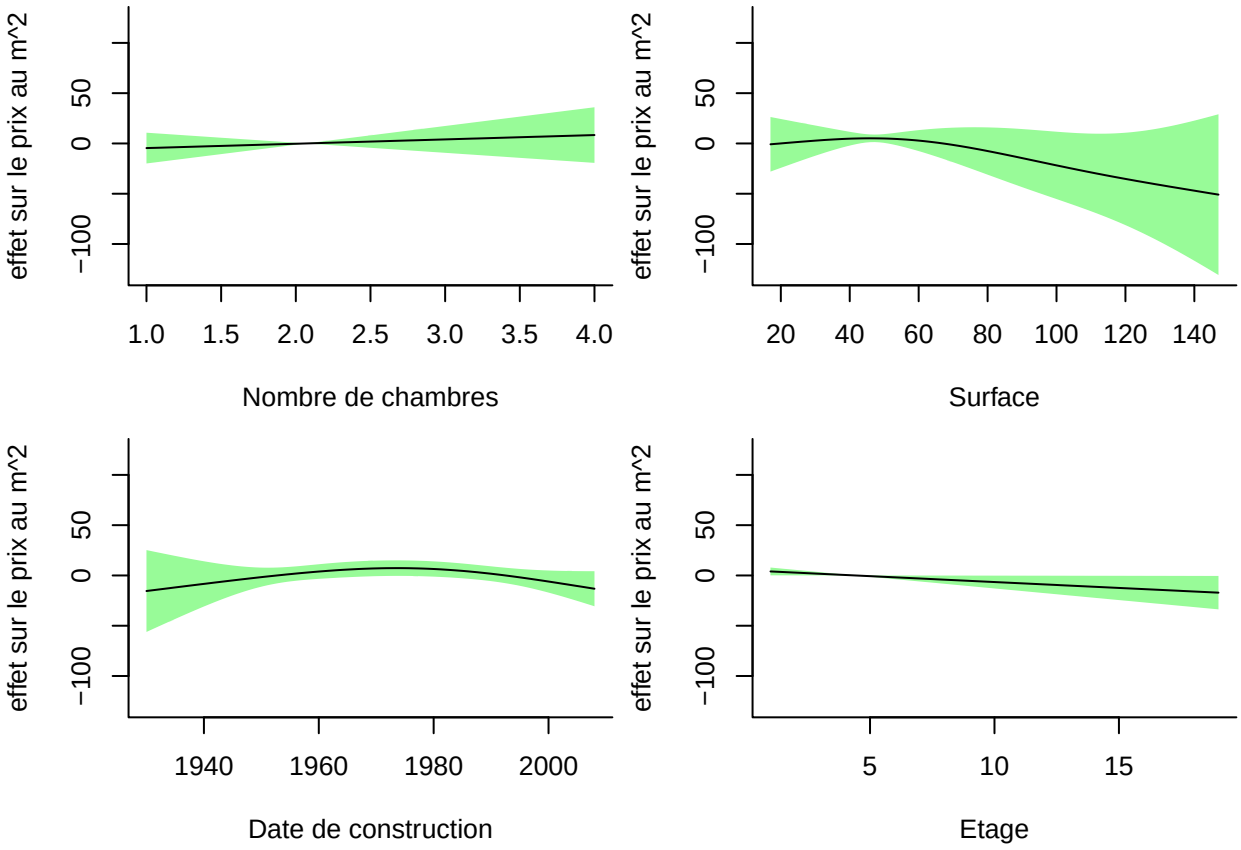
Pour ce district, on constate que l'a date de construction l'étage n'a presque aucun effet sur le prix du M2. La date de construction a quant-à elle un véritable impact.

- District="Srodmiescie" :



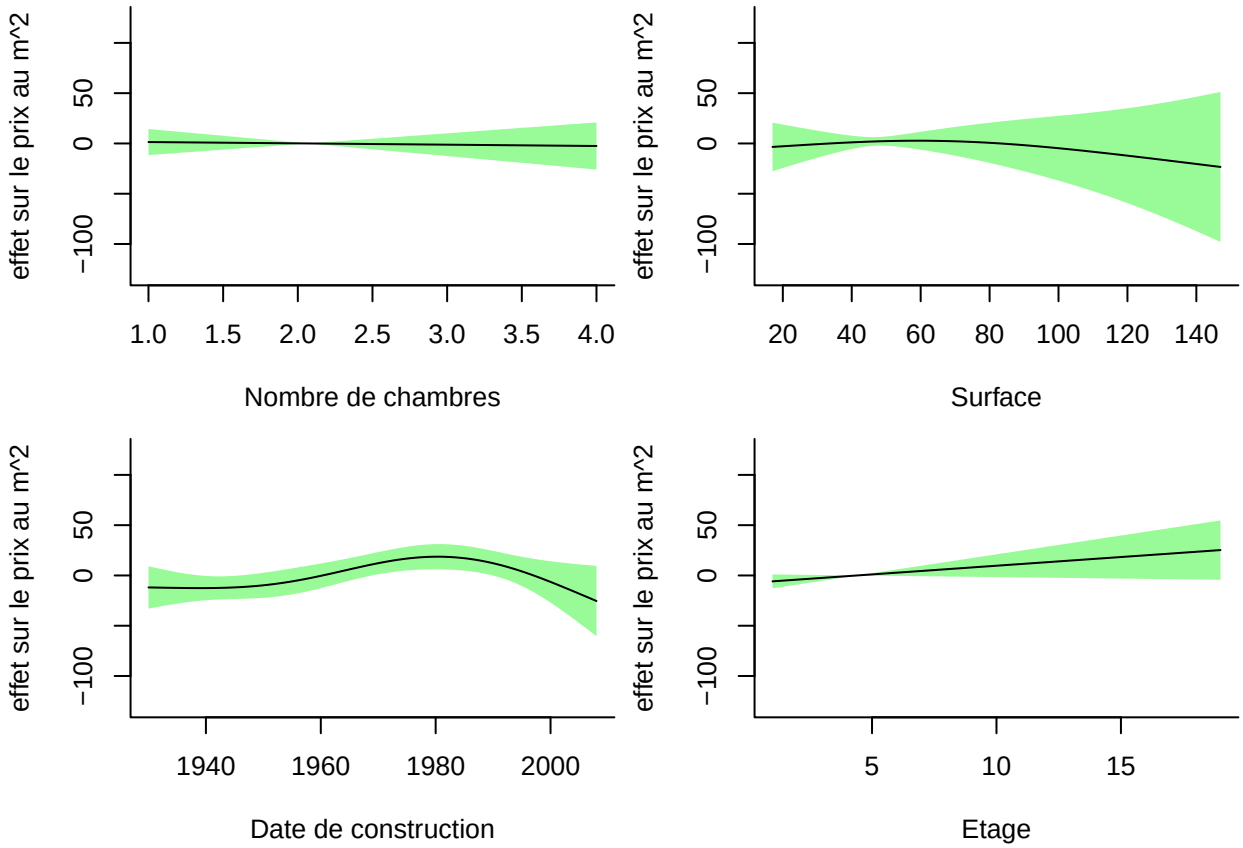
Ici, on remarque que l'a date de construction l'étage adopte une relation très spéciale, avec une variance qui augmente énormément quand l'appartemment est "haut".

- District="Wola" :



Pour ce district, la relation entre la surface et le prix du M2 adopte une grande variance

- District="Zoliborv" :



On voit sur ces graphes que le nombre de chambres a toujours une tendance presque constante : son effet est donc linéaire sur le prix au mètre carré. Les 3 autres variables, elles, ont des tendances variables : leur effet n'est donc pas linéaire sur le prix au mètre carré.

On peut maintenant faire un modèle semi-paramétrique. On distingue toujours les 2 cas.

- Cas ou la variable District a un effet linéaire :
- Cas ou la variable District a un effet non-linéaire :

On compare les deux modèles emboités via une analyse de variance (un test de Fisher) :

```
## Analysis of Deviance Table
##
## Model 1: areaPerMzloty ~ district + n.rooms + s(surface, k = 4) + s(construction.date,
##     k = 4) + s(floor, k = 4)
## Model 2: areaPerMzloty ~ district + n.rooms + s(surface, k = 4, by = district) +
##     s(construction.date, k = 4, by = district) + s(floor, k = 4,
##     by = district)
##   Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
## 1    395.80     115097
## 2    376.92     102251 18.882    12846 2.5193 0.0005008 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette le modèle où on considère l'effet de la variable District comme linéaire.

On compare le meilleur modèle semi-paramétrique au meilleur modèle non-paramétrique trouvé au dessus.

```
## Analysis of Deviance Table
##
## Model 1: areaPerMzloty ~ district + s(n.rooms, k = 4, by = district) +
##     s(surface, k = 4, by = district) + s(construction.date, k = 4,
##     by = district) + s(floor, k = 4, by = district)
## Model 2: areaPerMzloty ~ district + n.rooms + s(surface, k = 4, by = district) +
##     s(construction.date, k = 4, by = district) + s(floor, k = 4,
##     by = district)
##   Resid. Df Resid. Dev      Df Deviance      F  Pr(>F)
## 1    372.61      99669
## 2    376.92     102251 -4.3133  -2582.3 2.2517 0.05812 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On obtient une p-valeur <6% donc on rejette le modèle non-paramétrique au seuil de 6%. Le modèle semi-paramétrique est donc meilleur.

Enfin, on peut encore comparer les 2 modèles en calculant leur AIC : le meilleur modèle est celui qui a l'AIC le plus faible.

```
##              df      AIC
## resNP2 35.14089 3619.838
## resSP2 31.36119 3607.570
```

On retrouve à nouveau que le modèle semi-paramétrique est meilleur puisque son AIC est plus faible. Ici, la différence d'AIC n'est pas énorme mais elle est tout de même significative.

4 Conclusion

La première partie nous permet de conclure que la meilleure méthode pour prédire le prix au mètre carré d'un appartement de Varsovie avec la date de construction de celui-ci est la méthode des splines pénalisées et son MSE vaut 352.4423532.

La deuxième partie nous permet de conclure que c'est un modèle semi-paramétrique qui est le plus adapté à nos données pour prédire le prix au mètre carré avec la surface, le nombre de chambres, le district, l'étage et la date de construction.

Pour conclure, ce TP nous a permis d'appliquer les notions vues en cours et de voir quelles méthodes s'appropriaient le plus à notre jeu de données. Il est toujours très intéressant de voir comment se comportent les méthodes dans la pratique. On a plus pu construire un rapport ou on présente les méthodes de manière plus qualitative.