

# Projet : Méthodes de traitement de données manquantes

Olga Silva / Marlène Chevalier

30/11/2019

## Sujet : Valeur du logement en banlieue de Boston

Il s'agit de traiter les données manquantes du fichier Boston Housing. Ce fichier décrit la situation des logements dans les villes de la banlieue de Boston. Il est constitué de 506 enregistrements et de 14 variables quantitatives (soit 7084 données) :

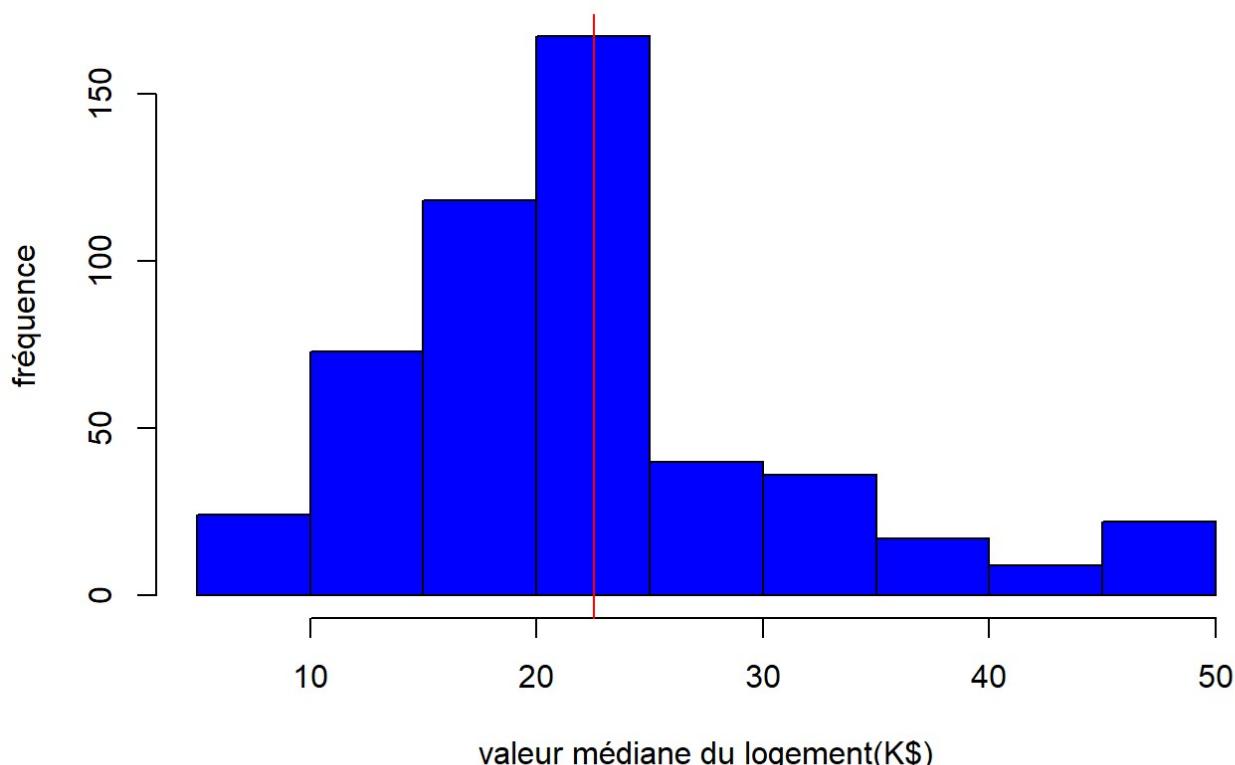
- **CRIM** : Taux de criminalité par habitant.
- **ZN** : Proportion de terrains résidentiels pour des lots de plus de 25000 pieds carré (environ 2300m<sup>2</sup>).
- **INDUS** : Proportion d'espace, en acres, consacré aux affaires non commerciales (1 acre environ 4000m<sup>2</sup>).
- **CHAS** : Proximité avec la rivière Charles (=1 si en bord de rivière / =0 si éloigné de la rivière)
- **NOX** : Concentration en oxyde d'azote (1 pour 10 millions)
- **RM** : Nombre moyen de chambres par logement
- **AGE** : Proportion des propriétés construites avant 1940
- **DIS** : Moyenne des distances aux 5 centres d'emploi de Boston
- **RAD** : Indice d'accessibilité aux autoroutes (de 1 à 8 et 24)
- **TAX** : Taux d'imposition foncier (1 pour 10 000\$)
- **PTRATIO** : Ratio d'élèves-enseignants
- **B** : Proportion de population afro-américaine
- **LSTAT** : Proportion de population précaire
- **MDEV** : Valeur médiane des habitations privées (en K\$)

Nous utiliserons ces données pour tenter d'expliquer la valeur médiane des habitations privées (MDEV) en fonction des autres variables du fichier.

## 1.Exploration des données incomplètes

### Graphiques sur les données incomplètes

La valeur médiane du logement à Boston est comprise entre 5 K\$ et 50 K\$ , en moyenne de 22.5 K\$.



Graphiquement (cf. annexe 1.1), on observe que :

- Le **taux de criminalité faible** (inférieur à 10%) est le plus fréquent. La valeur du logement a tendance à diminué lorsque le taux de criminalité augmente. Mais la corrélation entre les 2 reste faible (0.4).
- La **proportion de terrains résidentiels** est majoritairement faible (inférieur à 10%), mais lorsqu'elle augmente, la valeur du logement a tendance à augmenté.
- La **proportion de surface d'activité industrielle** la plus fréquente est entre 18 et 20 acres (entre 72000m<sup>2</sup> et 80000m<sup>2</sup>). A ce niveau, la valeur moyenne est bien souvent inférieure à la valeur médiane moyenne des logements (22.5K\$).
- La \*\*concentration d'oxyde d'azote\*\* est le plus souvent entre 0.4 et 0.6. Plus la concentration augmente, plus la valeur des logements diminue.
- Le **nombre moyen de chambre** est le plus souvent entre 5 et 7. Plus le nombre de chambre augmente, plus les logements ont de la valeur.
- La **proportion de propriétés construites avant 1940** est très importante (majoritairement autour de 90%). Plus cette proportion augmente, plus les logements ont de la valeur.
- La **moyenne des distances aux centres d'emploi** est fréquemment faible (<4). Cette variable influence peu la valeur des logements (corrélation=0.28).
- L'**imposition foncière** prend la plus importante partie de ses valeurs entre 200 et 500. Puis une autre série importante de ses valeurs est autour de 666 ; à ce niveau d'impôt, la valeur du logement est plus faible (<moyenne 22.5).
- Le **ratio élèves-enseignants** est réparti quasi-équitablement autour de la valeur moyenne du logement. Une hausse de ce ratio a tendance à faire baisser le prix du logement (corrélation =-0.54)
- La **proportion de population afro-américaine** est importante; mais son influence n'est pas significative sur la valeur du logement (cor =0.35)
- La **proportion de population précaire** influence négativement la valeur des logements (cor =-0.74).
- L'**indice d'accessibilité aux autoroutes** à 24 donne les valeurs des logements les plus basses (15K\$ en moyenne) . Les indices 3 et 8 donnent les valeurs de logement les plus élevées.
- La **proximité avec la rivière Charles** augmente légèrement la valeur du logement.

## Corrélation des variables

Les corrélations les plus significatives de la valeur du logement sont avec :

- RM (0.72) : plus le nombre de chambre est important, plus la valeur du biens est forte.
- INDUS et RAD (-0.51) : plus l'espace d'affaires non commerciales ou plus l'indice d'accessibilité aux autoroutes seront importants, moins la valeur du bien sera élevée.
- PTRATIO (-0.54) : plus le ratio élèves-enseignants est fort, moins la valeur des biens est élevée.
- LSTAT (-0.74) : une forte proportion de population précaire réduira très fortement la valeur des biens.  
(cf. annexe 1.2)

## Modélisation sur les données incomplètes

Nous commençons par examiner les résultats d'une regression linéaire de MEDV sur les autres variables. Le modèle est correctement ajusté ( $R^2=0.75$ ) mais il semble que les variables explicatives INDUS et AGE ne soient pas pertinentes pour ce modèle. (cf. annexe 1.3)

Nous allons appliquer une méthode "step AIC" pour choisir le meilleur modèle avec le risque de perte d'information et d'introduction de biais au dataset.

Ce modèle écarte les variables INDUS et AGE avec un R2 ajusté très proche de l'original. (cf. annexe 1.4)

Cependant, en regardant le graphique des résidus (cf. annexe 1.5) nous observons que le modèle n'est pas optimal, il semblerait qu'il y a une relation non linéaire avec une ou plusieurs variables. Pour la trouver, nous faisons un graphique par pairs (cf. annexe 1.6) et observons que la variable LSTAT n'a pas une relation polynomiale d'ordre 2. Si nous corigeons notre modèle en ajoutant un ordre quadratique, les résidus s'améliorent nettement. (cf. annexe 1.7)

## En conclusion sur l'exploration de données

La valeur médiane du logement à Boston et ses environs est comprise entre 5K\$ et 50K. *Sa distribution est croissante jusqu'à sa valeur moyenne (22.5K) puis décroît fortement à partir de 25 K\$.*

**La valeur du logement est influencée (p\_value <5%):**

- **positivement** principalement (coefficients estimés = 3.58) **par le nombre de chambres** et plus faiblement (coefficients estimés de 0.25 et 0.03) **par l'accessibilité aux autoroutes et la proportion de terrain résidentiel**

- **négativement** principalement (coefficients estimés = -14.71) **par la concentration en oxyde d'azote** et plus faiblement (coefficients estimés entre de -1.34, -0.76, -1.45, -0.01) **par la distance aux centres d'emplois, le ratio élèves-enseignants, la proportion de population précaire et le taux foncier**

Le R<sup>2</sup> ajusté du modèle de regression sur jeu de données incomplet est de **0.7965 (R<sup>2</sup> de référence)**.  
(cf. annexe 1.7 : résultat de la regression linéaire après sélection de variable)

Les variables INDUS et AGE ne sont pas significatives dans l'explication de la valeur du logement (p-value>5% cf. annexe 1.3), et LSTAT a une relation quadratique avec MEDV.

Nous allons comparer maintenant ce modèle obtenu avec des données manquantes et les nouveaux modèles que nous allons obtenir avec des datasets complets à partir de différentes méthodes.

## 2. Inventaire des données manquantes

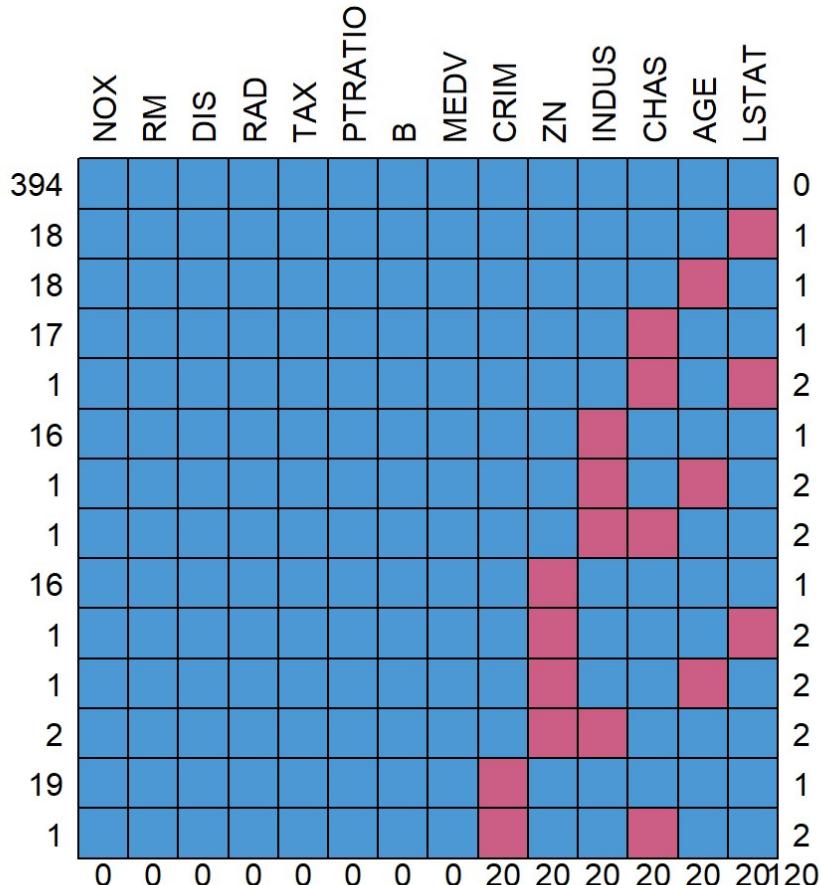
Il s'agit ici d'identifier les données manquantes par variable, représenter leur structure dans le jeu de données.

### Structure des données manquantes

La fonction **md.pattern** du package MICE a pour résultat une matrice, dans laquelle chaque ligne correspond à des structures de données manquantes et chaque colonne à une variable du fichier. Les lignes et les colonnes sont triées selon le niveau de complétude des données.

A chaque ligne de la matrice (qui définit une structure de données manquantes du jeu de données) :

- la première colonne indique le nombre d'observations correspondant à la structure de données manquantes décrite ;
- la dernière colonne donne le nombre de variables incomplètes.



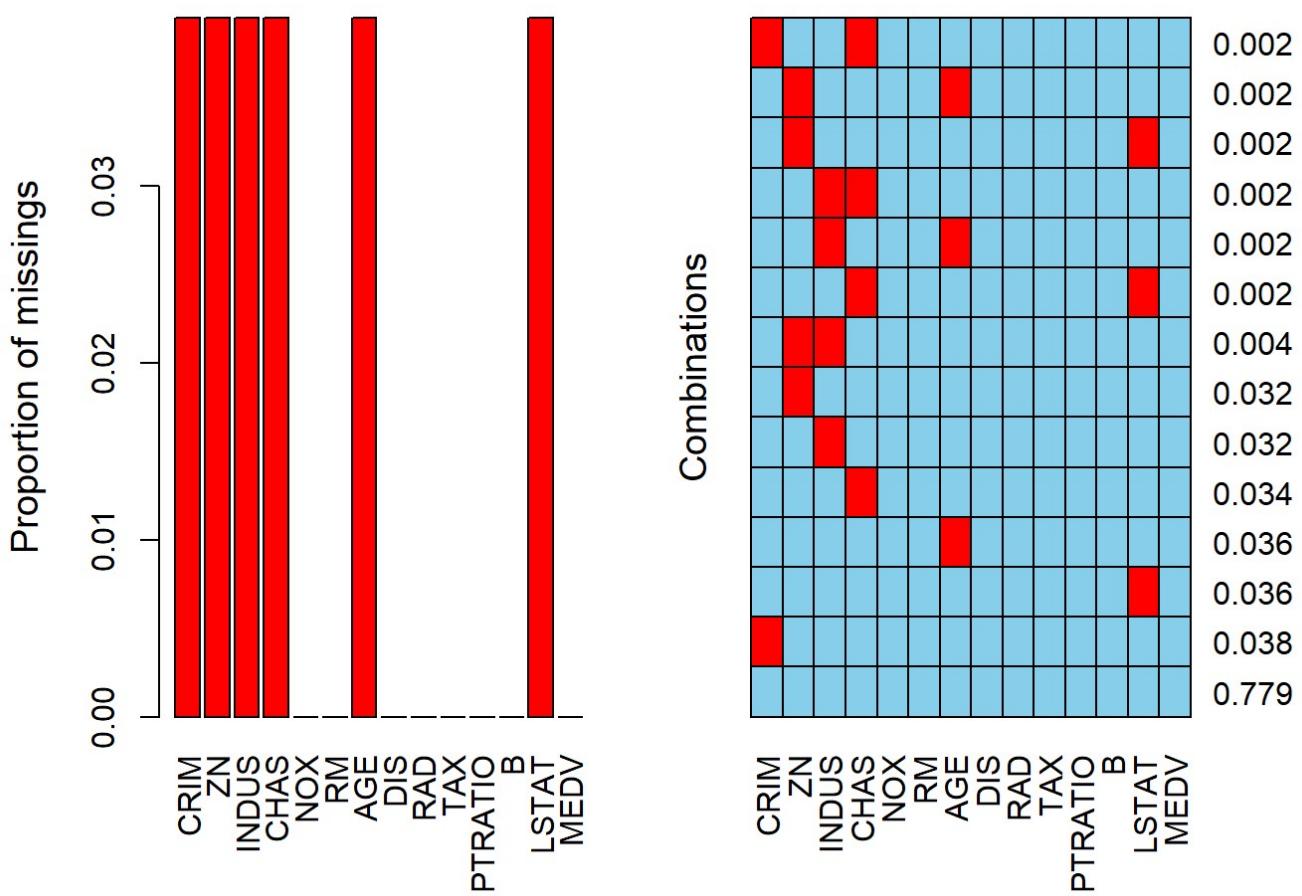
Au total, 120 observations sont manquantes : 20 pour chacune des variables CRIM, ZN, INDUS, CHAS, AGE, LSTAT.

- structure 1 : 394 observations pour lesquelles aucune donnée n'est manquante,
- structure 2 : 18 observations pour lesquelles seule la donnée de la variable LSTAT est absente,
- structure 3 : 18 observations pour lesquelles seule la donnée de la variable AGE est absente,
- ...
- structure 14 (dernière) : 1 observation pour laquelle les données des variables CRIM et CHAS sont absentes.

Une deuxième représentation des données manquantes obtenue avec **matrixplot** peut être observé en **annexe 2.1**. Elle permet d'identifier des dépendances des valeurs extrêmes et des données manquantes. Les données observées sont en gris et les manquantes en rouge. Nous pouvons observer que les valeurs manquantes se répartissent bien dans l'ensemble du jeu de données.

## Proportion de données manquantes

La fonction **aggr** permet d'appréhender les données par leur proportion dans le jeu complet.



En sortie, 2 graphiques :

- Le graphique de gauche donne la proportion de données manquantes de chaque variable : ici on retrouve des proportions égales pour les variables CRIM, ZN, INDUS, CHAS, LSTAT (autour de 4%), les autres variables sont complètes.
- Le graphique de droite donne la proportion de chaque structure de données.

Ici 78% du jeu de données est complet pour toutes les variables, 3.8% des individus ont uniquement la variable CRIM qui n'est pas renseignée, ...

## Catégories de données manquantes

Rubin (1976) a classé les problèmes des données manquantes en trois catégories :

- **Données manquantes de façon complètement aléatoire : MCAR** (missing completely at random). L'absence de données est dûe au hasard, à la malchance. Cette hypothèse est peu réaliste.
- **Données manquantes de façon aléatoire : MAR** (missing at random). La probabilité d'absence de la valeur d'une variable dépend des valeurs prises par d'autres variables qui ont été observées. MAR est plus générale et plus réaliste que MCAR.
- **Données manquantes de façon non aléatoire : MNAR** (missing not at random). La cause d'absence de la valeur d'une variable est de raison inconnue. MNAR est le cas le plus complexe.

La plupart des méthodes modernes de traitement des données manquantes partent de la supposition MAR. Dans le cas du jeu de données Boston Housing, on peut aussi partir de cette hypothèse.

## En conclusion sur l'inventaire des données manquantes

- 6 variables sont incomplètes, avec chacune 20 données manquantes (soit 120 au total) :
  - **CRIM** : Taux de criminalité par habitant
  - **ZN** : Proportion de terrains résidentiels
  - **INDUS** : Proportion d'espace consacré aux affaires non commerciales
  - **CHAS** : Proximité avec la rivière Charles
  - **AGE** : Proportion des propriétés construites avant 1940

- **LSTAT** : Proportion de population précaire

Sur le jeu de données, 22% des individus sont incomplets

- Nous supposons qu'on est en situation **MAR (Données manquantes de façon aléatoire)**

## 3.Traitement des données manquantes

### Imputation multiple

Les méthodes d'imputation simple consistent à remplacer chaque valeur manquante par une valeur unique prédite ou simulée.

Plusieurs solutions sont possibles (remplacer les données manquantes par la moyenne de la variable, faire une régression avec les données observées...). Ces solutions rapides sont à éviter car cela peut dégrader l'information (corrélation entre les variables, modification de la distribution des variables, variables sous-estimées ...). Nous allons utiliser uniquement l'imputation multiple pour notre projet.

L'imputation multiple va créer m datasets complets, au lieu d'imputer une seule fois comme l'imputation simple. Les méthodes d'imputation multiple suivent trois grandes étapes :

**Etape 1** : Imputation des données manquantes m fois

**Etape 2** : Analyse de m datasets imputés

**Etape 3** : Mise en commun des paramètres à travers m analyses

Plus de détail sur la méthodologie se trouve sur l'**annexe 3.1**

Nous allons ici tester plusieurs méthodes d'imputation multiple : l'imputation par regression stochastique, les forêts aléatoires, predictive mean matching et l'ACP. Nous allons appliquer ces méthodes à des modèles sans INDUS et AGE, car elles ne semblent pas pertinentes, selon les résultats des analyses précédentes.

### Imputation par régression stochastique

Cette méthode consiste à imputer les données manquantes en utilisant la regression à laquelle on a ajouté du bruit. Cela permet de corriger le biais de corrélation qui existe par les méthodes plus rapides d'imputation simple. Nous fixons m=6, en suivant les conseils de Stef Van Buuren dans son livre "flexible imputation data".

Pour faire cette imputation, nous utilisons la fonction **mice** ( avec **method = "norm.nob"**)

```

## 
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##      TAX + PTRATIO + B + LSTAT + I(LSTAT^2), data = dHBcompl.regsto)
## 
## Residuals:
##       Min     1Q Median     3Q    Max 
## -17.6161 -2.6113 -0.2792  1.9124 25.0548 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 42.160817   4.667547   9.033 < 2e-16 ***
## CRIM        -0.143846   0.029790  -4.829 1.84e-06 ***
## ZN          0.024394   0.012435   1.962 0.050357 .  
## CHAS        2.941981   0.774452   3.799 0.000164 *** 
## NOX         -13.994501  3.221078  -4.345 1.69e-05 *** 
## RM          3.218430   0.375564   8.570 < 2e-16 *** 
## DIS         -1.363581   0.169172  -8.060 5.80e-15 *** 
## RAD         0.269899   0.057885   4.663 4.02e-06 *** 
## TAX         -0.009235  0.003082  -2.996 0.002872 **  
## PTRATIO     -0.805602  0.119076  -6.765 3.78e-11 *** 
## B           0.008243   0.002441   3.378 0.000789 *** 
## LSTAT       -1.619560  0.115421 -14.032 < 2e-16 *** 
## I(LSTAT^2)  0.031862   0.003111  10.243 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.314 on 493 degrees of freedom
## Multiple R-squared:  0.7852, Adjusted R-squared:  0.7799 
## F-statistic: 150.2 on 12 and 493 DF,  p-value: < 2.2e-16

```

Avec cette méthode, on observe un R2 ajusté de **0.7806** après la régression linéaire faite avec le dataset completé. C'est une R2 ajusté proche de celui obtenu avec la régression faite avec les données manquantes.

La description de la nouvelle structure des données se trouve sur l'**annexe 3.2**. Sur cette description nous observons que pour deux variables (CRIM et LSTAT), l'imputation propose des valeurs négatifs en sachant que sur le dataset d'origine ces deux variables sont toujours positives.

### Imputation par forêts aléatoires

Dans l'imputation par forêts aléatoire, les valeurs sont imputées en faisant des tirages aléatoires à partir des distributions gaussiennes indépendantes, centrées en les moyennes prédictes par les forêts aléatoires. Pour ce faire , nous utilisons la fonction **mice ( avec method = "rf")**, avec m=6

```

## 
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##      TAX + PTRATIO + B + LSTAT + I(LSTAT^2), data = dHBcompl.rf)
##
## Residuals:
##       Min     1Q Median     3Q    Max
## -17.4885 -2.5991 -0.2677  1.9179 25.1679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.534622   4.643549   9.160 < 2e-16 ***
## CRIM        -0.144928   0.030052  -4.823 1.89e-06 ***
## ZN          0.021046   0.012057   1.746 0.081517 .
## CHAS        2.696889   0.763237   3.533 0.000449 ***
## NOX         -14.541976  3.215675  -4.522 7.67e-06 ***
## RM          3.301033   0.371172   8.894 < 2e-16 ***
## DIS         -1.363469   0.166253  -8.201 2.08e-15 ***
## RAD         0.269279   0.057831   4.656 4.15e-06 ***
## TAX         -0.009645   0.003077  -3.135 0.001822 **
## PTRATIO     -0.813213   0.117194  -6.939 1.25e-11 ***
## B           0.008215   0.002435   3.374 0.000800 ***
## LSTAT       -1.650979  0.118185 -13.969 < 2e-16 ***
## I(LSTAT^2)  0.033031   0.003180   10.388 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.309 on 493 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7805
## F-statistic: 150.7 on 12 and 493 DF,  p-value: < 2.2e-16

```

Avec cette méthode on observe un R<sup>2</sup> ajusté de **0.783** avec le dataset completé, proche de celui obtenu avec la méthode d'imputation stochastique. Les valeurs estimées et les résidus restent aussi très proches.

La description de la nouvelle structure des données se trouve sur l'**annexe 3.3**. Ici le nouveau dataset completé est cohérent avec celui d'origine.

### Imputation par predictive mean matching

La méthode pmm commence par faire une regression linéaire de variable pour estimer les paramètres de la distribution qui lie les variables avec données manquantes X et les variables complètes Y du dataset. A l'aide de cette distribution, on simule des prédictions de valeurs manquantes et présentes de la variable X. Ensuite on identifie pour chaque donnée manquante de X un ensemble de prédition sur les valeurs présentes de X, proches de la prédition sur la valeur manquante. Parmi cet ensemble de cas proches, on en choisit un au hasard et on attribue alors la valeur observée pour remplacer la valeur manquante.

Pour faire cette imputation, nous utilisons la fonction **mice** ( avec **method = "pmm"**), m=6

```

## 
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##      TAX + PTRATIO + B + LSTAT + I(LSTAT^2), data = dHBcompl.pmm)
##
## Residuals:
##       Min     1Q Median     3Q    Max 
## -17.2881 -2.6058 -0.2266  1.9320 24.9769 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 42.620768   4.627390   9.211 < 2e-16 ***
## CRIM        -0.124973   0.025799  -4.844 1.71e-06 ***
## ZN          0.021619   0.012362   1.749 0.080941 .  
## CHAS        2.515246   0.744798   3.377 0.000791 *** 
## NOX         -13.215203  3.227311  -4.095 4.94e-05 *** 
## RM          3.179371   0.373429   8.514 < 2e-16 *** 
## DIS         -1.364792   0.169861  -8.035 6.97e-15 *** 
## RAD         0.262876   0.057533   4.569 6.20e-06 *** 
## TAX         -0.009773   0.003074  -3.179 0.001569 **  
## PTRATIO     -0.791904   0.117977  -6.712 5.28e-11 *** 
## B           0.007816   0.002434   3.211 0.001407 **  
## LSTAT       -1.669251   0.118536 -14.082 < 2e-16 *** 
## I(LSTAT^2)  0.032705   0.003168  10.322 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.295 on 493 degrees of freedom
## Multiple R-squared:  0.7871, Adjusted R-squared:  0.7819 
## F-statistic: 151.8 on 12 and 493 DF,  p-value: < 2.2e-16

```

De même que pour les deux méthodes précédentes, le R2 ajusté est très proche du modèle avec des données manquantes (**0.7808**). Les valeurs estimées et les résidus restent aussi très proches.

La description de la nouvelle structure des données se trouve sur l'**annexe 3.4** et on observe que les valeurs minimum, maximum et médiane sont presque identiques à celles obtenues avec les forêts aléatoires.

## Imputation à l'aide d'un traitement bayésien de l'ACP

Les données manquantes sont imputées en utilisant la ACP (Analyse des composantes principales).

Il s'agit d'abord d'estimer le nombre de composantes utilisés pour compléter les données avec la méthode de ACP : fonction **estim\_ncpPCA(dHB,method.cv = "Kfold")**

Selon le graphique (cf. **annexe 3.5**), le nombre de dimensions à retenir sera de 3.

- générer les ensembles de données imputées avec la fonction MIPCA en utilisant le nombre de dimensions précédemment calculé et la méthode bayésienne.  
fonction **MIPCA (avec method.mi = "Bayes")**

## Régression linéaire sur le jeu de données complété par MIPCA

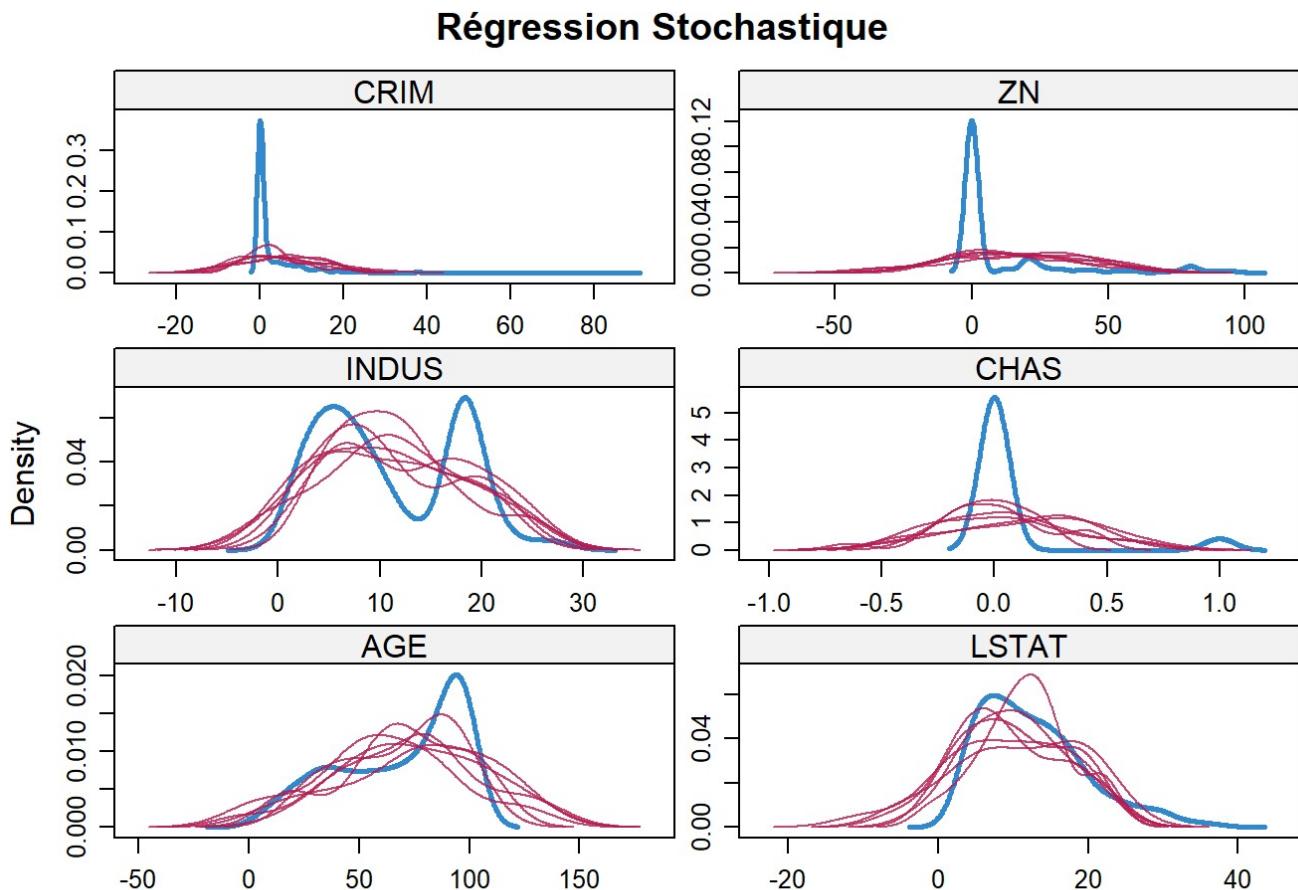
	estimate	std.error	statistic	df	p.value
## (Intercept)	35.951923596	5.155438880	6.973591	481.7192	1.023892e-11
## CRIM	-0.111829650	0.033347220	-3.353492	465.8459	8.630828e-04
## ZN	0.045202734	0.013834093	3.267488	461.3412	1.166036e-03
## CHAS	2.777197166	0.871486458	3.186736	471.0318	1.534501e-03
## NOX	-17.573579507	3.570054394	-4.922496	485.6421	1.172625e-06
## RM	3.846233820	0.415359407	9.260014	472.4972	0.000000e+00
## DIS	-1.472465198	0.186582154	-7.891779	479.0497	2.042810e-14
## RAD	0.293513628	0.064114477	4.577962	486.1031	5.970885e-06
## TAX	-0.011664463	0.003401937	-3.428771	487.6940	6.577037e-04
## PTRATIO	-0.942867225	0.131298453	-7.181099	485.4173	2.613909e-12
## B	0.009070683	0.002707828	3.349800	487.6963	8.713828e-04
## LSTAT	-0.508073031	0.048630938	-10.447527	453.7065	0.000000e+00

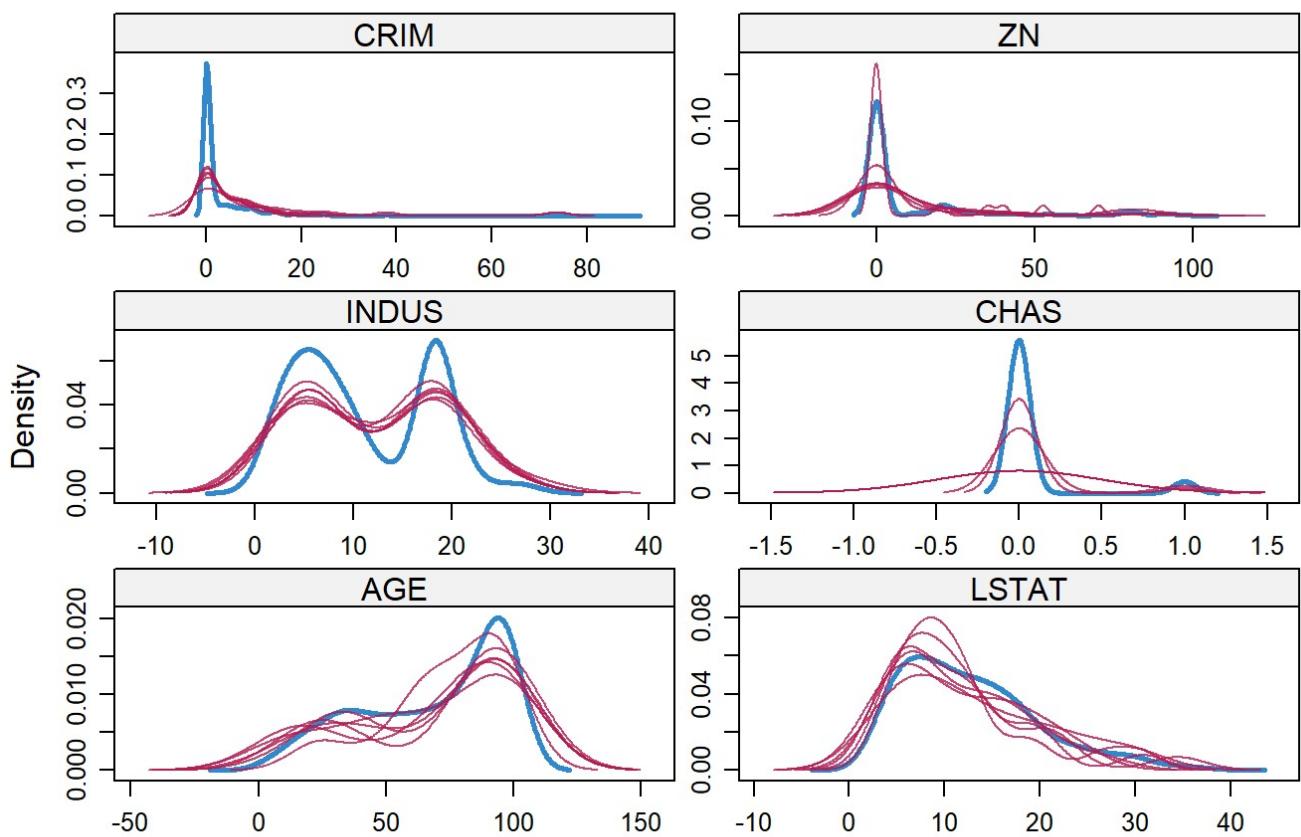
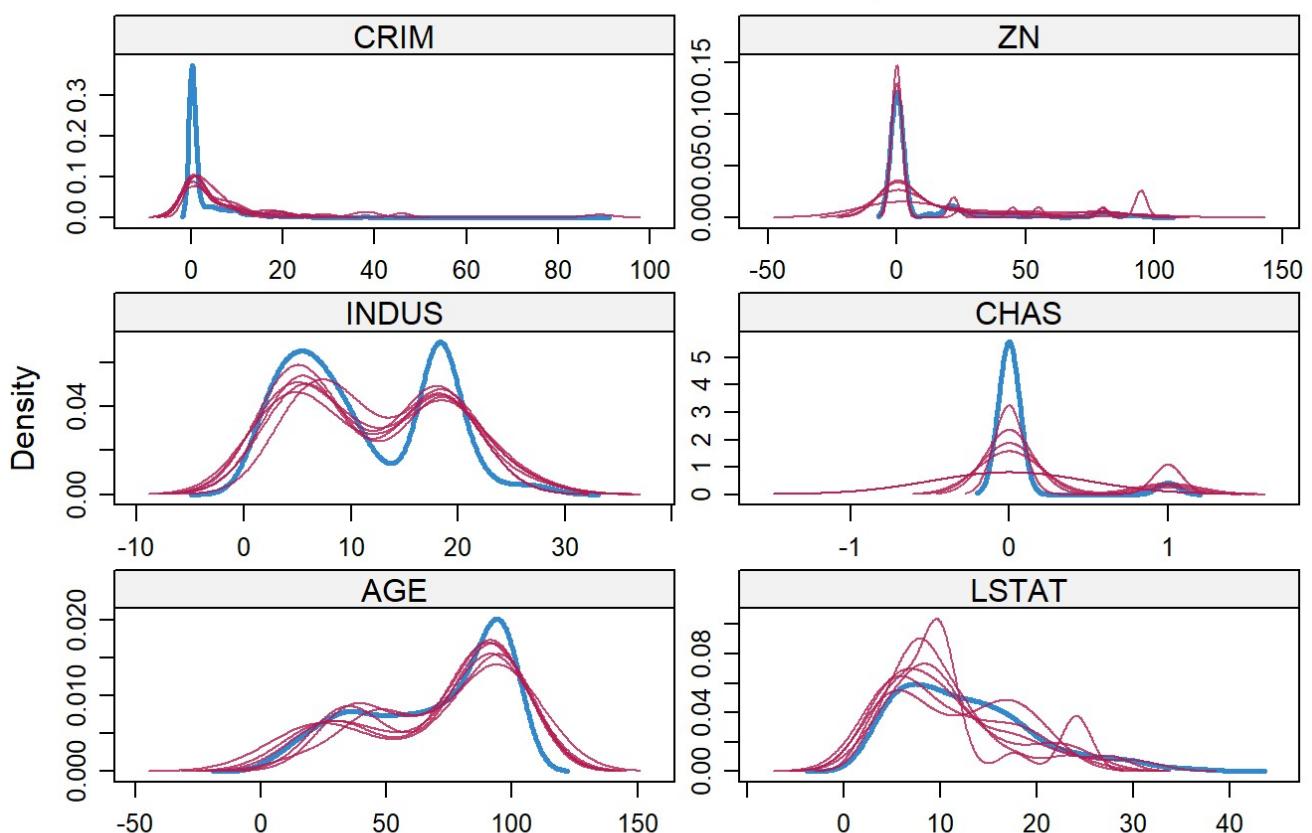
Les valeurs des estimateurs sont très proches de celles trouvées par les autres méthodes d'imputation multiple.

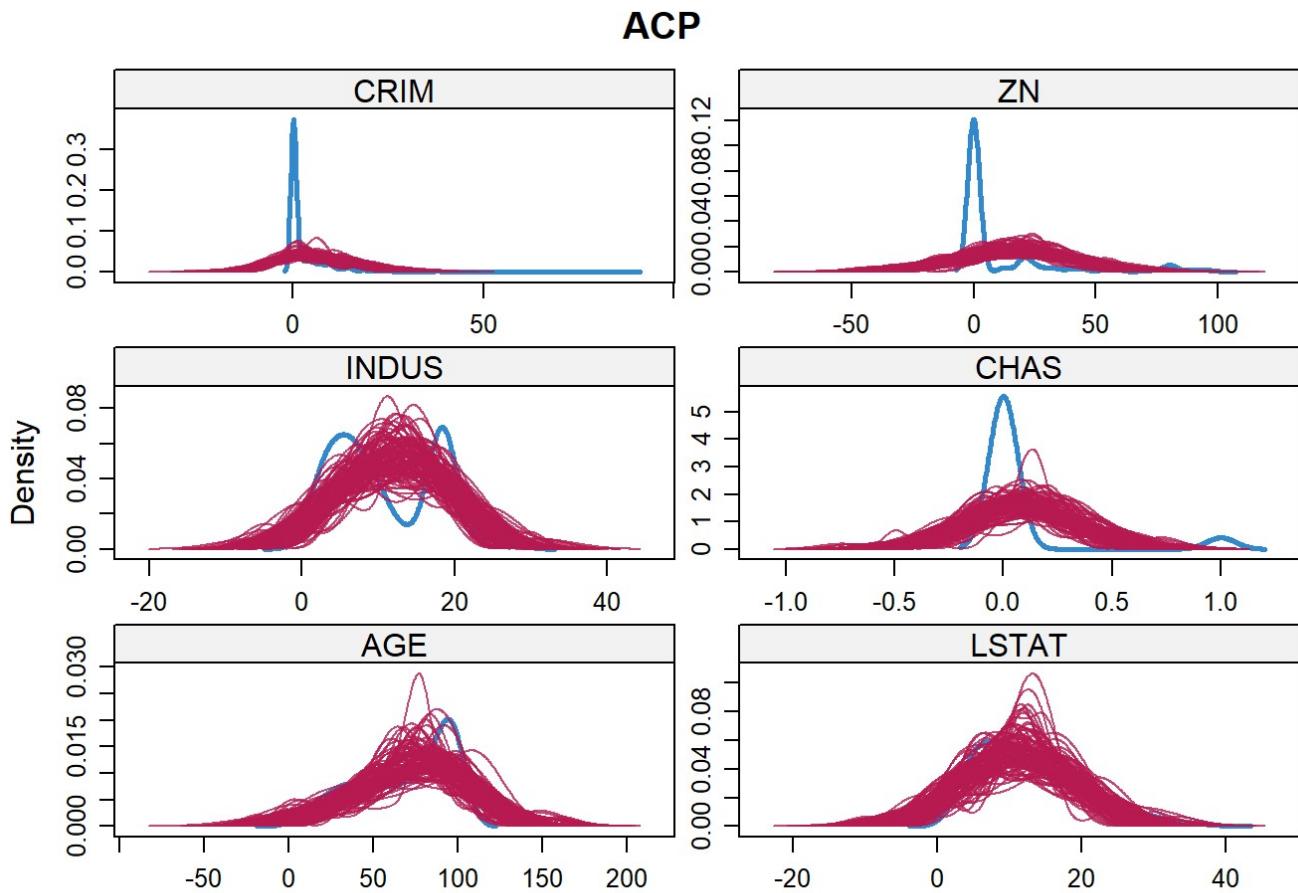
Nous avons testé 4 méthodes d'imputation, qui nous ont permis de faire des régressions avec le jeu de données complet. Ces régressions ont des résultats très proches en terme de R2 ajusté, des estimateurs et des résidus. Nous allons faire des diagnostics pour choisir le meilleur modèle d'imputation.

## 4. Diagnostics et conclusion

**Diagnostic 1 : vérifier que la distribution des données imputées est similaire à celle des données d'origine.**



**Forêts aléatoires****Predictive mean matching**



Pour la variable CRIM, aucune méthode ne semble parvenir à reproduire la même distribution. Par contre pour INDUS, AGE et LSTAT, toutes les méthodes semblent réussir à le faire.

Pour CHAS et ZN, les forêts aléatoires sont les plus proches de la vraie distribution.

Les méthodes de régression stochastique et d'ACP imputent des valeurs plus basses que celles observées dans la distribution originale des variables CRIM, ZN et AGE. Par contre, les imputations par les méthodes pmm et les forêts aléatoires respectent mieux la distribution originale des variables.

Pour la variable CRIM, les valeurs observées plus grandes se trouvent uniquement sur l'imputation par forêt aléatoire.

Pour la variable AGE, la régression stochastique et ACP imputent des valeurs plus grandes (proches de 200) que celles observées (autour de 100).

## Diagnostic 2 : vérifier la convergence des algorithmes.

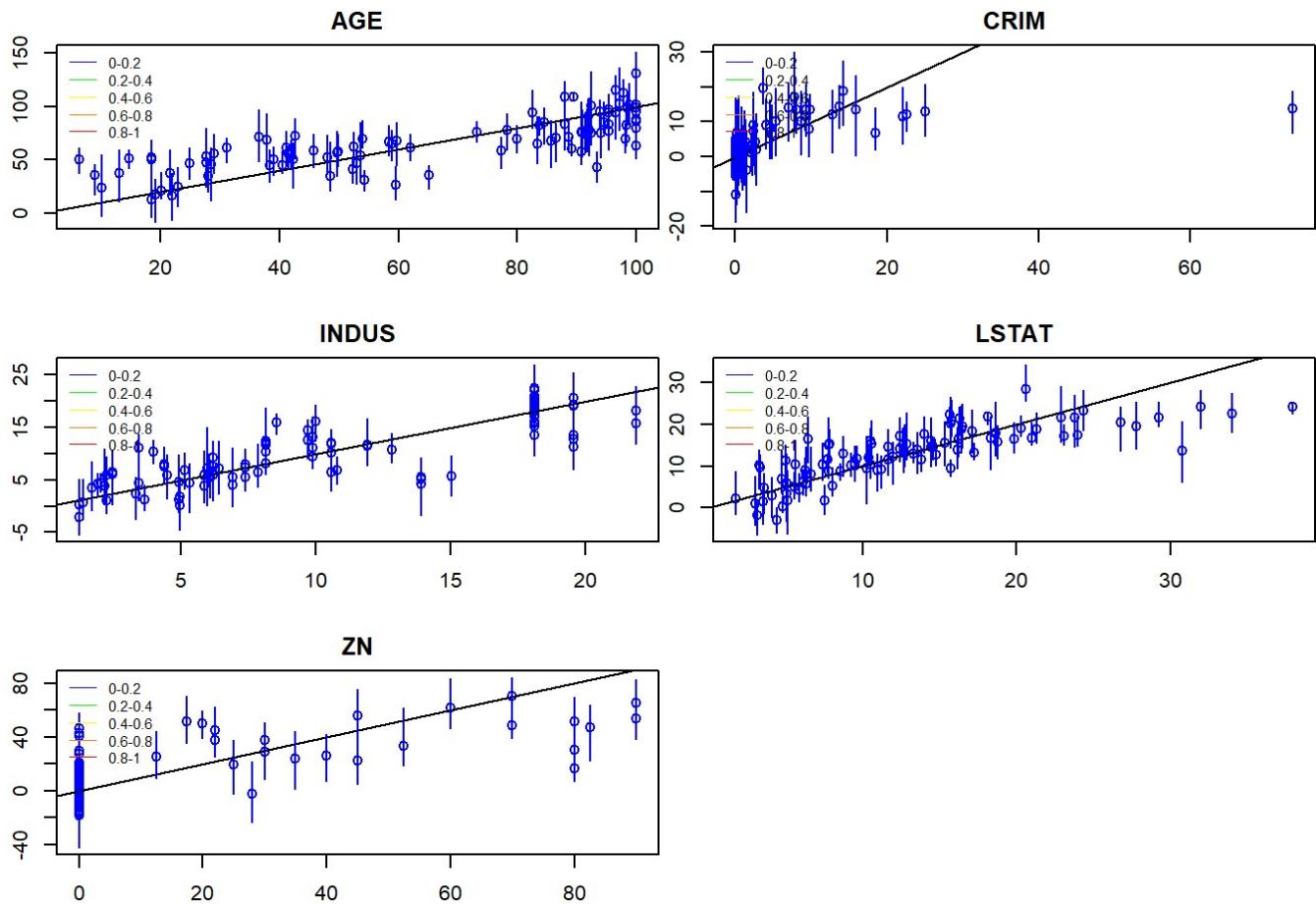
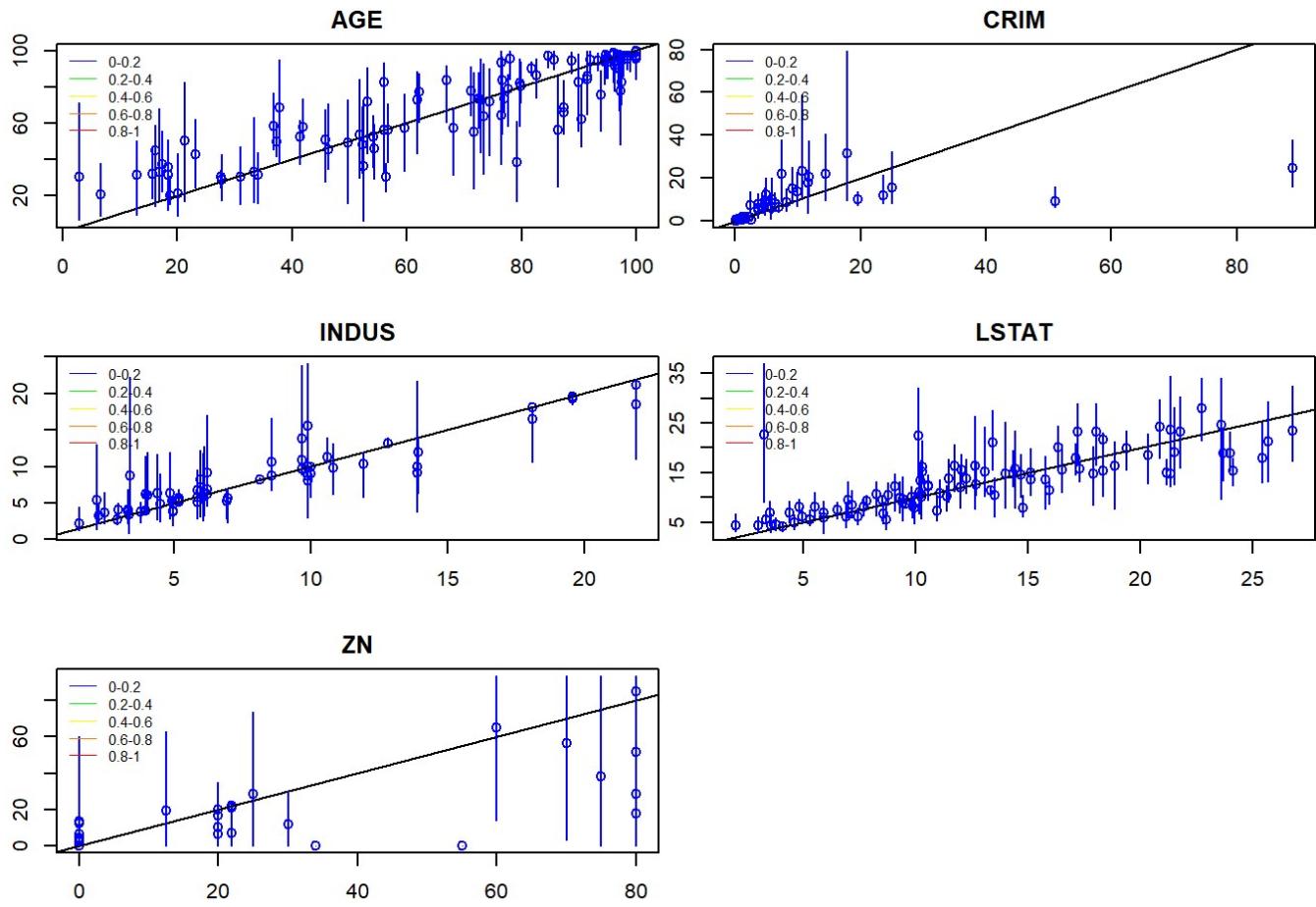
La vérification de la convergence se fait à partir des graphiques de variation de la moyenne et de l'écart type, pour chaque méthode, pour chaque itération et chaque donnée imputée. Pour que la convergence soit vérifiée, il faut que les différentes courbes se mélangent, sans une tendance particulière. C'est bien le cas pour nos graphiques; donc, nous n'avons pas de problème de convergence. (**cf.annexe 4.1**)

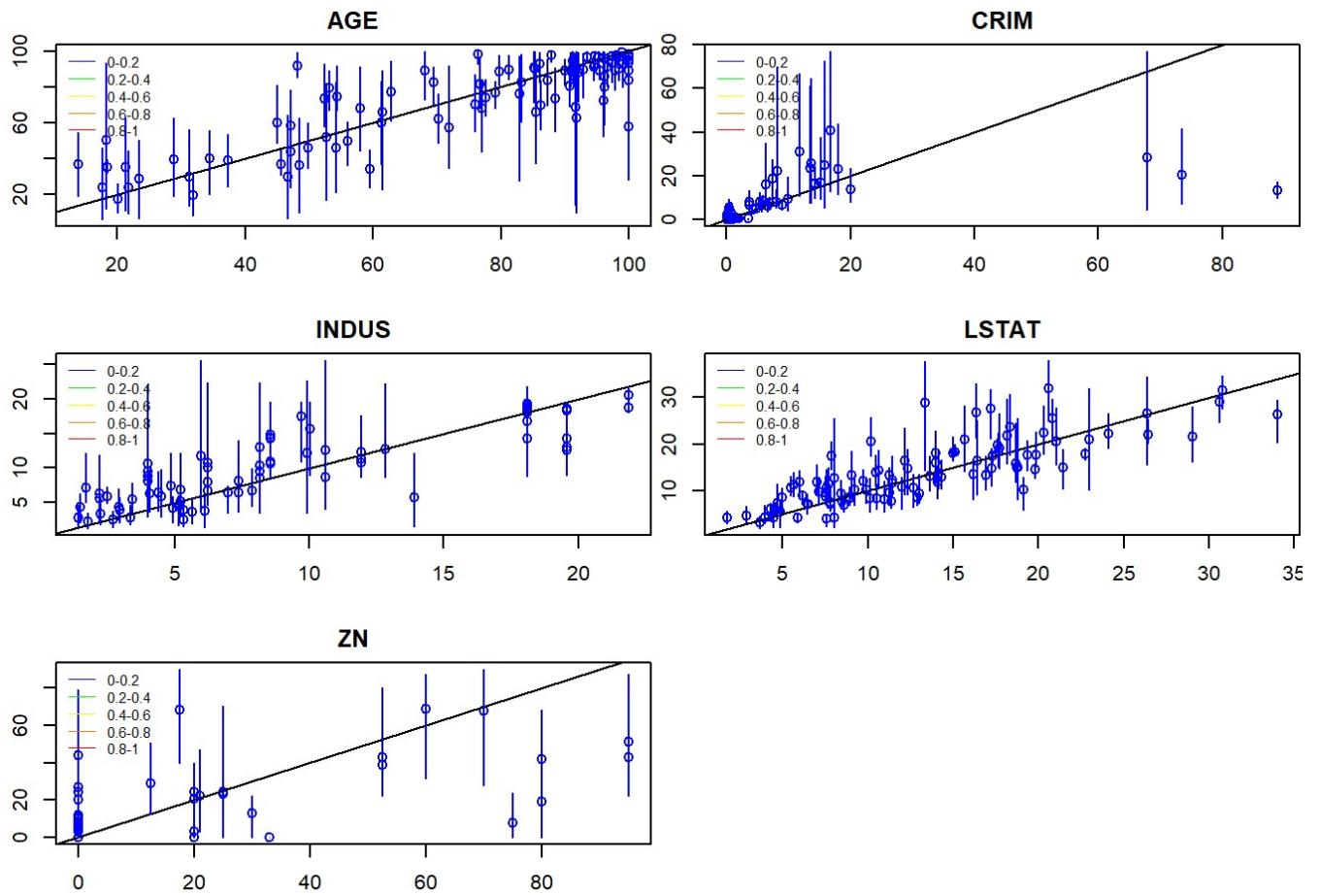
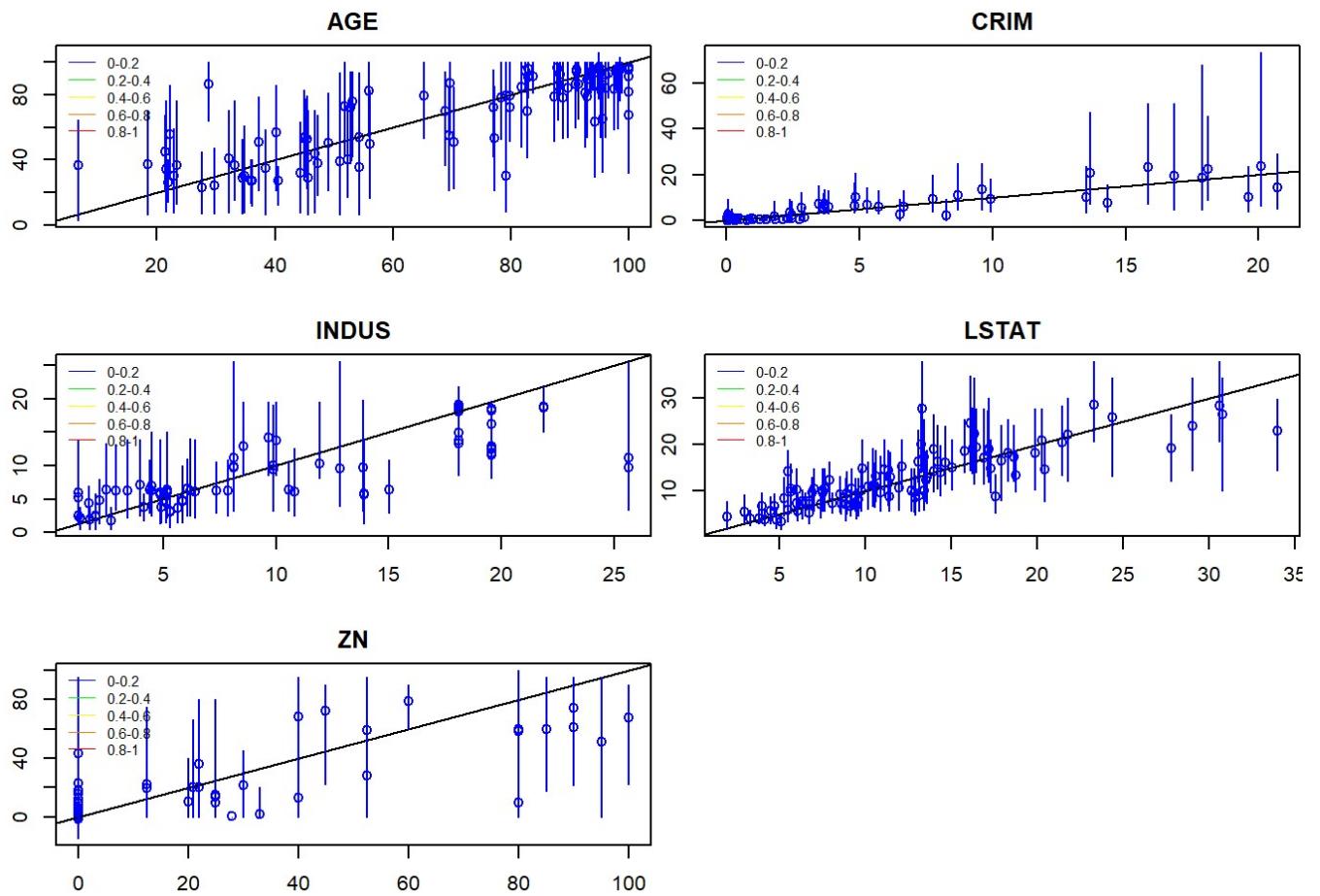
## Diagnostic 3 : vérifier l'ajustement du modèle d'imputation

Pour cela, nous traçons le graphe d'overimputation. Chaque donnée observée est supprimée et pour chacune d'entre elles, 100 valeurs sont prédites (en utilisant la même méthode d'imputation choisie); la moyenne et des intervalles de confiance de 90% sont calculés pour ces valeurs prédites.

Sur ces graphiques, la 1ère bissectrice ( $y=x$ ) représente l'imputation parfaite. La qualité de l'imputation se mesure en observant la proximité des intervalles de confiance avec cette droite. On espère que 90% des intervalles traversent la 1ère bissectrice. La couleur des intervalles représente la fraction de données manquantes (entre 0 - 20% pour notre cas).

**Ajustement de la regression stochastique :**

**Ajustement des forêts aléatoires :****Ajustement de la predictive mean matching :**

**Ajustement de l'ACP :**

Comme CHAS n'est pas une variable continue, elle prend uniquement les valeurs 0 et 1, elle n'apparaît pas dans les

graphiques.

On observe que pour les variables INDUS, AGE et LSTAT la plupart des intervalles de confiance contiennent bien la ligne diagonale pour chacune des méthodes. Cependant, ce n'est pas le cas pour les variables CRIM et ZN.

## Conclusion

D'après les trois diagnostics réalisés, les 4 méthodes d'imputation présentées:

- Respectent bien la distribution des variables, même si la régression stochastique et l'ACP ne le font pas si bien que les autres.
- Ne présentent pas de problème de convergence
- Ne permettent pas de bien reconstruire les variables CRIM et ZN, qui ne sont pas toujours bien représentées (cf. overimputation).

Par contre, si on compare les méthodes d'imputation en terme de rapidité machine, la méthode APC est celle qui demande plus de temps, même sur un dataset plutôt petit.

Les résidus sont très proches quelques soient les modèles (cf.[annexe 4.2](#)).

Donc pour notre petit dataset, nous aurions obtenu les mêmes résultats pour la régression linéaire sans l'imputation.

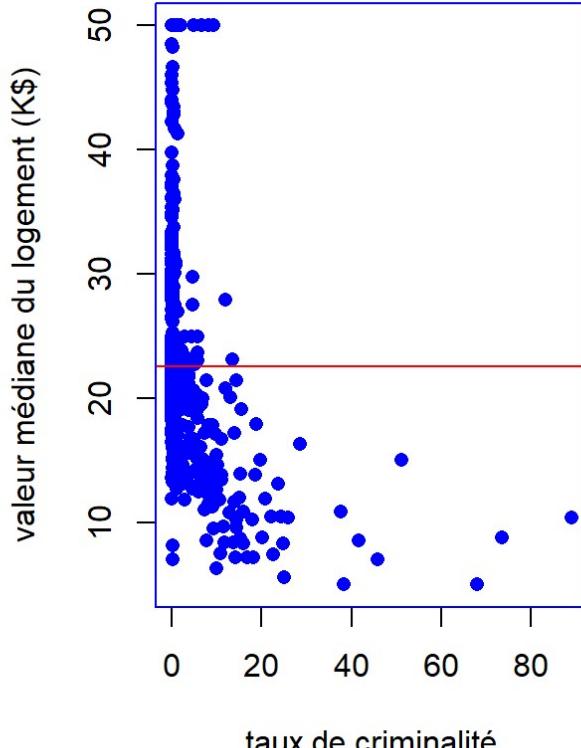
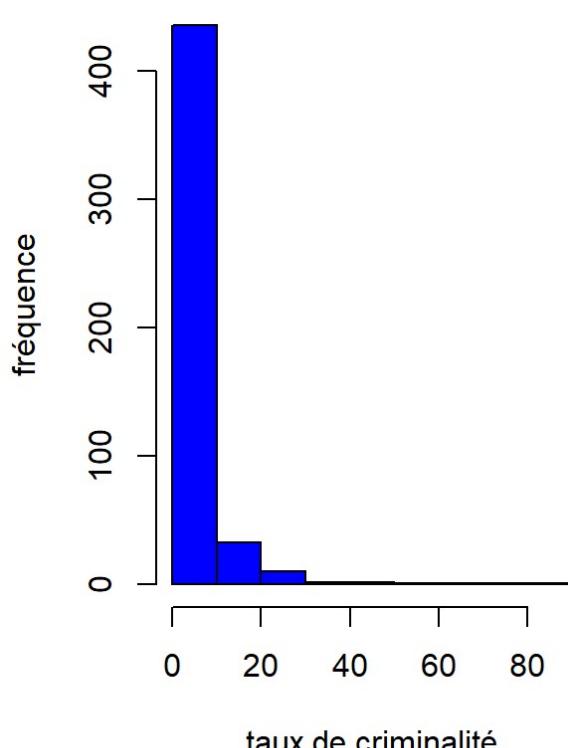
---

## Annexes

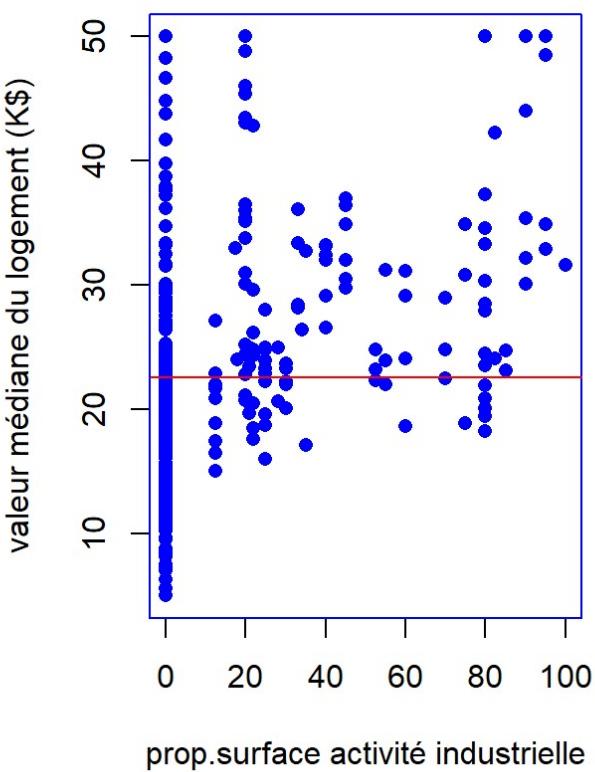
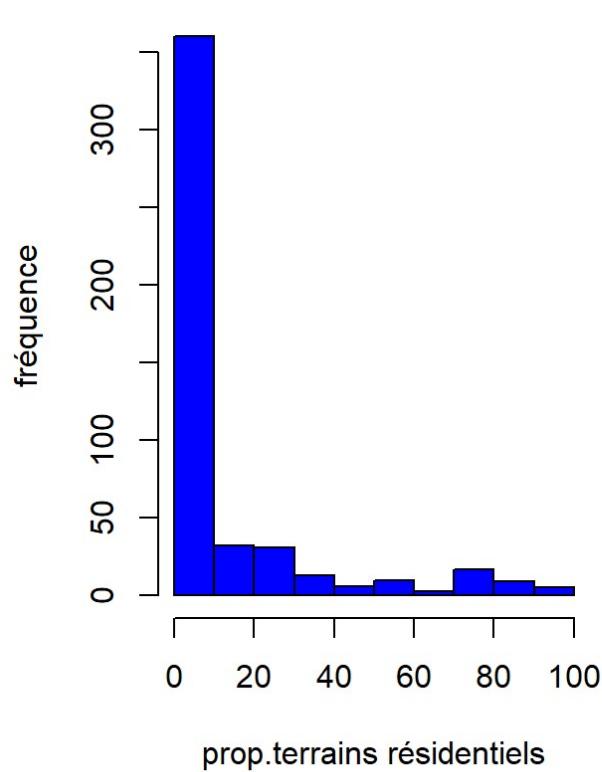
### Annexe 1.1 : graphiques exploratoires du jeu de données

Les représentations graphiques donnent, pour chaque variable explicative, la distribution sous forme d'histogramme et le nuage de point de la variable vs la valeur médiane du logement (en rouge la valeur moyenne = 22.5K\$).

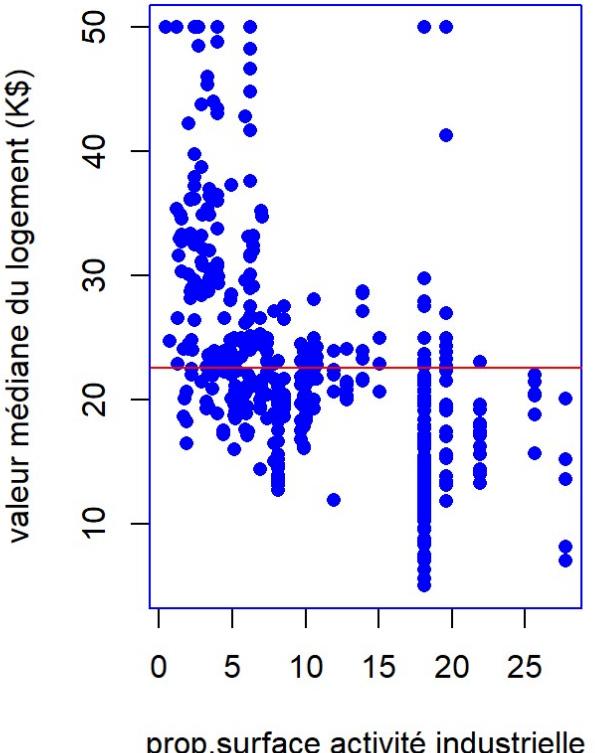
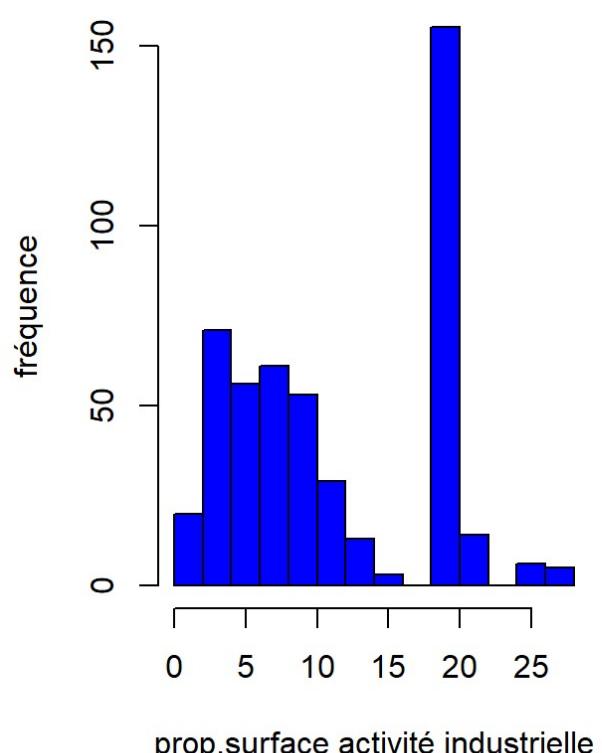
#### Valeur médiane du logement et taux de criminalité



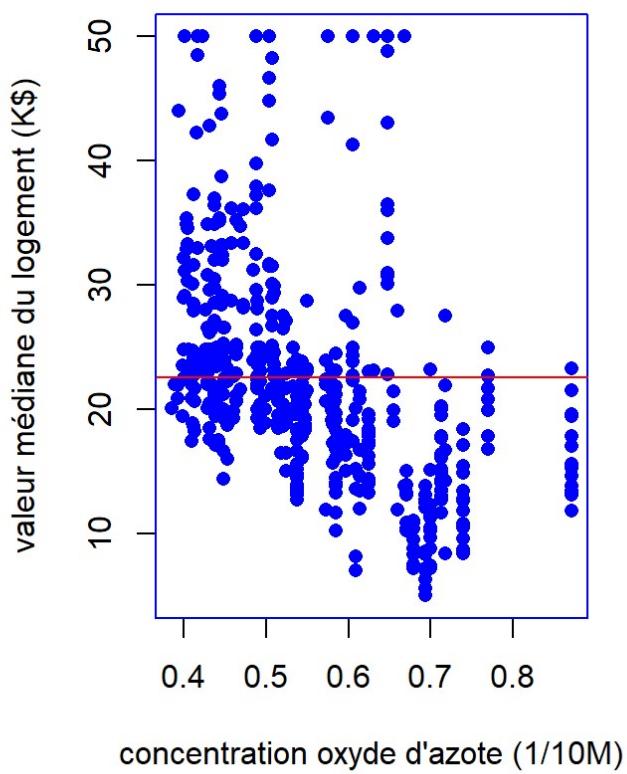
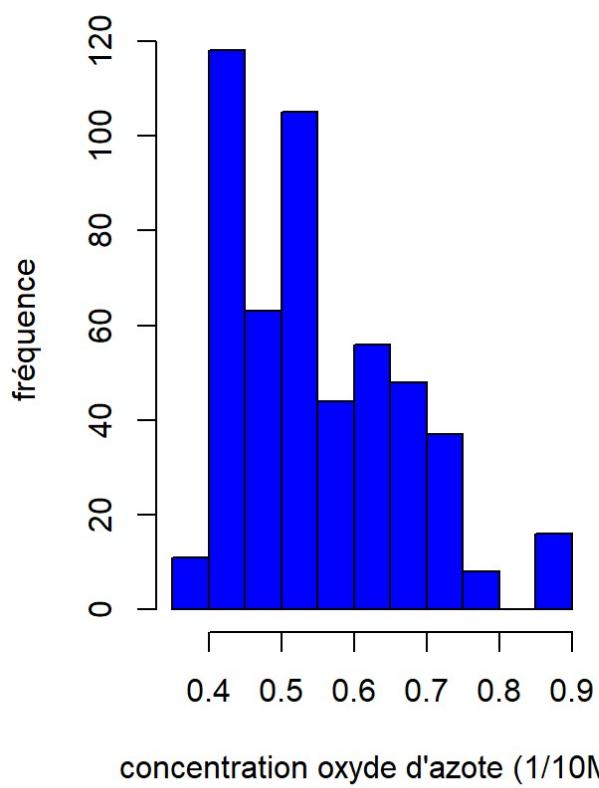
### Valeur médiane du logement et proportion de terrains résidentiels



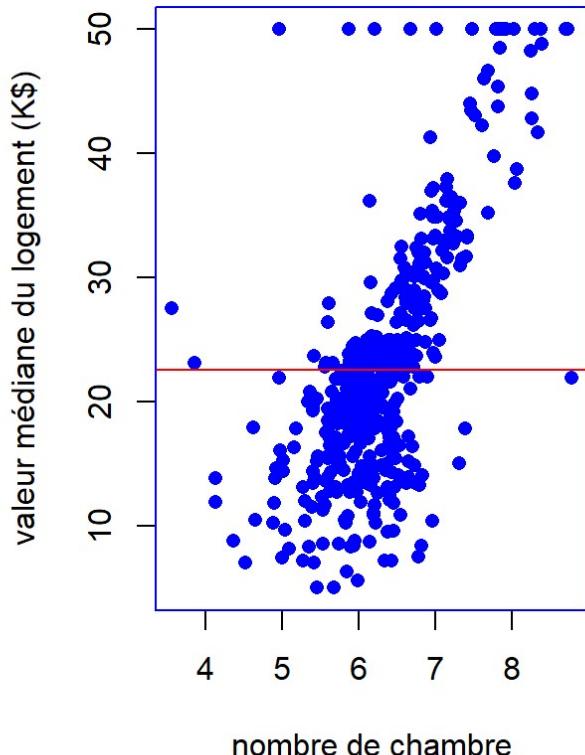
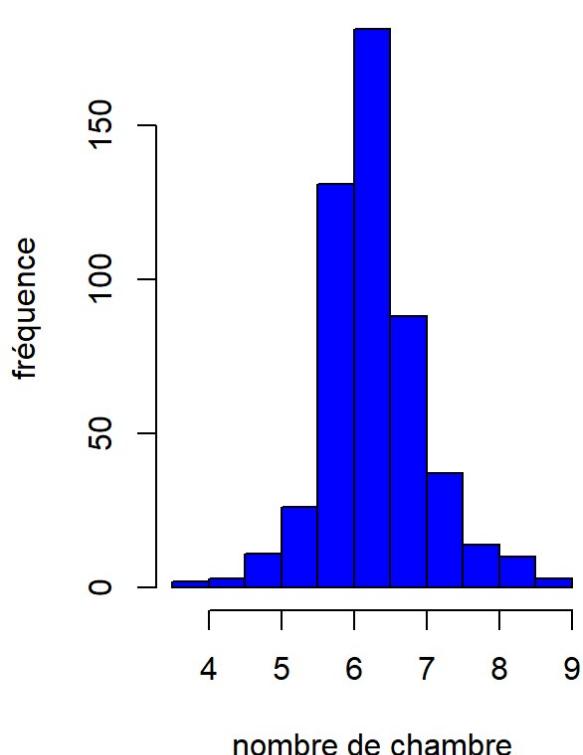
### Valeur médiane du logement et proportion de surface d'activité industrie



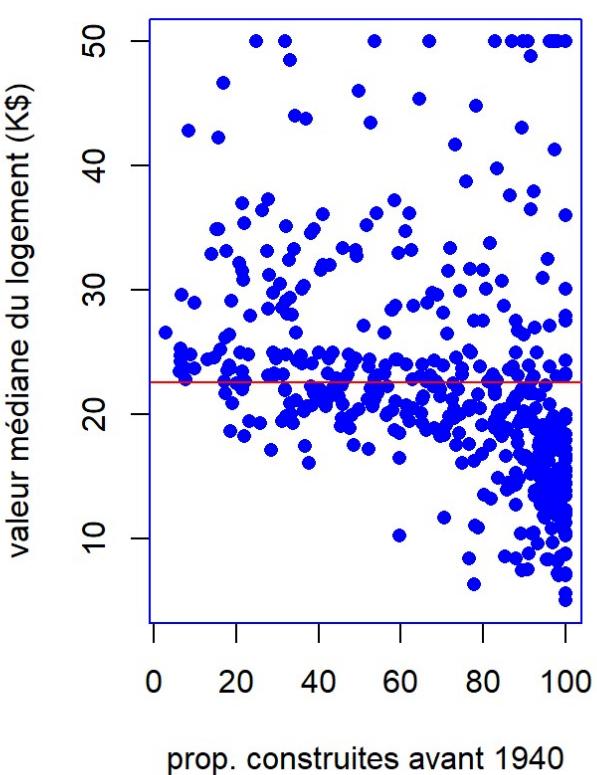
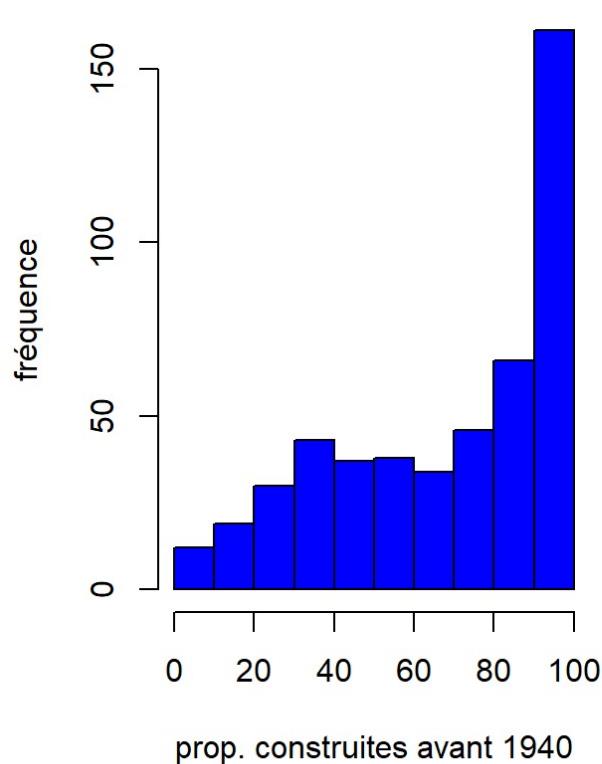
### Valeur médiane du logement et concentration en oxyde d'azote



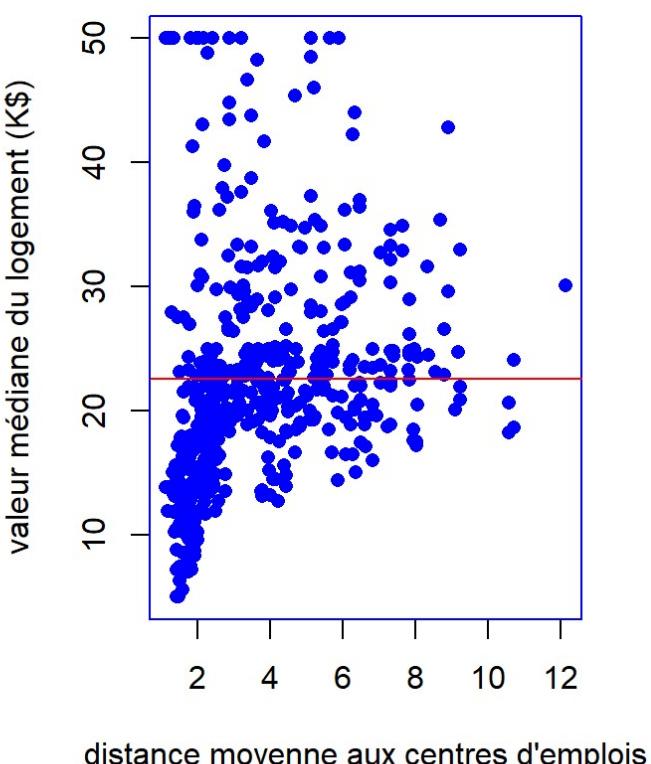
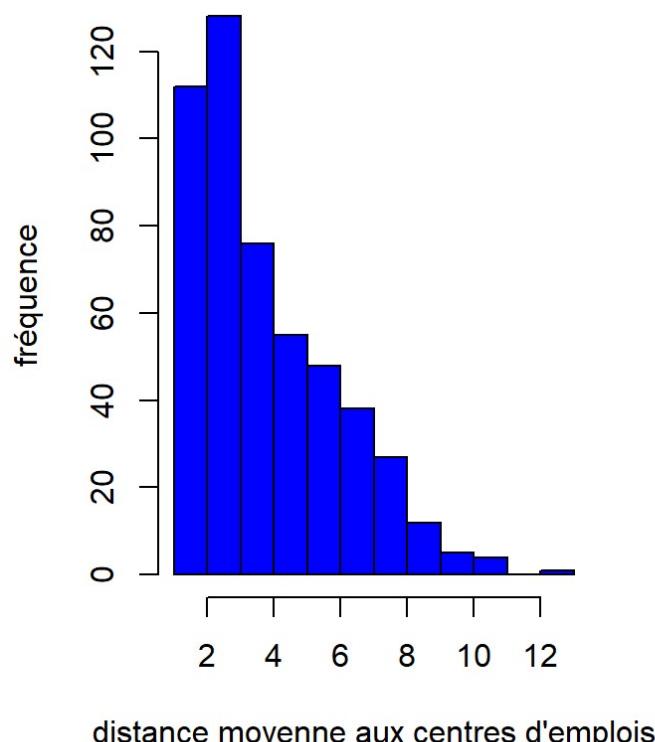
### Valeur médiane du logement et nombre moyen de chambre



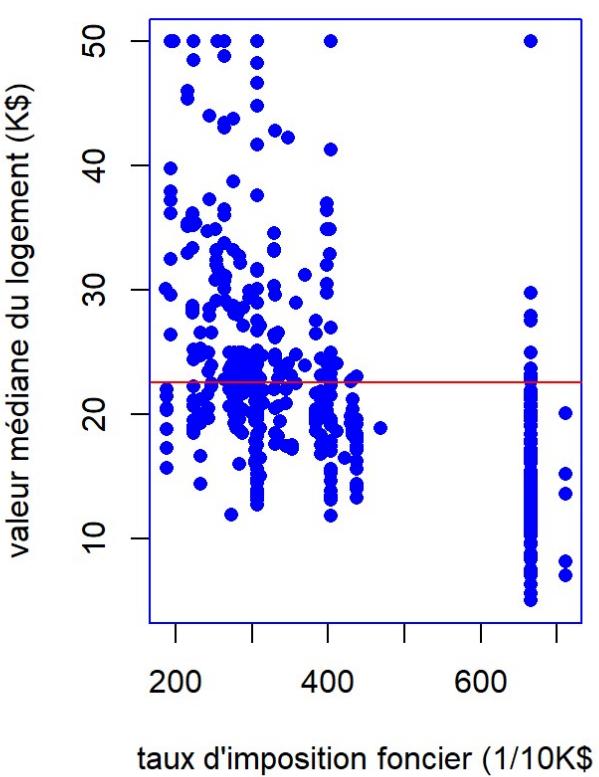
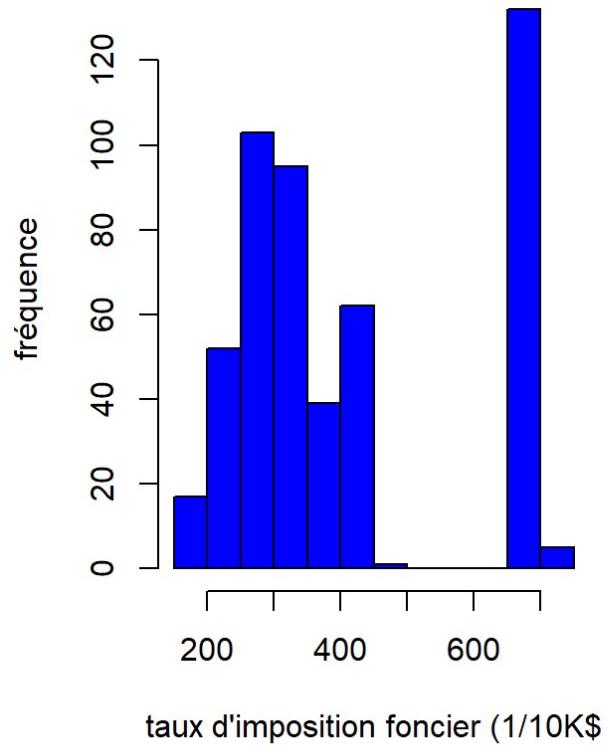
### Valeur du logement et proportion de propriétés construites avant 1940



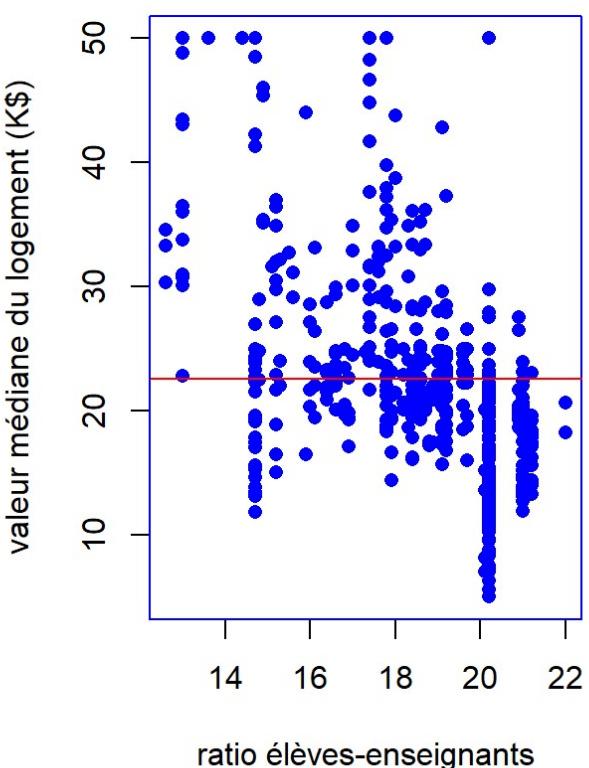
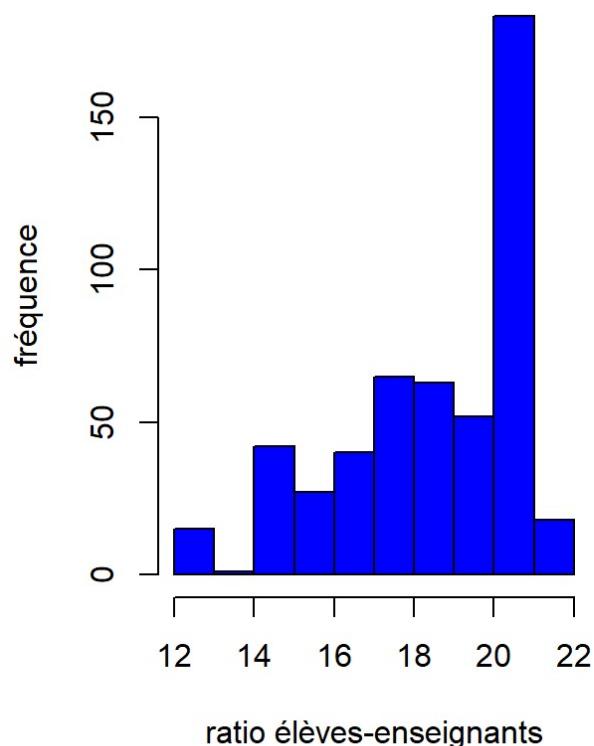
### Valeur du logement et distance moyenne aux 5 centres d'emplois



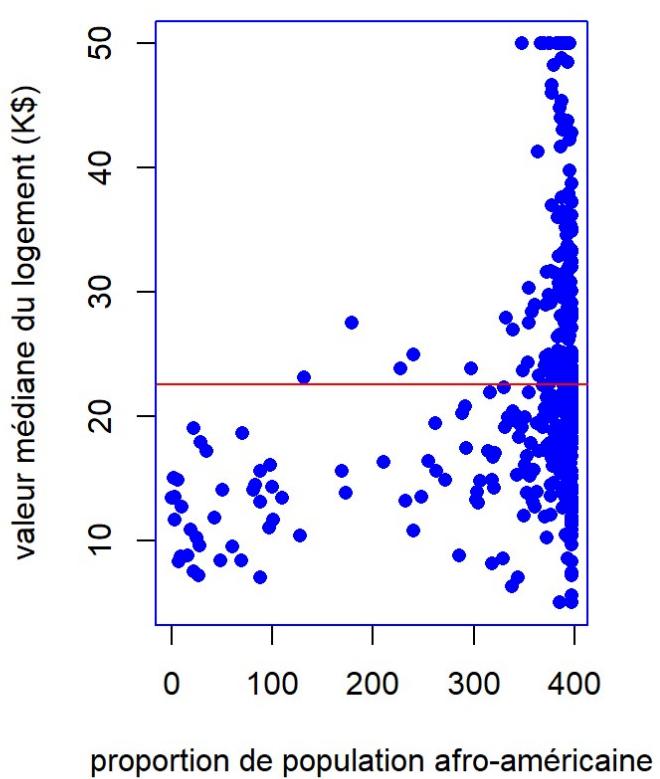
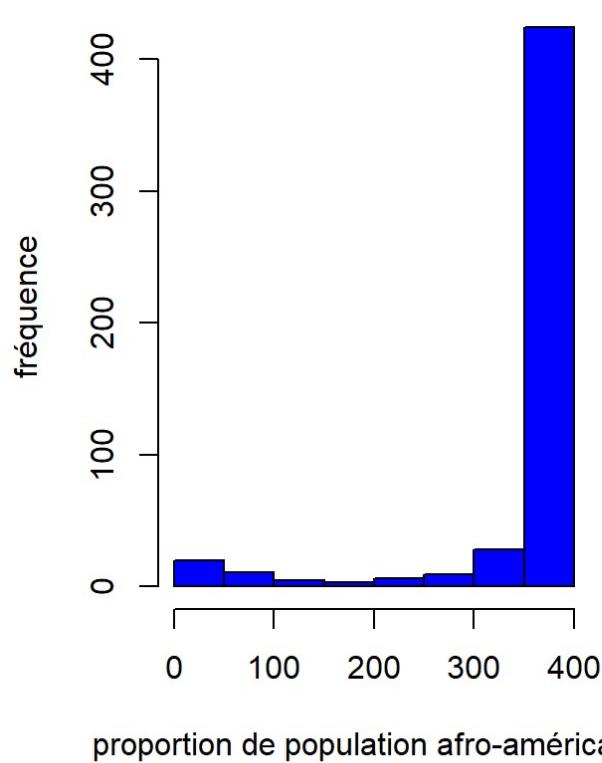
### Valeur médiane du logement et taux d'imposition foncier



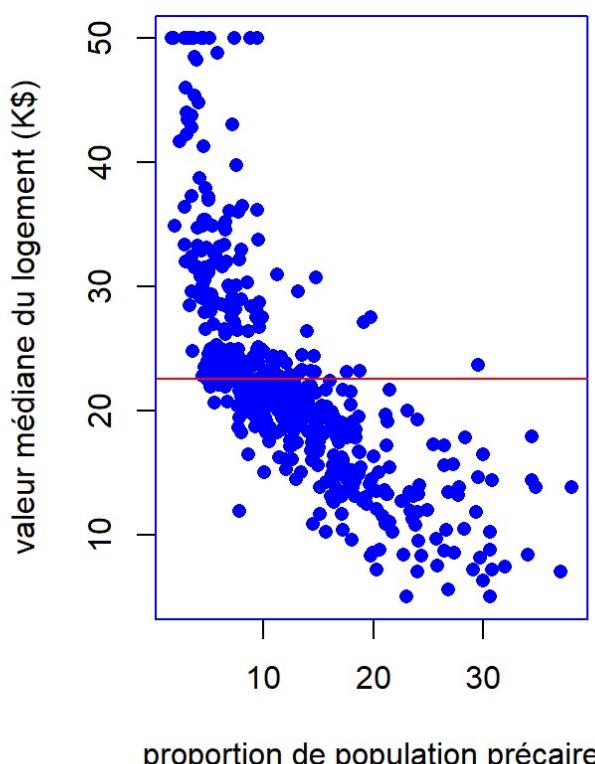
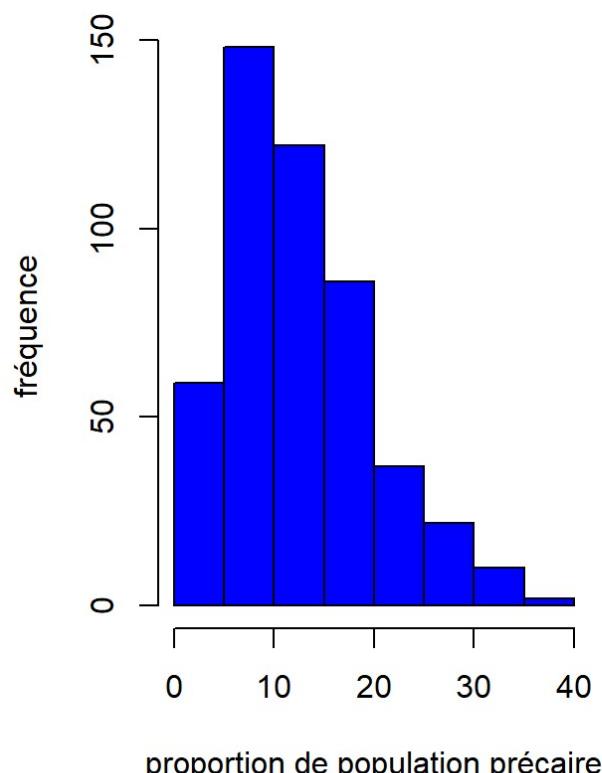
### Valeur médiane du logement et ratio élèves-enseignants

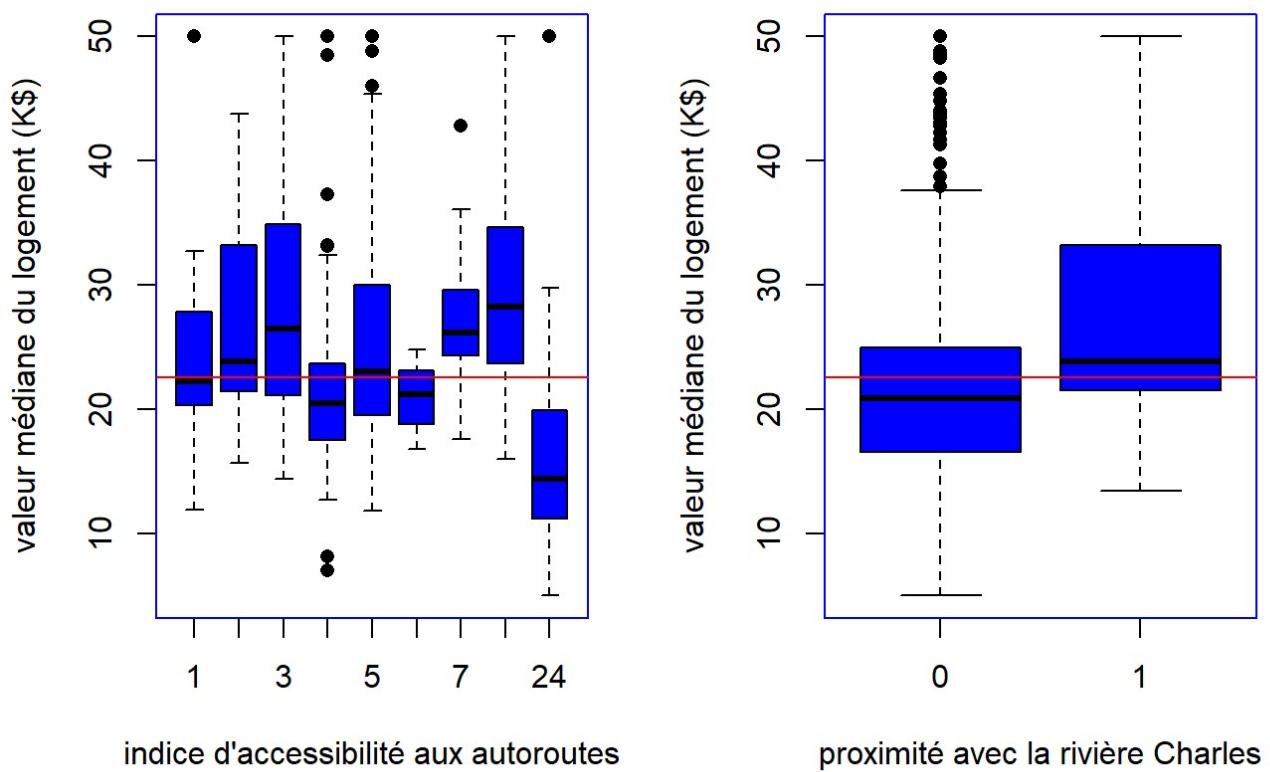


### Valeur du logement et proportion de population afro-américaine



### Valeur médiane du logement et proportion de population précaire





### Annexe 1.2 : corrélation entre les variables du jeu de données

Les corrélations les plus significatives apparaissent en rouge.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	<b>1</b>	-0.19	0.39	-0.05	0.42	-0.23	0.34	-0.37	<b>0.61</b>	<b>0.56</b>	0.27	-0.39	0.46	-0.4
ZN	-0.19	<b>1</b>	<b>-0.52</b>	-0.03	<b>-0.52</b>	0.34	<b>-0.57</b>	<b>0.65</b>	-0.3	-0.31	-0.42	0.17	-0.42	0.41
INDUS	0.39	<b>-0.52</b>	<b>1</b>	0.05	<b>0.76</b>	-0.4	<b>0.64</b>	<b>-0.7</b>	<b>0.59</b>	<b>0.73</b>	0.4	-0.34	<b>0.6</b>	<b>-0.51</b>
CHAS	-0.05	-0.03	0.05	<b>1</b>	0.08	0.1	0.07	-0.1	0.01	-0.03	-0.1	0.07	-0.04	0.17
NOX	0.42	<b>-0.52</b>	<b>0.76</b>	0.08	<b>1</b>	-0.32	<b>0.73</b>	<b>-0.77</b>	<b>0.63</b>	<b>0.68</b>	0.21	-0.38	<b>0.59</b>	-0.46
RM	-0.23	0.34	-0.4	0.1	-0.32	<b>1</b>	-0.25	0.22	-0.24	-0.32	-0.39	0.12	<b>-0.64</b>	<b>0.72</b>
AGE	0.34	<b>-0.57</b>	<b>0.64</b>	0.07	<b>0.73</b>	-0.25	<b>1</b>	<b>-0.75</b>	0.44	0.5	0.26	-0.28	<b>0.6</b>	-0.41
DIS	-0.37	<b>0.65</b>	<b>-0.7</b>	-0.1	<b>-0.77</b>	0.22	<b>-0.75</b>	<b>1</b>	-0.48	<b>-0.53</b>	-0.23	0.29	<b>-0.51</b>	0.28
RAD	<b>0.61</b>	-0.3	<b>0.59</b>	0.01	<b>0.63</b>	-0.24	0.44	-0.48	<b>1</b>	<b>0.9</b>	0.44	-0.44	<b>0.51</b>	-0.42
TAX	<b>0.56</b>	-0.31	<b>0.73</b>	-0.03	<b>0.68</b>	-0.32	0.5	<b>-0.53</b>	<b>0.9</b>	<b>1</b>	0.45	-0.44	<b>0.57</b>	<b>-0.51</b>
PTRATIO	0.27	-0.42	0.4	-0.1	0.21	-0.39	0.26	-0.23	0.44	0.45	<b>1</b>	-0.18	0.4	<b>-0.54</b>
B	-0.39	0.17	-0.34	0.07	-0.38	0.12	-0.28	0.29	-0.44	-0.44	-0.18	<b>1</b>	-0.38	0.35
LSTAT	0.46	-0.42	<b>0.6</b>	-0.04	<b>0.59</b>	<b>-0.64</b>	<b>0.6</b>	<b>-0.51</b>	<b>0.51</b>	<b>0.57</b>	0.4	-0.38	<b>1</b>	<b>-0.74</b>
MEDV	-0.4	0.41	<b>-0.51</b>	0.17	-0.46	<b>0.72</b>	-0.41	0.28	-0.42	<b>-0.51</b>	<b>-0.54</b>	0.35	<b>-0.74</b>	<b>1</b>

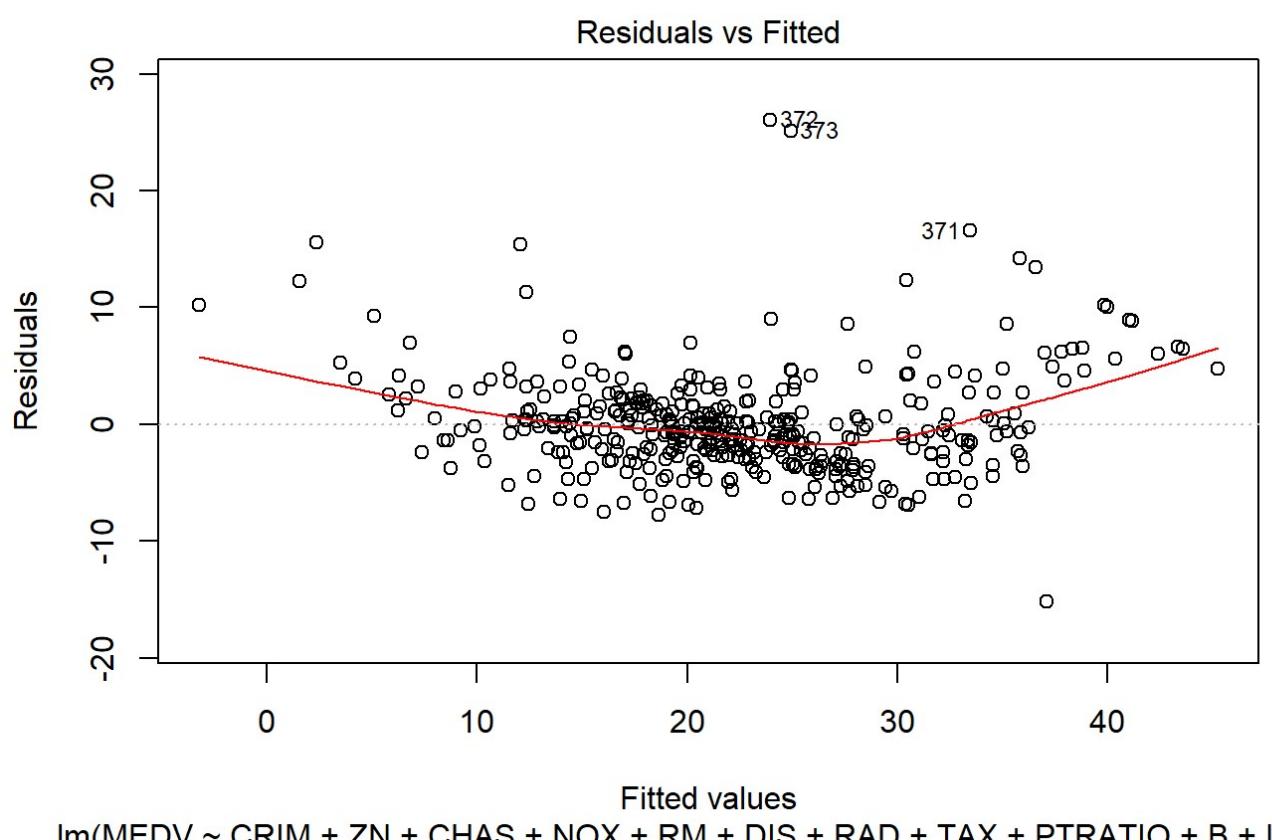
### Annexe 1.3 : régression linéaire MEDV sur l'ensemble des variables

```
##  
## Call:  
## lm(formula = MEDV ~ ., data = dHB)  
##  
## Residuals:  
##       Min        1Q    Median        3Q       Max  
## -15.4234 -2.5830 -0.5079  1.6681 26.2604  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 32.680059   5.681290  5.752 1.81e-08 ***  
## CRIM        -0.097594   0.032457 -3.007 0.002815 **  
## ZN          0.048905   0.014398  3.397 0.000754 ***  
## INDUS       0.030379   0.065933  0.461 0.645237  
## CHAS         2.769378   0.925171  2.993 0.002940 **  
## NOX        -17.969028   4.242856 -4.235 2.87e-05 ***  
## RM          4.283252   0.470710  9.100 < 2e-16 ***  
## AGE        -0.012991   0.014459 -0.898 0.369504  
## DIS         -1.458510   0.211007 -6.912 2.03e-11 ***  
## RAD         0.285866   0.069298  4.125 4.55e-05 ***  
## TAX        -0.013146   0.003955 -3.324 0.000975 ***  
## PTRATIO     -0.914582   0.140581 -6.506 2.44e-10 ***  
## B           0.009656   0.002970  3.251 0.001251 **  
## LSTAT      -0.423661   0.055022 -7.700 1.19e-13 ***  
## ---  
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.487 on 380 degrees of freedom  
##   (112 observations deleted due to missingness)  
## Multiple R-squared:  0.7671, Adjusted R-squared:  0.7591  
## F-statistic: 96.29 on 13 and 380 DF,  p-value: < 2.2e-16
```

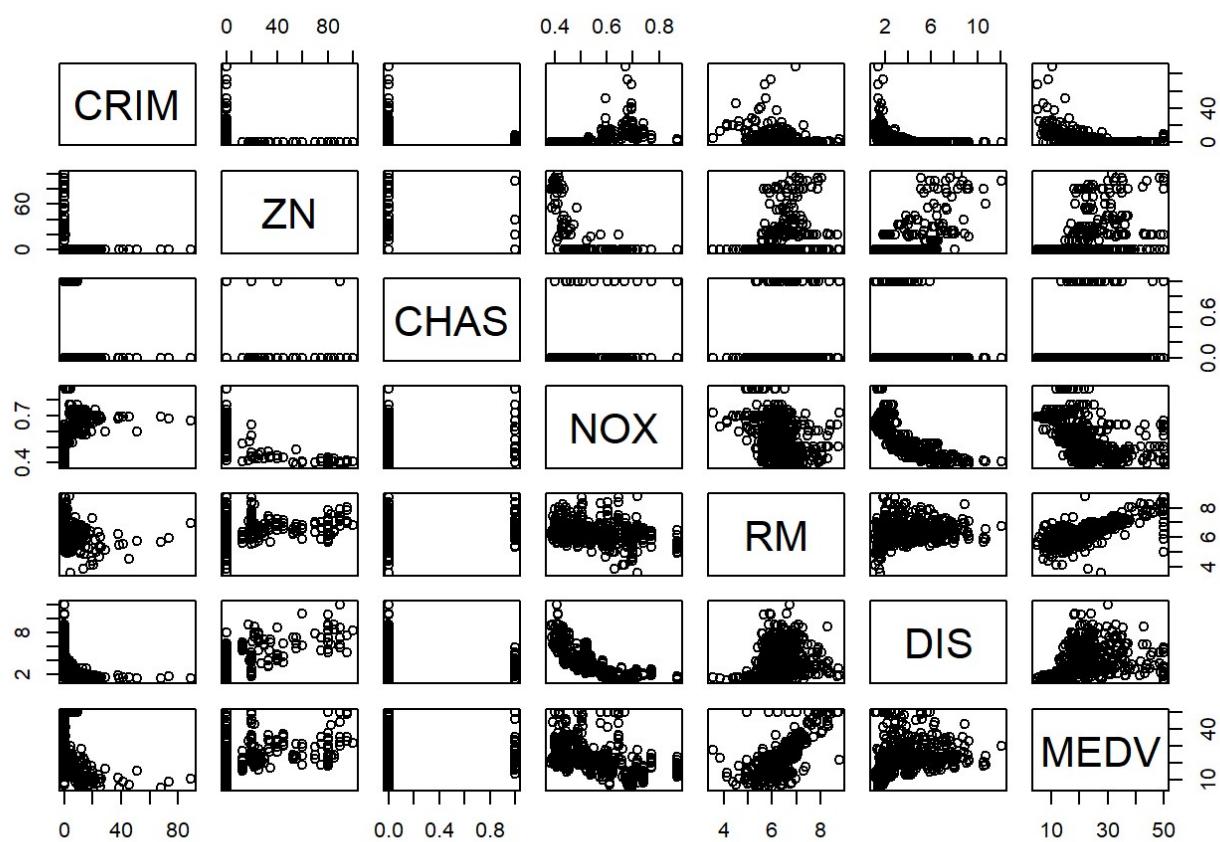
#### Annexe 1.4 : regression linéaire MEDV sur variables choisies par stepAIC

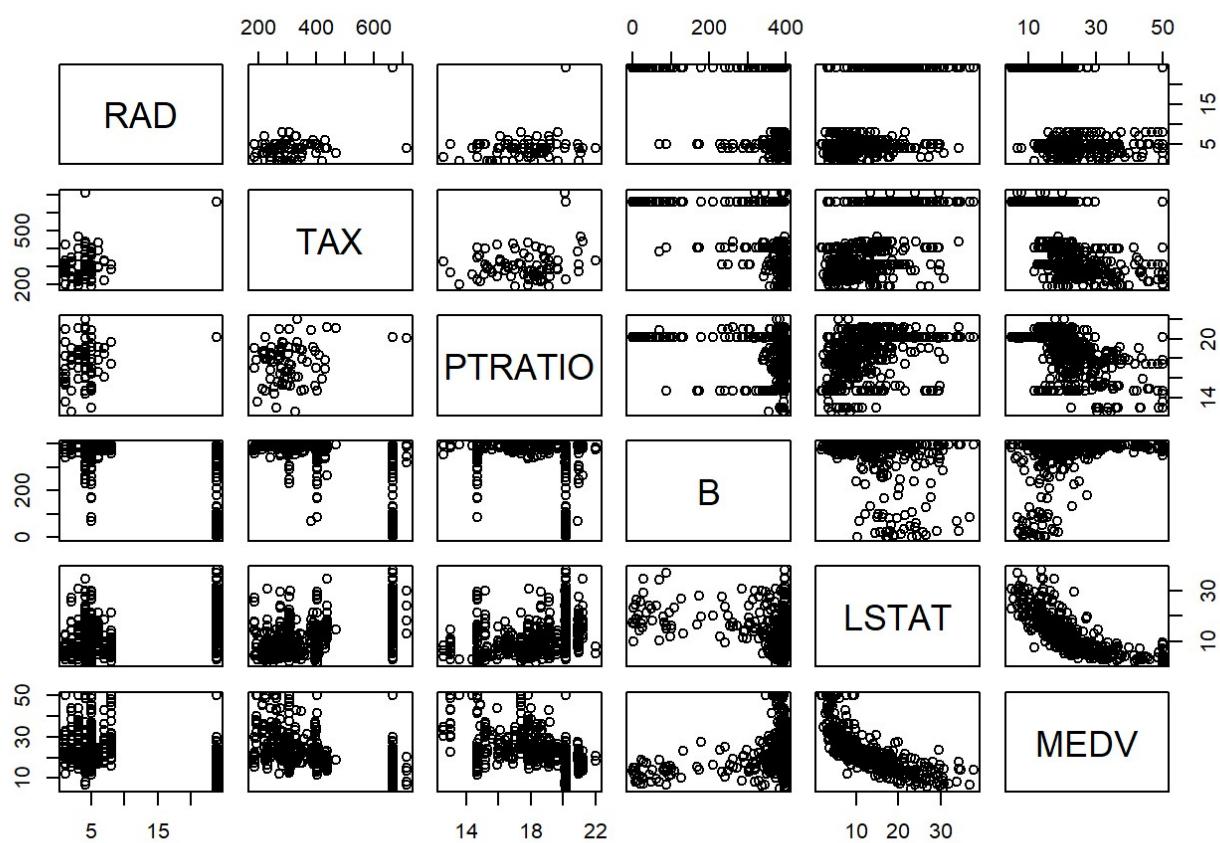
```
##  
## Call:  
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +  
##      TAX + PTRATIO + B + LSTAT, data = dHB2)  
##  
## Residuals:  
##       Min     1Q Median     3Q    Max  
## -15.214 -2.552 -0.503  1.768 26.027  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 32.975051   5.630782   5.856 1.02e-08 ***  
## CRIM        -0.098151   0.032405  -3.029 0.002621 **  
## ZN          0.049962   0.014169   3.526 0.000473 ***  
## CHAS         2.788061   0.919721   3.031 0.002600 **  
## NOX        -18.467815   3.895303  -4.741 3.01e-06 ***  
## RM          4.166982   0.455473   9.149 < 2e-16 ***  
## DIS         -1.420599   0.197272  -7.201 3.20e-12 ***  
## RAD         0.282322   0.065525   4.309 2.09e-05 ***  
## TAX        -0.012400   0.003471  -3.573 0.000398 ***  
## PTRATIO     -0.914756   0.138631  -6.599 1.39e-10 ***  
## B           0.009477   0.002961   3.201 0.001483 **  
## LSTAT       -0.439994   0.051567  -8.532 3.41e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.481 on 382 degrees of freedom  
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.7598  
## F-statistic: 114 on 11 and 382 DF,  p-value: < 2.2e-16
```

## Annexe 1.5 : Graphique des résidus du modèle

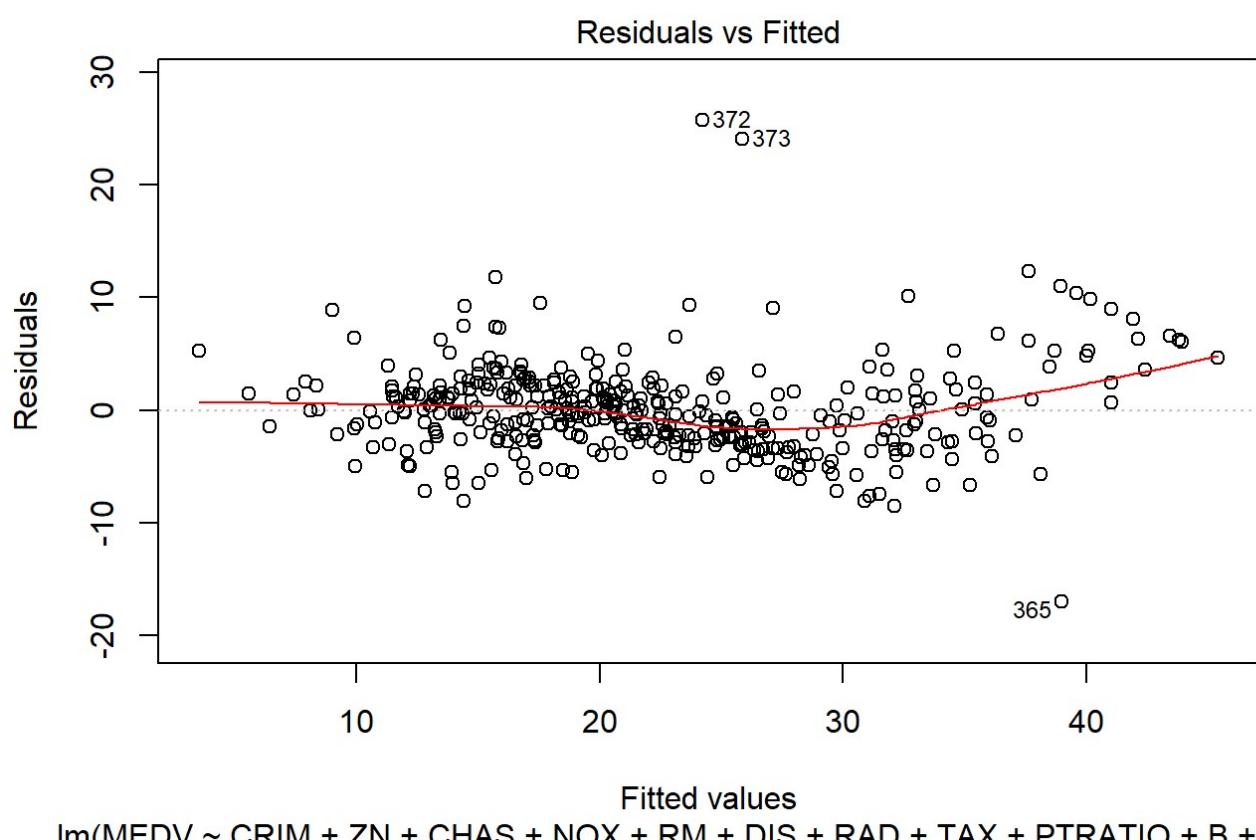


### Annexe 1.6 : pairs plot

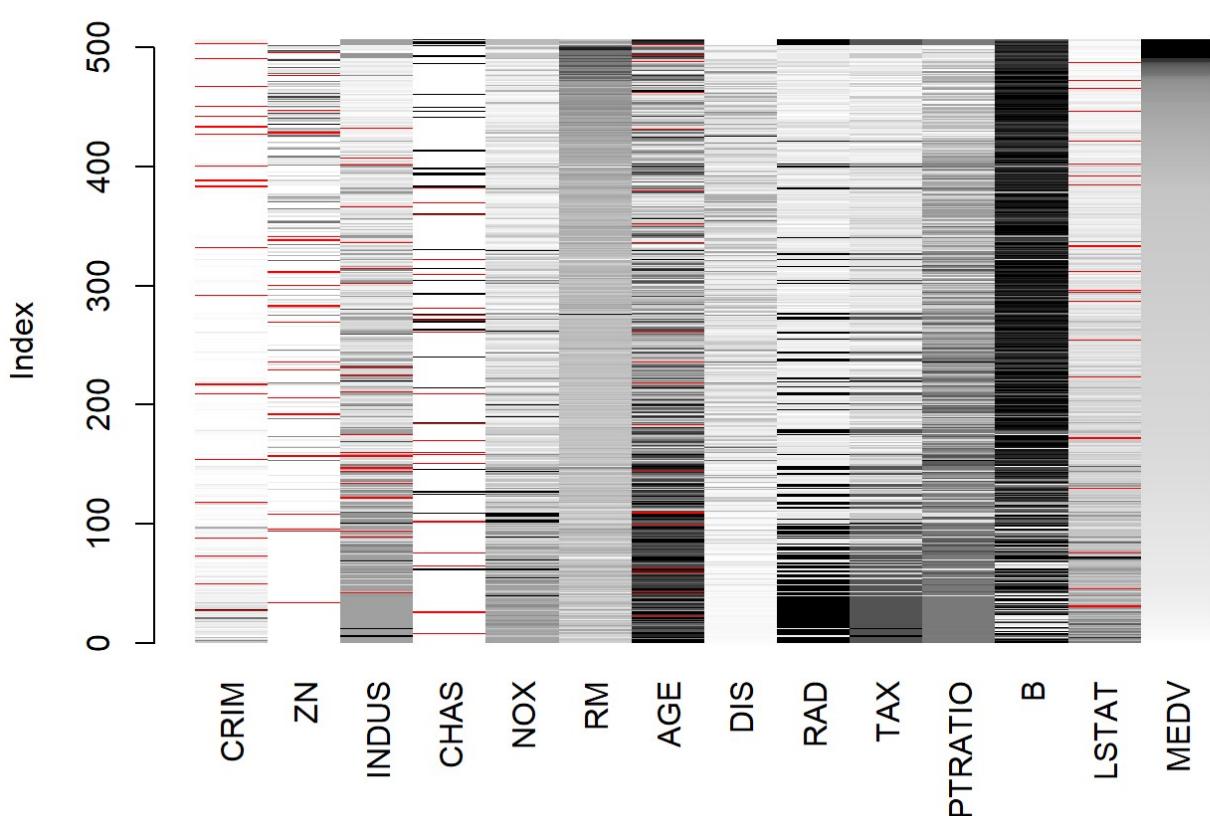


**Annexe 1.7 : régression linéaire multiple avec ordre 2**

```
##  
## Call:  
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +  
##       TAX + PTRATIO + B + LSTAT + I(LSTAT^2), data = dHB2)  
##  
## Residuals:  
##       Min     1Q Median     3Q    Max  
## -17.0589 -2.4364 -0.2657  1.8691 25.7721  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 38.104411   5.218784   7.301 1.68e-12 ***  
## CRIM        -0.130258   0.030072  -4.332 1.90e-05 ***  
## ZN          0.030435   0.013249   2.297  0.02215 *  
## CHAS         2.778685   0.846518   3.282  0.00112 **  
## NOX        -14.717306   3.613208  -4.073 5.64e-05 ***  
## RM          3.585063   0.424957   8.436 6.88e-16 ***  
## DIS         -1.346050   0.181789  -7.404 8.53e-13 ***  
## RAD         0.249627   0.060436   4.130 4.45e-05 ***  
## TAX        -0.010198   0.003205  -3.182  0.00158 **  
## PTRATIO     -0.759891   0.128934  -5.894 8.32e-09 ***  
## B           0.009128   0.002725   3.349  0.00089 ***  
## LSTAT        -1.455094   0.130342 -11.164 < 2e-16 ***  
## I(LSTAT^2)    0.028375   0.003393   8.362 1.17e-15 ***  
## ---  
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.125 on 381 degrees of freedom  
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7965  
## F-statistic: 129.2 on 12 and 381 DF,  p-value: < 2.2e-16
```



### Annexe 2.1 : Matrixplot données manquantes



### Annexe 3.1 : Imputation - méthodologie

L'imputation correspond à l'action de convertir un échantillon incomplet en un échantillon complet. Le but de l'imputation multiple est d'affecter plusieurs fois des données manquantes, d'analyser les données complétées et ensuite d'intégrer les résultats des analyses.

Les 7 étapes de l'imputation:

**Etape 1** - Décider si supposition de MAR est plausible.

(vu en partie 2)

**Etape 2** - Identifier la forme du modèle d'imputation.

Le choix sera orienté par l'échelle de la variable à imputer, et intègre de préférence la connaissance de la relation entre les variables. L'algorithme MICE a besoin d'avoir une méthode univariée d'imputation pour chaque variable incomplète.

**Etape 3** - Sélectionner le groupe de variables à inclure comme prédicteurs dans le modèle d'imputation (fonction **mice**)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	0	1	1	1	1	1	1	1	1	1	1	1	1	1
ZN	1	0	1	1	1	1	1	1	1	1	1	1	1	1
INDUS	1	1	0	1	1	1	1	1	1	1	1	1	1	1
CHAS	1	1	1	0	1	1	1	1	1	1	1	1	1	1
NOX	1	1	1	1	0	1	1	1	1	1	1	1	1	1
RM	1	1	1	1	1	0	1	1	1	1	1	1	1	1
AGE	1	1	1	1	1	1	0	1	1	1	1	1	1	1
DIS	1	1	1	1	1	1	1	0	1	1	1	1	1	1
RAD	1	1	1	1	1	1	1	1	0	1	1	1	1	1
TAX	1	1	1	1	1	1	1	1	1	0	1	1	1	1
PTRATIO	1	1	1	1	1	1	1	1	1	1	0	1	1	1
B	1	1	1	1	1	1	1	1	1	1	1	0	1	1
LSTAT	1	1	1	1	1	1	1	1	1	1	1	1	0	1
MEDV	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Selon la matrice de résultat, CRIM sera prédit à partir de toutes les autres variables (indicateur = 1); idem pour ZN, INDUS, CHAS, AGE et LSTAT. Nous allons utiliser toutes les variables comme prédictrices. Cela est possible car le dataset est encore de taille raisonnable, (difficile sur les grands datasets , à cause de la multicolinearité ou de la capacité des machines)

**Etape 4** - Imputer ou non des variables qui sont des fonctions d'autres variables incomplètes.

Dans le cas de notre dataset, les variables avec des données manquantes ne sont pas des fonctions d'autres variables du dataset. Chacune représente une thématique différente, utile pour l'estimation de la valeur de la maison.

**Etape 5** - Définir l'ordre d'imputation des variables (influence sur la convergence de l'algorithme). Par défaut, l'algorithme MICE imprime les données incomplètes du dataset de gauche à droite. L'ordre est à changer si on a des soucis de convergence des algorithmes.

**Etape 6** - Définir les imputations de départ et le nombre d'itérations

**Etape 7** - Imputer et ajuster le modèle L'imputation du dataset demande de faire des "essais-erreur", pour adapter et améliorer le modèle. Pour démarrer, il est conseillé de mettre m = 5 et l'augmenter lors de la dernière étape si on est déjà satisfait avec le modèle.

### Annexe 3.2 : Structure des données imputées par régression stochastique

```
##      CRIM          ZN          INDUS          CHAS
## Min. :-11.28958   Min. :-22.78   Min. : 0.460   Min. :-0.64922
## 1st Qu.: 0.08057   1st Qu.: 0.00   1st Qu.: 5.190   1st Qu.: 0.00000
## Median : 0.26600   Median : 0.00   Median : 9.489   Median : 0.00000
## Mean   : 3.61817   Mean   : 11.50   Mean   :11.078   Mean   : 0.06542
## 3rd Qu.: 3.68939   3rd Qu.: 12.50   3rd Qu.:18.100   3rd Qu.: 0.00000
## Max.  : 88.97620   Max.  :100.00   Max.  :27.740   Max.  : 1.00000
##      NOX          RM          AGE          DIS
## Min. :0.3850     Min. :3.561    Min. : 2.90    Min. : 1.130
## 1st Qu.:0.4490    1st Qu.:5.886    1st Qu.: 45.45   1st Qu.: 2.100
## Median :0.5380    Median :6.208    Median : 76.80   Median : 3.207
## Mean   :0.5547    Mean   :6.285    Mean   : 68.54   Mean   : 3.795
## 3rd Qu.:0.6240    3rd Qu.:6.623    3rd Qu.: 93.80   3rd Qu.: 5.188
## Max.  :0.8710    Max.  :8.780    Max.  :105.91   Max.  :12.127
##      RAD          TAX          PTRATIO         B
## Min. : 1.000     Min. :187.0    Min. :12.60    Min. : 0.32
## 1st Qu.: 4.000     1st Qu.:279.0    1st Qu.:17.40   1st Qu.:375.38
## Median : 5.000     Median :330.0    Median :19.05   Median :391.44
## Mean   : 9.549     Mean   :408.2    Mean   :18.46   Mean   :356.67
## 3rd Qu.:24.000     3rd Qu.:666.0    3rd Qu.:20.20   3rd Qu.:396.23
## Max.  :24.000     Max.  :711.0    Max.  :22.00   Max.  :396.90
##      LSTAT        MEDV
## Min. :-2.211     Min. : 5.00
## 1st Qu.: 6.950     1st Qu.:17.02
## Median :11.360     Median :21.20
## Mean   :12.629     Mean   :22.53
## 3rd Qu.:16.930     3rd Qu.:25.00
## Max.  :37.970     Max.  :50.00
```

```

##      CRIM          ZN          INDUS          CHAS
## Min.   : 0.00632  Min.   : 0.00  Min.   : 0.46  Min.   :0.00000
## 1st Qu.: 0.08190  1st Qu.: 0.00  1st Qu.: 5.19  1st Qu.:0.00000
## Median : 0.25372  Median : 0.00  Median : 9.69  Median :0.00000
## Mean    : 3.61187  Mean   : 11.21  Mean   :11.08  Mean   :0.06996
## 3rd Qu.: 3.56026  3rd Qu.: 12.50  3rd Qu.:18.10  3rd Qu.:0.00000
## Max.   :88.97620  Max.   :100.00  Max.   :27.74  Max.   :1.00000
## NA's    :20        NA's   :20    NA's   :20    NA's   :20

##      NOX          RM          AGE          DIS
## Min.   :0.3850    Min.   :3.561  Min.   : 2.90  Min.   : 1.130
## 1st Qu.:0.4490    1st Qu.:5.886  1st Qu.: 45.17  1st Qu.: 2.100
## Median :0.5380    Median :6.208  Median : 76.80  Median : 3.207
## Mean   :0.5547    Mean   :6.285  Mean   : 68.52  Mean   : 3.795
## 3rd Qu.:0.6240    3rd Qu.:6.623  3rd Qu.: 93.97  3rd Qu.: 5.188
## Max.   :0.8710    Max.   :8.780  Max.   :100.00  Max.   :12.127
## NA's   :20

##      RAD          TAX          PTRATIO         B
## Min.   : 1.000  Min.   :187.0  Min.   :12.60  Min.   : 0.32
## 1st Qu.: 4.000  1st Qu.:279.0  1st Qu.:17.40  1st Qu.:375.38
## Median : 5.000  Median :330.0  Median :19.05  Median :391.44
## Mean   : 9.549  Mean   :408.2  Mean   :18.46  Mean   :356.67
## 3rd Qu.:24.000  3rd Qu.:666.0  3rd Qu.:20.20  3rd Qu.:396.23
## Max.   :24.000  Max.   :711.0  Max.   :22.00  Max.   :396.90
## NA's   :20

##      LSTAT          MEDV
## Min.   : 1.730  Min.   : 5.00
## 1st Qu.: 7.125  1st Qu.:17.02
## Median :11.430  Median :21.20
## Mean   :12.715  Mean   :22.53
## 3rd Qu.:16.955  3rd Qu.:25.00
## Max.   :37.970  Max.   :50.00
## NA's   :20

```

### Annexe 3.3 : Structure des données imputées par forêts aléatoires

```

##      CRIM          ZN          INDUS          CHAS
## Min.   : 0.00632  Min.   : 0.00  Min.   : 0.460  Min.   :0.00000
## 1st Qu.: 0.08204  1st Qu.: 0.00  1st Qu.: 5.145  1st Qu.:0.00000
## Median : 0.25651  Median : 0.00  Median : 9.690  Median :0.00000
## Mean    : 3.62417  Mean   : 11.03  Mean   :11.103  Mean   :0.07115
## 3rd Qu.: 3.68939  3rd Qu.: 12.50  3rd Qu.:18.100  3rd Qu.:0.00000
## Max.    :88.97620  Max.   :100.00  Max.   :27.740  Max.   :1.00000
##      NOX          RM          AGE          DIS
## Min.   :0.3850    Min.   :3.561   Min.   : 2.90   Min.   : 1.130
## 1st Qu.:0.4490    1st Qu.:5.886   1st Qu.: 45.17   1st Qu.: 2.100
## Median :0.5380    Median :6.208   Median : 77.15   Median : 3.207
## Mean    :0.5547    Mean   :6.285   Mean   : 68.58   Mean   : 3.795
## 3rd Qu.:0.6240    3rd Qu.:6.623   3rd Qu.: 94.10   3rd Qu.: 5.188
## Max.    :0.8710    Max.   :8.780   Max.   :100.00   Max.   :12.127
##      RAD          TAX          PTRATIO         B
## Min.   : 1.000    Min.   :187.0   Min.   :12.60   Min.   : 0.32
## 1st Qu.: 4.000    1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
## Median : 5.000    Median :330.0   Median :19.05   Median :391.44
## Mean    : 9.549    Mean   :408.2   Mean   :18.46   Mean   :356.67
## 3rd Qu.:24.000    3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
## Max.    :24.000    Max.   :711.0   Max.   :22.00   Max.   :396.90
##      LSTAT        MEDV
## Min.   : 1.73    Min.   : 5.00
## 1st Qu.: 6.95    1st Qu.:17.02
## Median :11.43    Median :21.20
## Mean    :12.73    Mean   :22.53
## 3rd Qu.:17.06    3rd Qu.:25.00
## Max.    :37.97    Max.   :50.00

```

#### Annexe 3.4 : Structure des données imputées par pmm

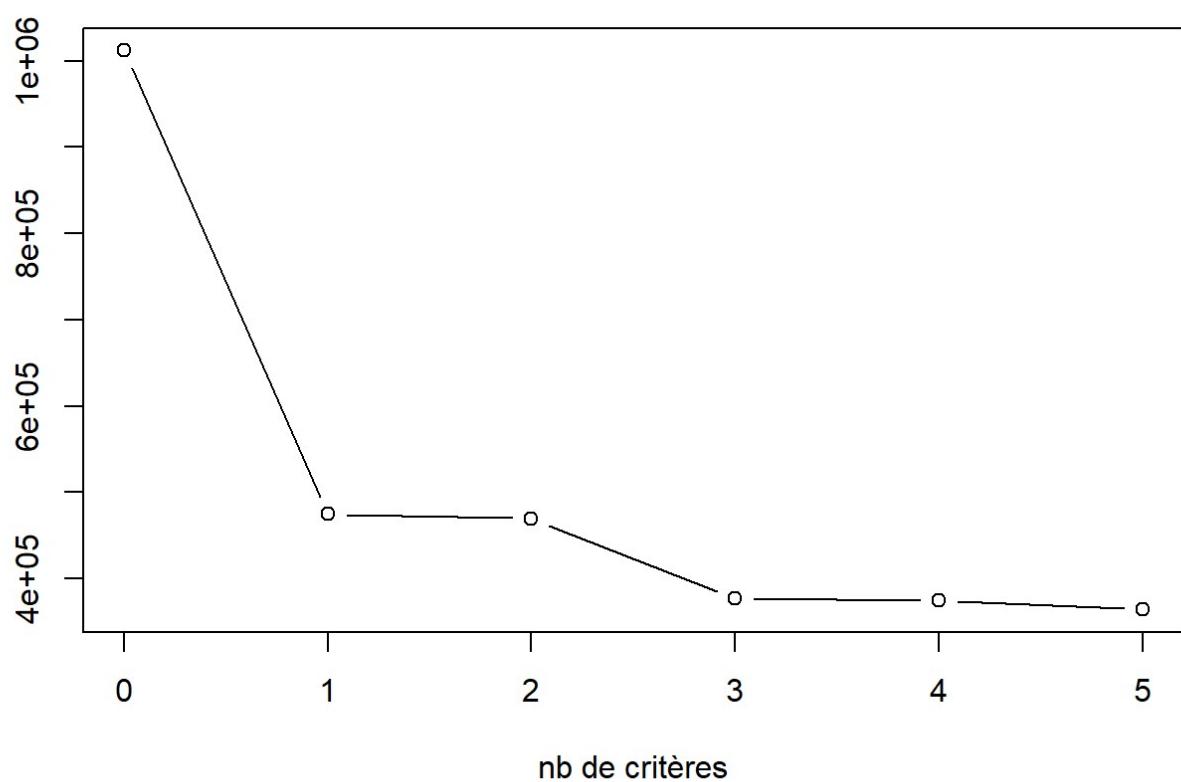
```

##      CRIM          ZN          INDUS          CHAS
## Min.   : 0.00632  Min.   : 0.00  Min.   : 0.46  Min.   :0.0000
## 1st Qu.: 0.08204  1st Qu.: 0.00  1st Qu.: 5.13  1st Qu.:0.0000
## Median : 0.25651  Median : 0.00  Median : 9.69  Median :0.0000
## Mean    : 3.90055  Mean   : 11.49  Mean   :11.07  Mean   :0.0751
## 3rd Qu.: 3.67708  3rd Qu.: 12.50  3rd Qu.:18.10  3rd Qu.:0.0000
## Max.   :88.97620  Max.   :100.00  Max.   :27.74  Max.   :1.0000
##      NOX          RM          AGE          DIS
## Min.   :0.3850    Min.   :3.561   Min.   : 2.90  Min.   : 1.130
## 1st Qu.:0.4490    1st Qu.:5.886   1st Qu.: 45.02  1st Qu.: 2.100
## Median :0.5380    Median :6.208   Median : 77.50  Median : 3.207
## Mean    :0.5547    Mean   :6.285   Mean   : 68.58  Mean   : 3.795
## 3rd Qu.:0.6240    3rd Qu.:6.623   3rd Qu.: 94.10  3rd Qu.: 5.188
## Max.   :0.8710    Max.   :8.780   Max.   :100.00  Max.   :12.127
##      RAD          TAX          PTRATIO         B
## Min.   : 1.000    Min.   :187.0   Min.   :12.60  Min.   : 0.32
## 1st Qu.: 4.000    1st Qu.:279.0   1st Qu.:17.40  1st Qu.:375.38
## Median : 5.000    Median :330.0   Median :19.05  Median :391.44
## Mean    : 9.549    Mean   :408.2   Mean   :18.46  Mean   :356.67
## 3rd Qu.:24.000    3rd Qu.:666.0   3rd Qu.:20.20  3rd Qu.:396.23
## Max.   :24.000    Max.   :711.0   Max.   :22.00  Max.   :396.90
##      LSTAT        MEDV
## Min.   : 1.73    Min.   : 5.00
## 1st Qu.: 6.95    1st Qu.:17.02
## Median :11.27    Median :21.20
## Mean    :12.63    Mean   :22.53
## 3rd Qu.:16.93    3rd Qu.:25.00
## Max.   :37.97    Max.   :50.00

```

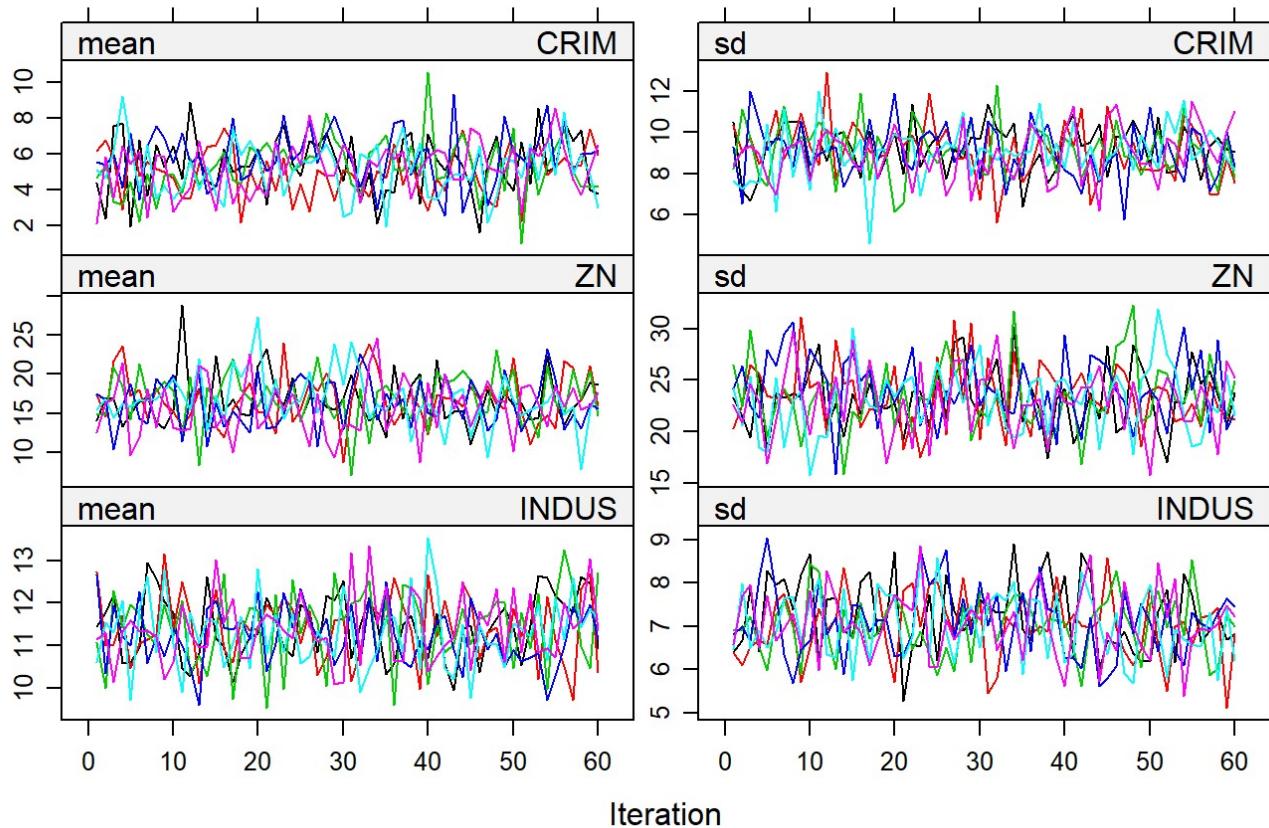
### Annexe 3.5 : Imputation par ACP

graphique des dimensions

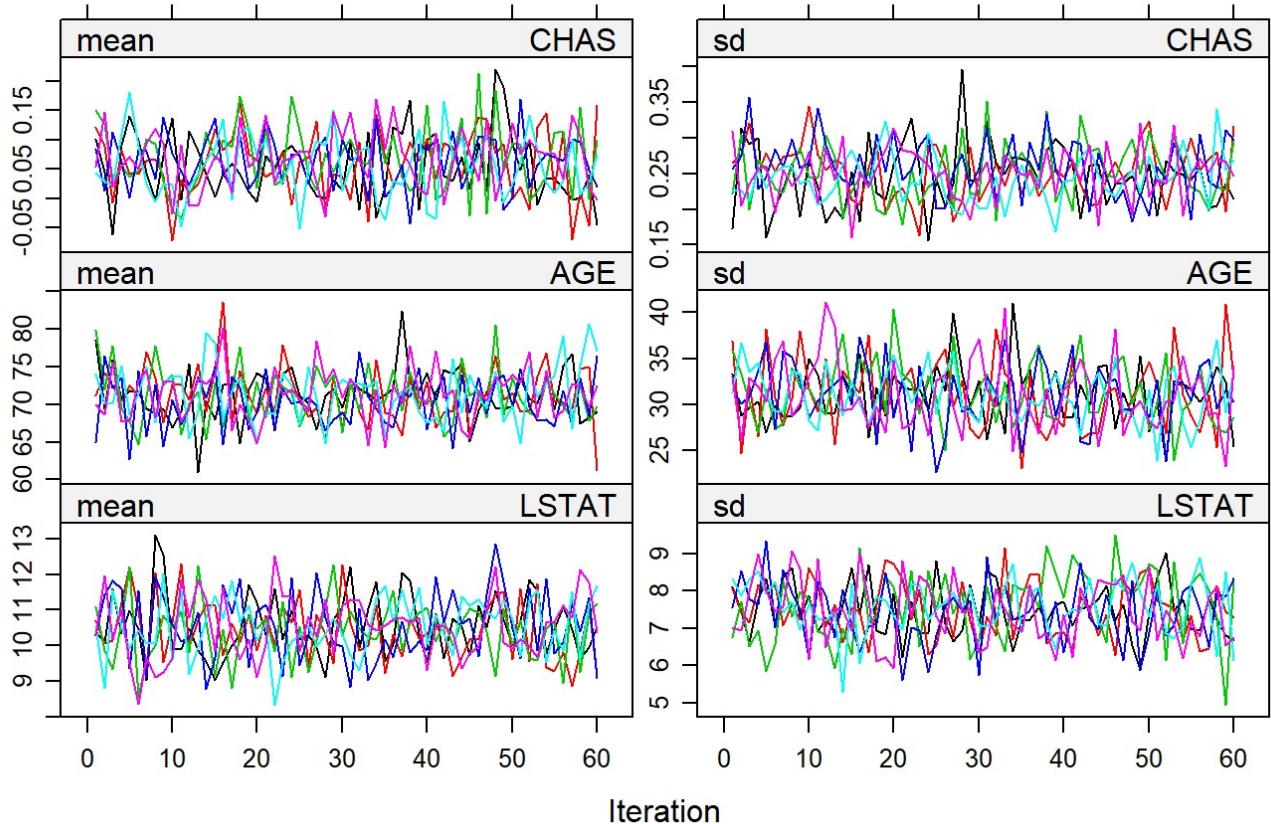


#### Annexe 4.1 Diagnostic 2 : convergence des algorithmes.

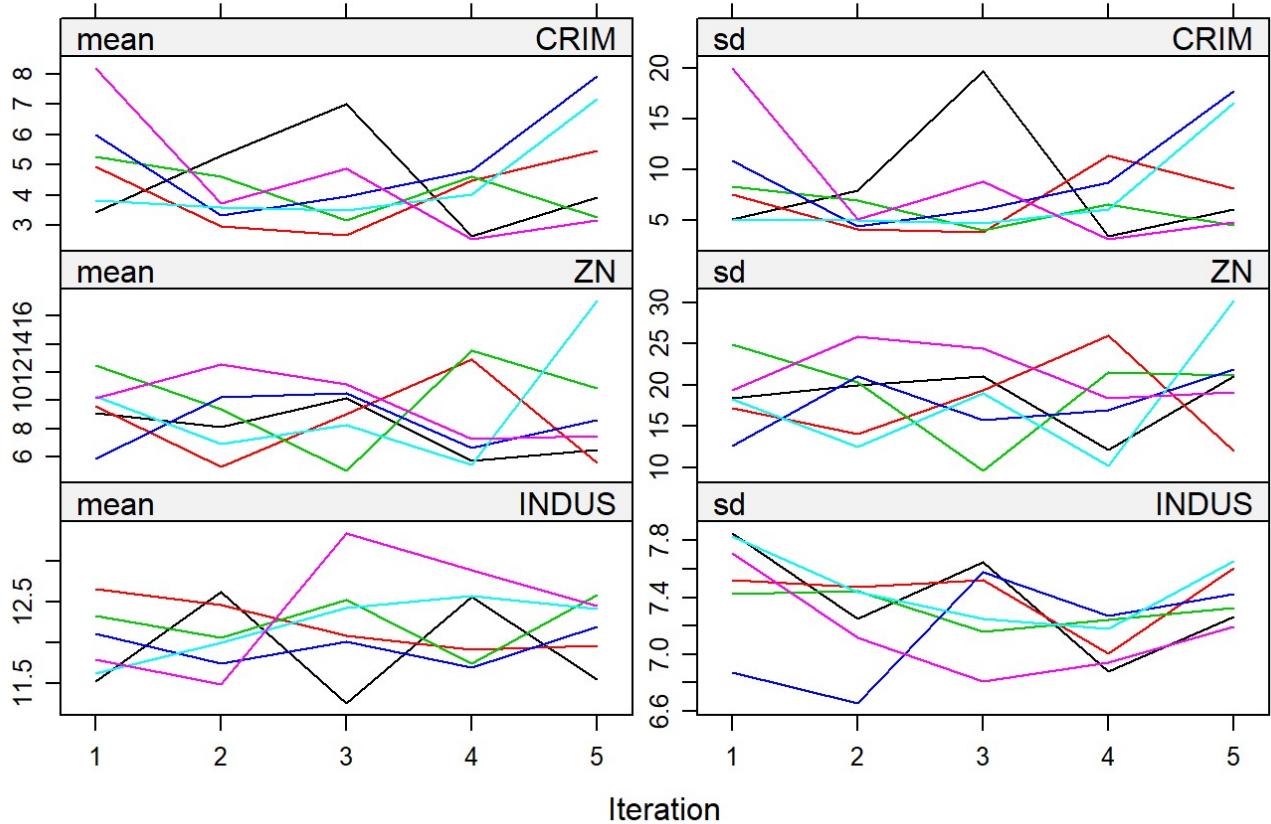
#### Régression Stochastique



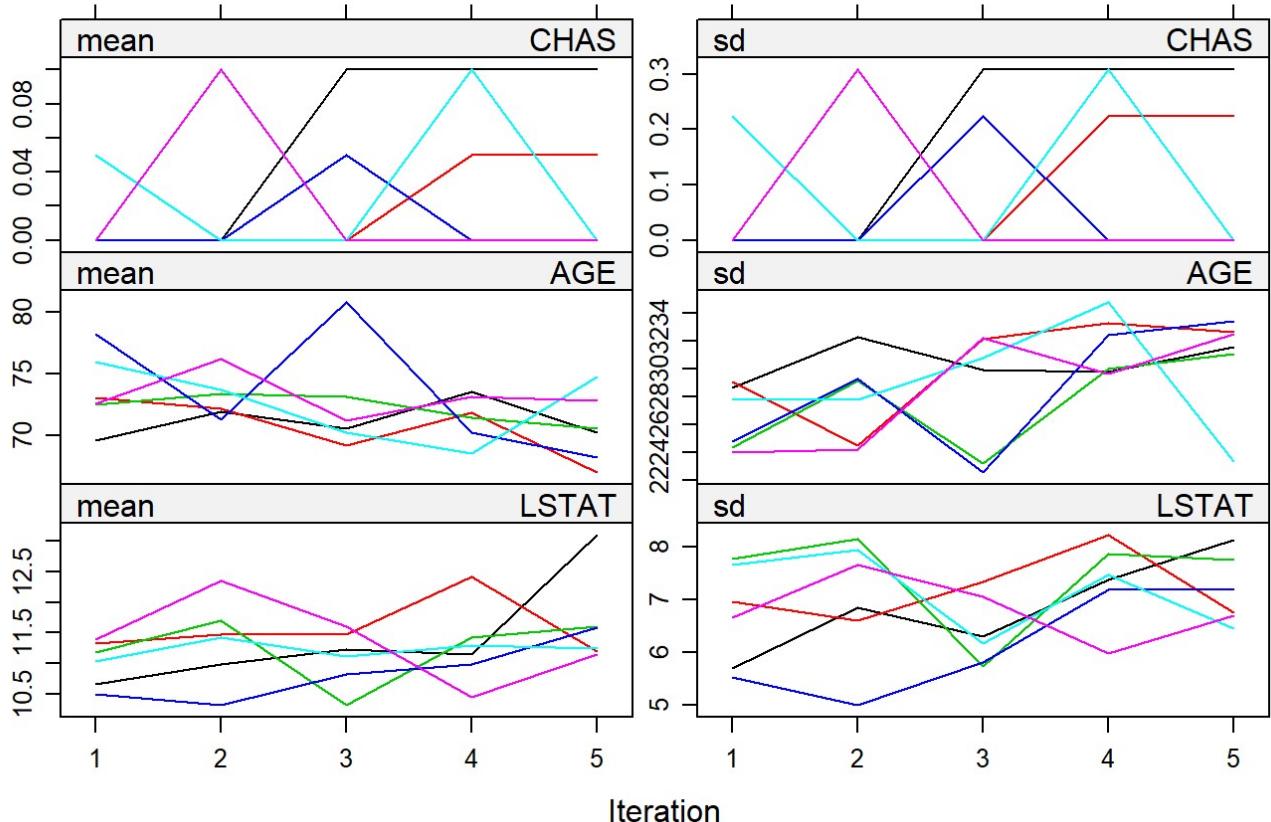
## Régression Stochastique



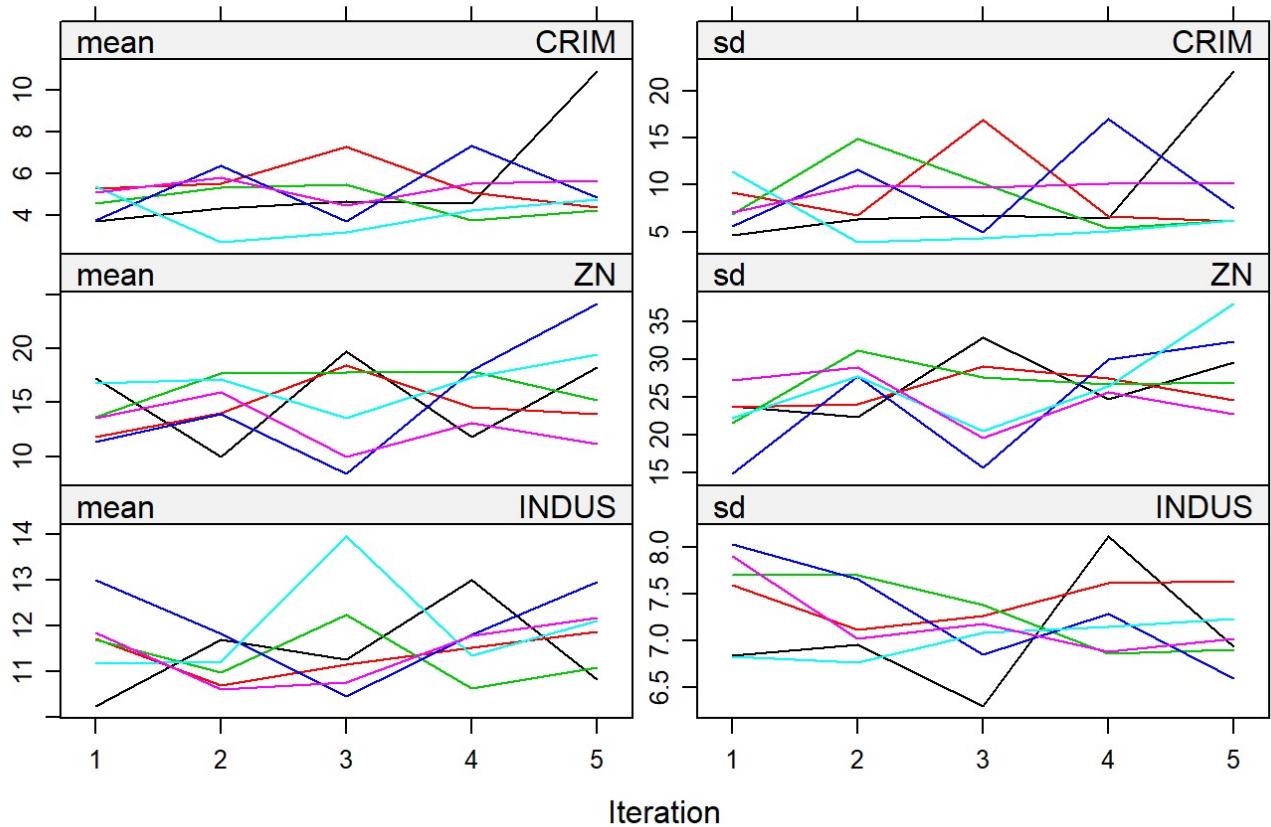
## Forêts aléatoires



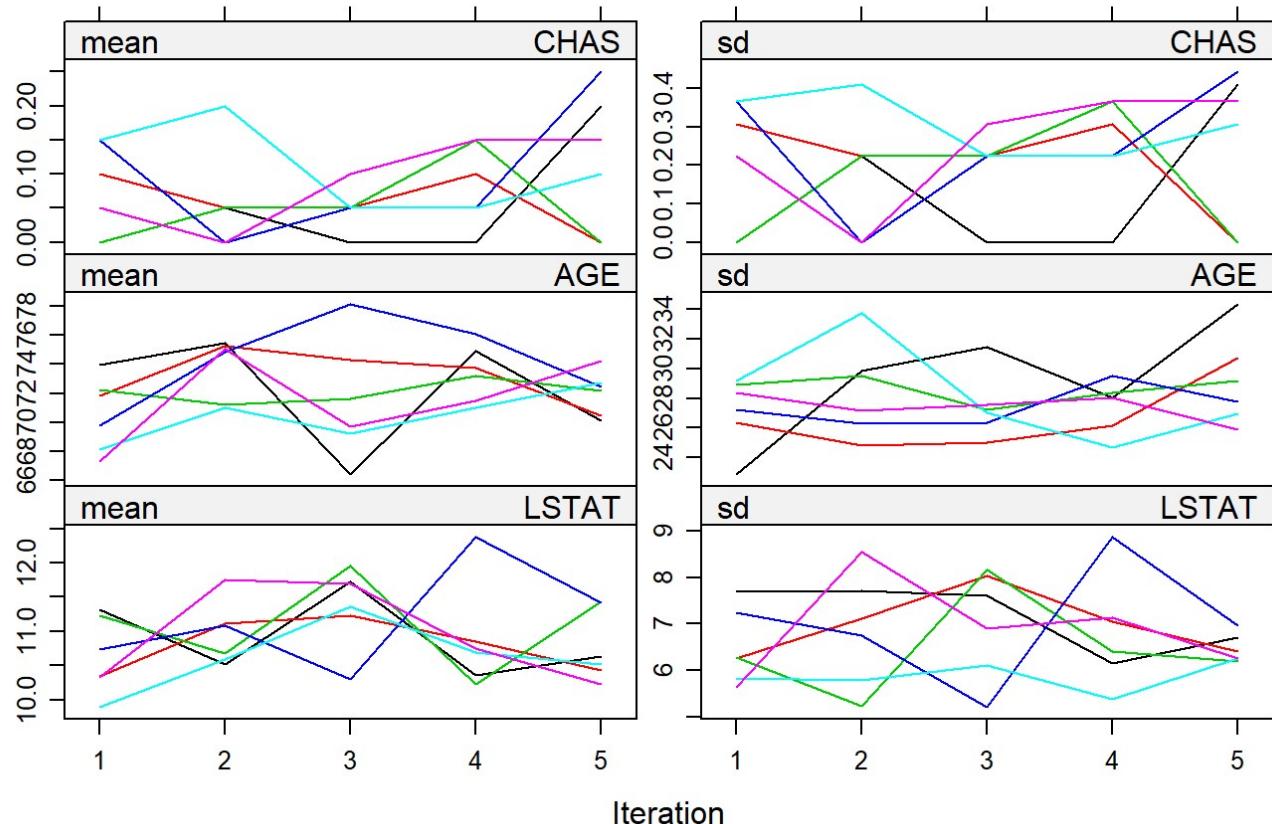
## Forêts aléatoires



## Predictive mean matching



## Predictive mean matching



### Annexe 4.2 : Comparaison des résidus

