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) :

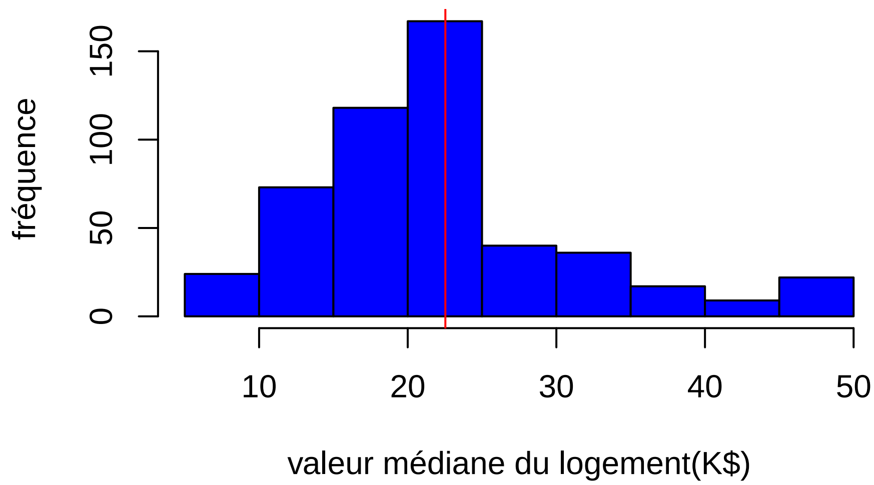
- **CRIM** : Taux de criminalité par habitant.
- **ZN** : Proportion de terrains résidentiels pour des lots de plus de 25000 pieds carré (environ 2300m^2).
- **INDUS** : Proportion d'espace, en acres, consacré aux affaires non commerciales (1 acre environ 4000m^2).
- **CHAS** : Proximité avec la rivière 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 1.2*), on observe que :

- Le **taux de criminalité** faible (inféreur à 10%) est le plus fréquent. La valeur du logement a tendance a 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 (inféreur à 10%), mais lorsqu'elle augmente, la valeur du logement a tendance a augmenté.

- La **proportion de surface d'activité industrielle** la plus fréquente est entre 18 et 20 acres (entre 72 000m² et 80 000m²). 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 logenment (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 elèves-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 1.3*)

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^2ajusté=0.7591) mais il semble que les variables explicatives INDUS et AGE ne soient pas pertinentes pour ce modèle. (**cf. annexe 1.4**)

Nous allons appliquer une méthode "step AIC" pour choisir le meilleur modèle. Cependant pour utiliser cette méthode, nous devons supprimer les lignes avec des données manquantes, avec le risque de perte d'information et d'introduction de biais au dataset. Ce modèle écarte aussi les variables INDUS et AGE avec un R^2 ajusté très proche (0.7598) de l'original. (**cf. annexe 1.5**)

Cependant, en regardant le graphique des résidus (**cf. annexe 1.6**) nous observons que le modèle n'est pas optimal : il semble qu'il y ait une relation non linéaire avec une ou plusieurs variables. Pour la trouver, nous faisons un graphique par pairs (**cf. annexe 1.7**) et observons une possible relation polynomiale d'ordre 2 entre LSTAT et MEDV. Si nous corrigeons notre modèle en ajoutant un ordre quadratique, le R^2 ajusté est légèrement meilleur (0.7965) et les résidus s'améliorent nettement. (**cf. annexe 1.8**)

En conclusion sur l'exploration de données

La valeur médiane du logement à Boston et ses environs est comprise entre 5 et 50K dollars. Sa distribution est croissante jusqu'à sa valeur moyenne 22.5K dollars puis décroit fortement à partir de 25K dollars.

La valeur du logement est influencée (p_value <5%):

- positivement principalement (coefficient estimé = 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 (coefficient estimé = -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^2 ajusté du modèle de regression sur jeu de données incomplet est de **0.7965 (R^2 de référence)**.
(**cf. annexe 1.8** : résultat de la regression linéaire après selection de variable)

Les variables INDUS et AGE ne sont pas significatives dans l'explication de la valeur du logement (p-value>5% **cf. annexe 1.4**), et LSTAT a une relation quadratique avec MEDV.

Le meilleur modèle sur dataset incomplet testé ici est la regression lineaire du prix médian du logement, MEDV, sur l'ensemble des variables explicatives, hormis INDUS et AGE et incluant LSTAT².

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

2.Inventaire des données manquantes

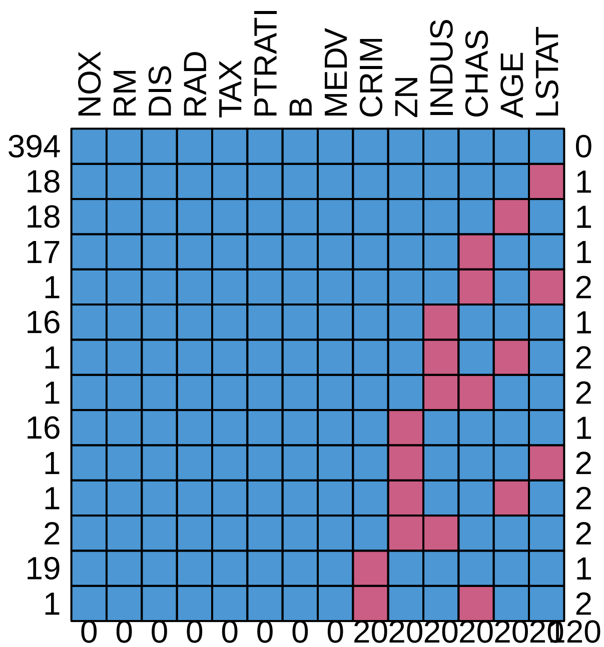
Il s'agit ici d'identifier les données manquantes par variable, représenter leur structure dans le jeu de données.

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 derniere colonne donne le nombre de variables incomplètes.



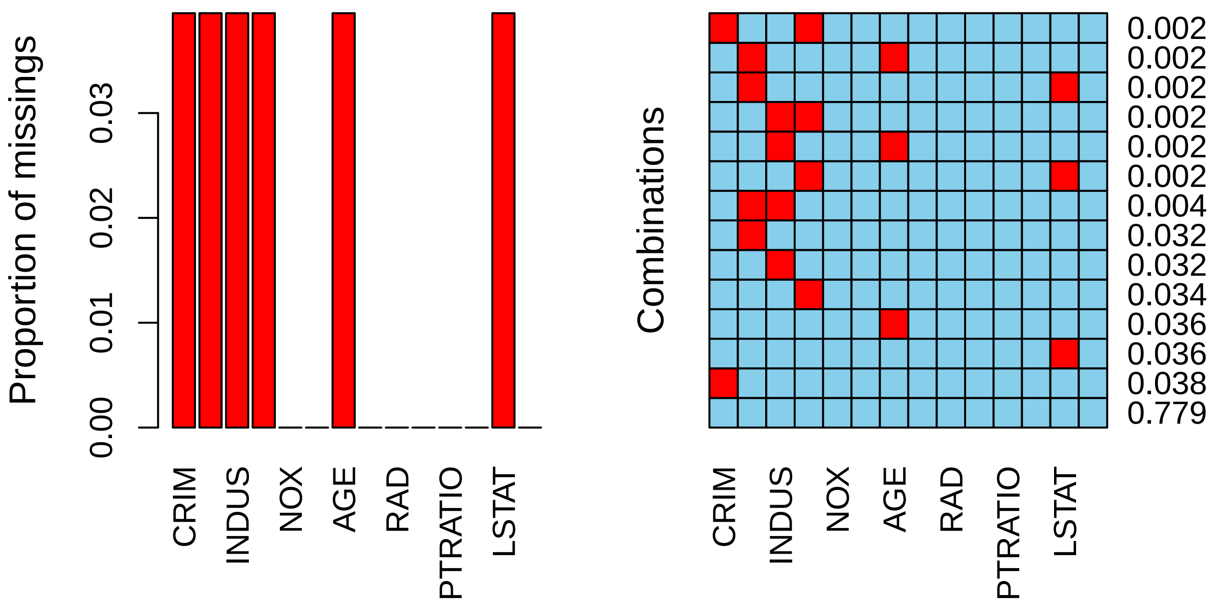
Au total, 120 observations sont manquantes : 20 pour chacune des variables CRIM, ZN, INDUS, CHAS, AGE, LSTAT.

- 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 dépendences des valeurs extrêmes et des données manquantes. Les données observées sont en bleu 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 raision 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, par rapport à notre exploration initial, nous pouvons 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 elle peuvent 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 détails 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 ces variables ne semblent pas pertinentes (cf.résultats de la partie 1).

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 pouvons fiver un m de taille moderé, en suivant les conseils de Stef Van Buuren dans son livre "flexible imputation data". Nous allons le fixer à 50, pour diminuer l'erreur dû à la simulation.

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
## -17.2946  -2.6367  -0.2356   1.8727  25.2018
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.119700   4.653524   8.836  < 2e-16 ***
## CRIM         -0.152360   0.029814  -5.110 4.61e-07 ***
## ZN            0.031747   0.012448   2.550 0.011064 *
## CHAS          3.014384   0.765785   3.936 9.46e-05 ***
## NOX         -13.460090   3.237709  -4.157 3.80e-05 ***
## RM            3.193761   0.376311   8.487 2.50e-16 ***
## DIS          -1.375621   0.168044  -8.186 2.32e-15 ***
## RAD           0.274140   0.057921   4.733 2.90e-06 ***
## TAX          -0.009799   0.003089  -3.173 0.001604 **
## PTRATIO      -0.788903   0.119001  -6.629 8.87e-11 ***
## B             0.008243   0.002444   3.373 0.000802 ***
## LSTAT        -1.517022   0.110083 -13.781  < 2e-16 ***
## I(LSTAT^2)    0.029246   0.002979   9.818  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.319 on 493 degrees of freedom
## Multiple R-squared:  0.7847, Adjusted R-squared:  0.7795
## F-statistic: 149.8 on 12 and 493 DF,  p-value: < 2.2e-16
```

Avec cette méthode, on observe un R^2 ajusté de **0.78** après la régression linéaire faite avec le dataset completé. Sa valeur est proche de celle obtenue pour 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 négatives en sachant que sur le dataset d'origine ces deux variables sont toujours positives.

Imputation par forêts aléatoires

Dans l'imputation par forêts aléatoires, les valeurs sont imputées en faisant des tirages aléatoires à partir des distributions gaussiennes independantes, centrées en les moyennes prédites par les forêts aléatoires. Pour ce faire , nous utilisons la fonction **mice (avec method = "rf")**, avec m=50

```
##
## 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
```

```
## -17.8191  -2.5753  -0.2496   1.9036  25.1538
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.042992   4.638382   9.064  < 2e-16 ***
## CRIM         -0.143591   0.029736  -4.829 1.84e-06 ***
## ZN            0.020152   0.012224   1.649 0.099883 .
## CHAS          2.811189   0.783758   3.587 0.000368 ***
## NOX         -13.554761   3.221807  -4.207 3.07e-05 ***
## RM            3.278444   0.373857   8.769  < 2e-16 ***
## DIS          -1.341981   0.166744  -8.048 6.33e-15 ***
## RAD           0.264960   0.057611   4.599 5.40e-06 ***
## TAX          -0.009474   0.003069  -3.087 0.002133 **
## PTRATIO      -0.791087   0.117771  -6.717 5.12e-11 ***
## B             0.008205   0.002430   3.376 0.000793 ***
## LSTAT        -1.717956   0.118245 -14.529  < 2e-16 ***
## I(LSTAT^2)    0.034908   0.003145  11.101  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.298 on 493 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.7816
## F-statistic: 151.6 on 12 and 493 DF,  p-value: < 2.2e-16
```

Avec cette méthode on observe un R^2 ajusté de **0.78** 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 prédiction sur les valeurs présentes de X, proches de la prédiction 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=50

```
##
## 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
## -17.453  -2.628  -0.251   1.871  25.092
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.588175   4.648709   9.161  < 2e-16 ***
## CRIM         -0.154235   0.029987  -5.143 3.90e-07 ***
## ZN            0.025804   0.012538   2.058 0.040101 *
```

```
## CHAS          2.662856   0.753478   3.534 0.000448 ***
## NOX         -13.563632   3.220018  -4.212 3.01e-05 ***
## RM            3.195735   0.374539   8.532  < 2e-16 ***
## DIS          -1.402110   0.169832  -8.256 1.39e-15 ***
## RAD           0.272979   0.057800   4.723 3.04e-06 ***
## TAX          -0.010061   0.003080  -3.267 0.001163 **
## PTRATIO      -0.778813   0.118053  -6.597 1.08e-10 ***
## B             0.007858   0.002437   3.224 0.001346 **
## LSTAT        -1.672289   0.118970 -14.056  < 2e-16 ***
## I(LSTAT^2)    0.033232   0.003188  10.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.3 on 493 degrees of freedom
## Multiple R-squared:  0.7866, Adjusted R-squared:  0.7814
## F-statistic: 151.4 on 12 and 493 DF,  p-value: < 2.2e-16
```

De même que pour les deux méthodes précédentes, le R^2 ajusté **0.78** est très proche du modèle avec des données manquantes. 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 mininimum, maximum et médianne sont presque idéntiques 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 l'ACP (Analyse des Composantes Principales).

Cette méthode fonctionne mieux quand la relation entre les variables continues est linéaire. Selon le pairplot que si trouve sur l'annexe 1.7, nous devrions transformer les variables CRIM et ZN, avant de faire l'imputation par ACP. Nous allons appliquer une tansformation avec la racine carré d'abord.

Ensuite, il s'agit d'abord d'estimer le nombre de composantes utilisées 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 est 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")**, avec nboot = 50 (m)

Régression linéaire sur le jeu de données complété par MIPCA

```
##                  estimate   std.error  statistic       df      p.value
## (Intercept)  41.268647323 4.831571080   8.541455 451.5238 2.220446e-16
## CRIM         -0.148237237 0.030186132  -4.910773 486.6777 1.240620e-06
## ZN            0.025199719 0.012693325   1.985273 459.3424 4.770833e-02
## CHAS          2.695275749 0.808709820   3.332809 444.7548 9.315613e-04
## NOX         -14.014037686 3.306919328  -4.237792 473.2182 2.714530e-05
## RM            3.308246616 0.388341096   8.518920 452.4183 2.220446e-16
## DIS          -1.355078182 0.173442939  -7.812818 434.9249 4.241052e-14
## RAD           0.262872872 0.059430792   4.423176 463.5434 1.213266e-05
## TAX          -0.009697194 0.003173840  -3.055350 459.2885 2.378836e-03
## PTRATIO      -0.800221148 0.121979516  -6.560291 468.9320 1.422866e-10
## B             0.008255401 0.002471537   3.340190 486.9780 9.014941e-04
## LSTAT        -1.562106163 0.129328514 -12.078590 287.1240 0.000000e+00
## I(LSTAT^2)    0.030677617 0.003436005   8.928281 318.5775 0.000000e+00
```

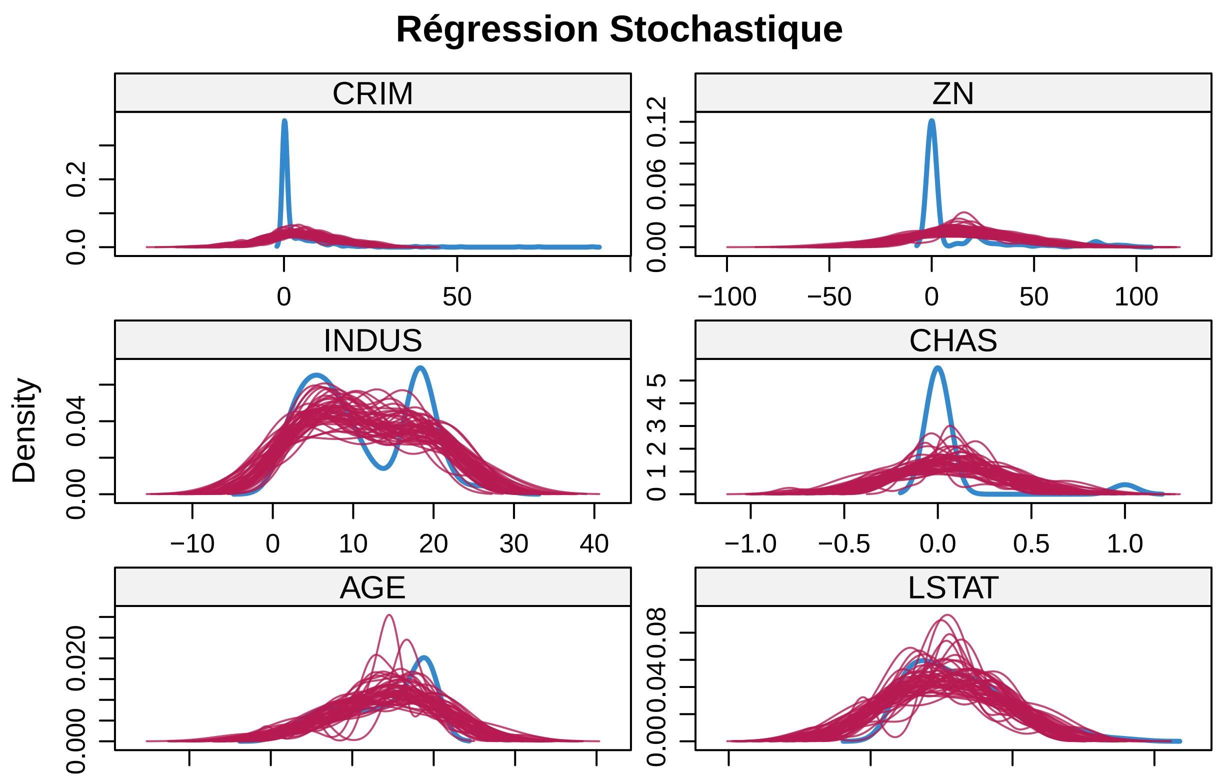
Les valeurs des estimateurs sont très proches de celles trouvées par les autres méthodes d'imputation multiple.

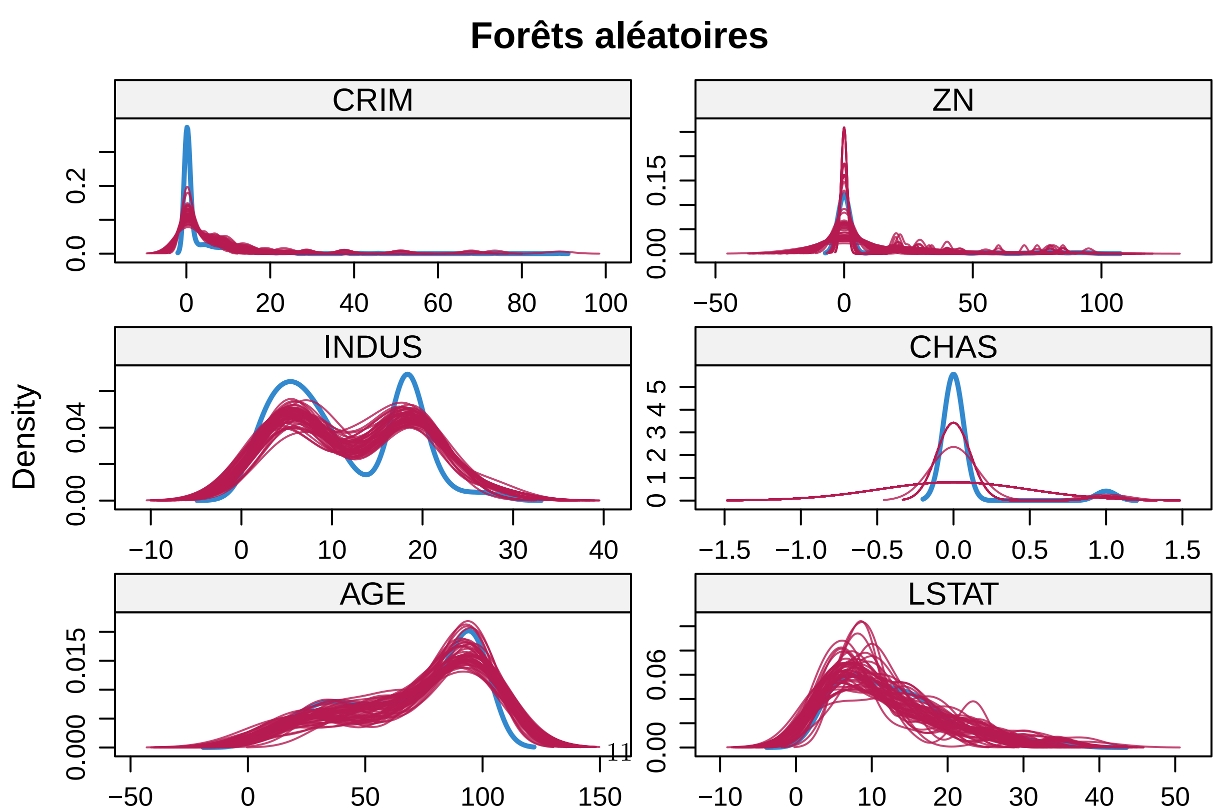
De même que pour les autres, le R^2 ajusté **0.78** est très proche du modèle avec des données manquantes. Les valeurs estimées et les résidus restent aussi très proches.

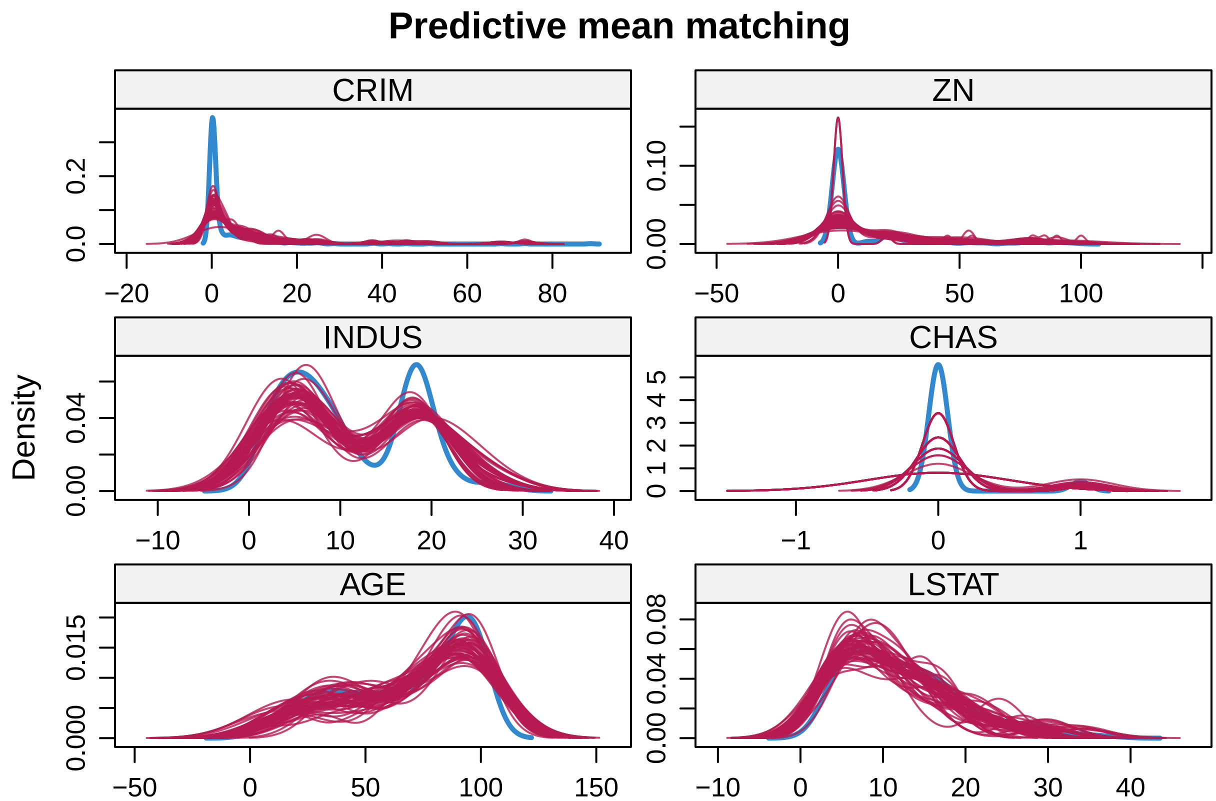
Nous avons testé 4 méthodes d'imputation, qui nous ont permis de faire des régressions avec le jeu de données complet. Ces régressions ont des résultats très proches en terme de R^2 ajusté, d'estimateurs et de résidus. Nous allons faire des diagnostics pour choisir le meilleur modèle d'imputation.

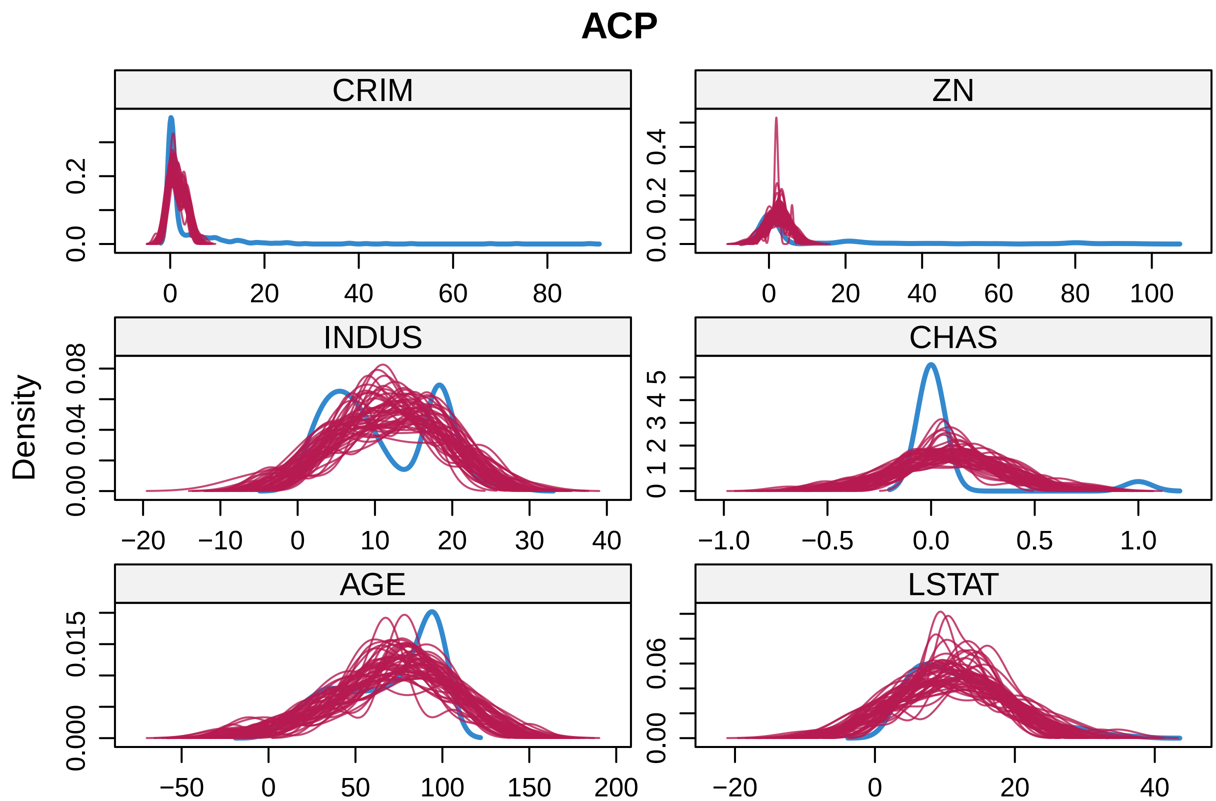
4. Diagnostics et conclusion

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






Predictive mean matching
CRIM
ZN
INDUS
CHAS
AGE
LSTAT
Density



Un analyse par variable: * Pour la variable CRIM, aucune méthode ne semble parvenir à reproduire la même distribution, mais les forêts aléatoires et pmm s'approchent plus. L'ACP arrive à bien le faire pour le début de la distribution, mais ne le fait pas pour la queue. * Pour INDUS, AGE et LSTAT, toutes les méthodes semblent réussir à le faire, mais les deux bosses sont mieux réproduites par forêts aléatoires et pmm * Pour CHAS et ZN, les forêts aléatoires sont les plus proches de la vraie distribution. Pour CHAS une imputation comme variable catégorielle pourrait changer le résultats.

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**)

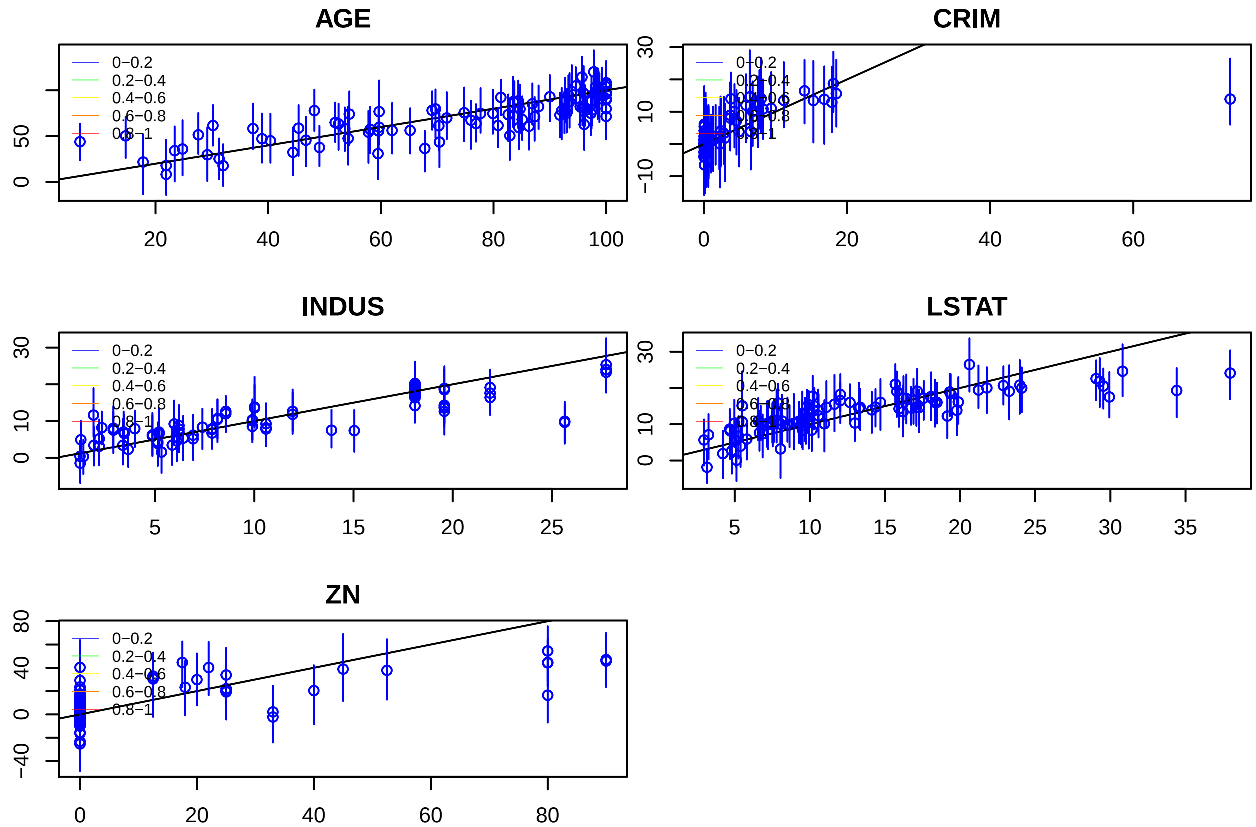
Diagnostic 3 : vérifier l'ajustement du modèle d'imputation

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

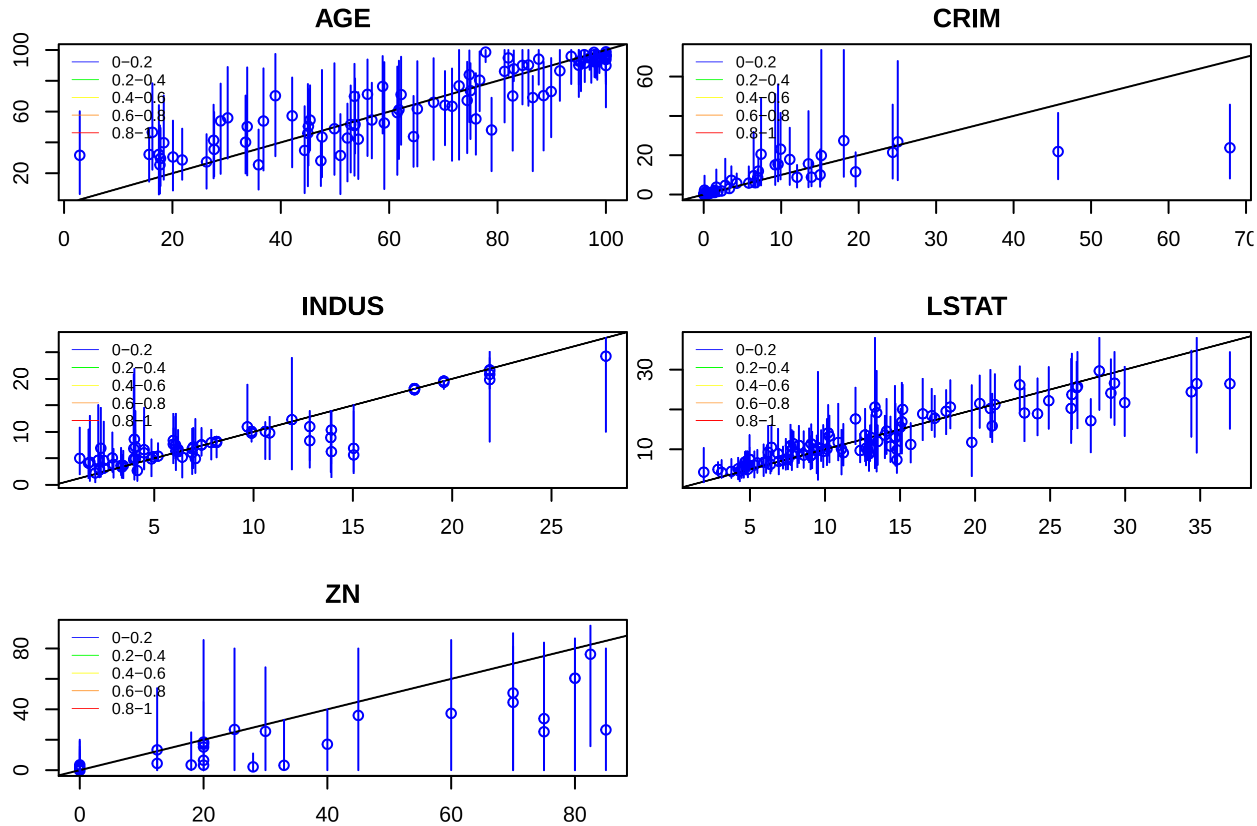
Sur ces graphiques, la 1ère bissectrice (y=x) représente l'imputation parfaite. La qualité de l'imputation se mesure en observant la proximité des intervalles de confiance avec cette droite. On espère que 90% des intervalles traversent la 1ere bissectrice. La couleur des intervalles represente la fraction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

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

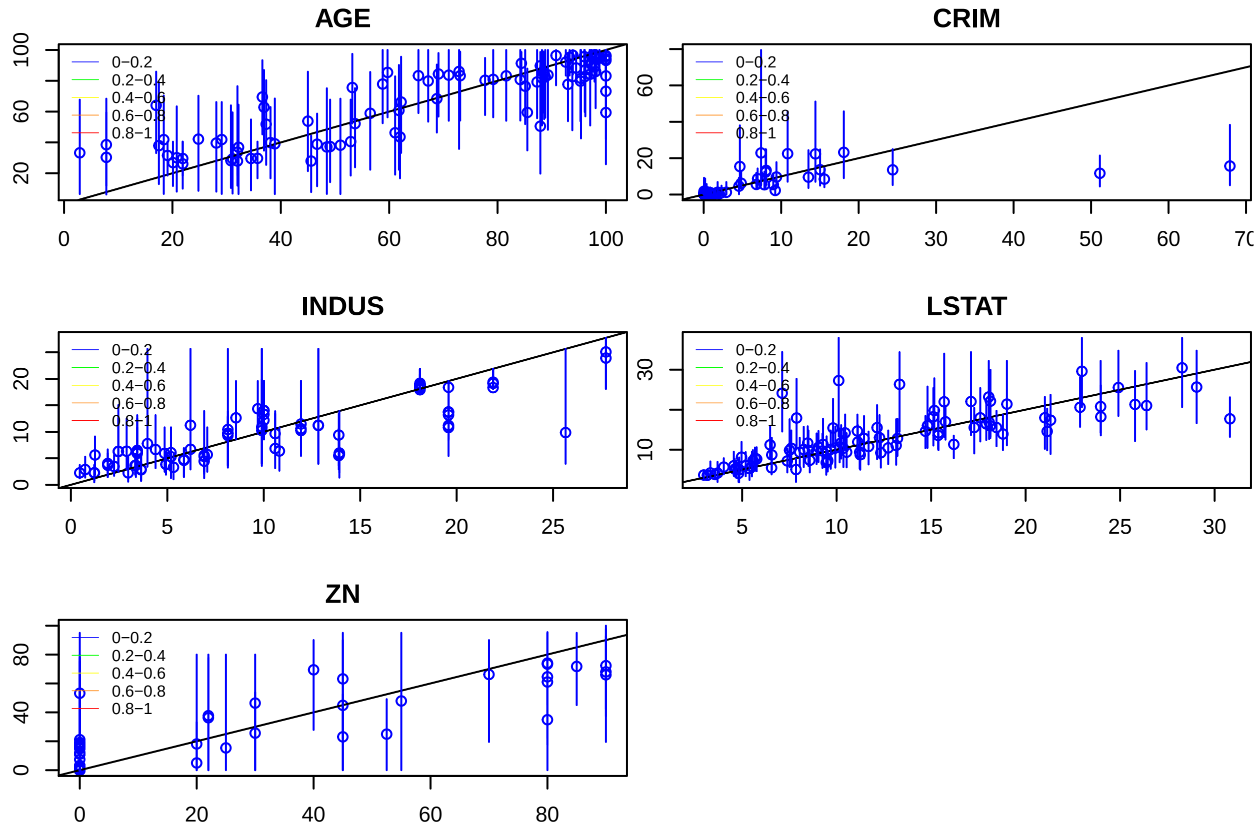
Ajustement de la regression stochastique :



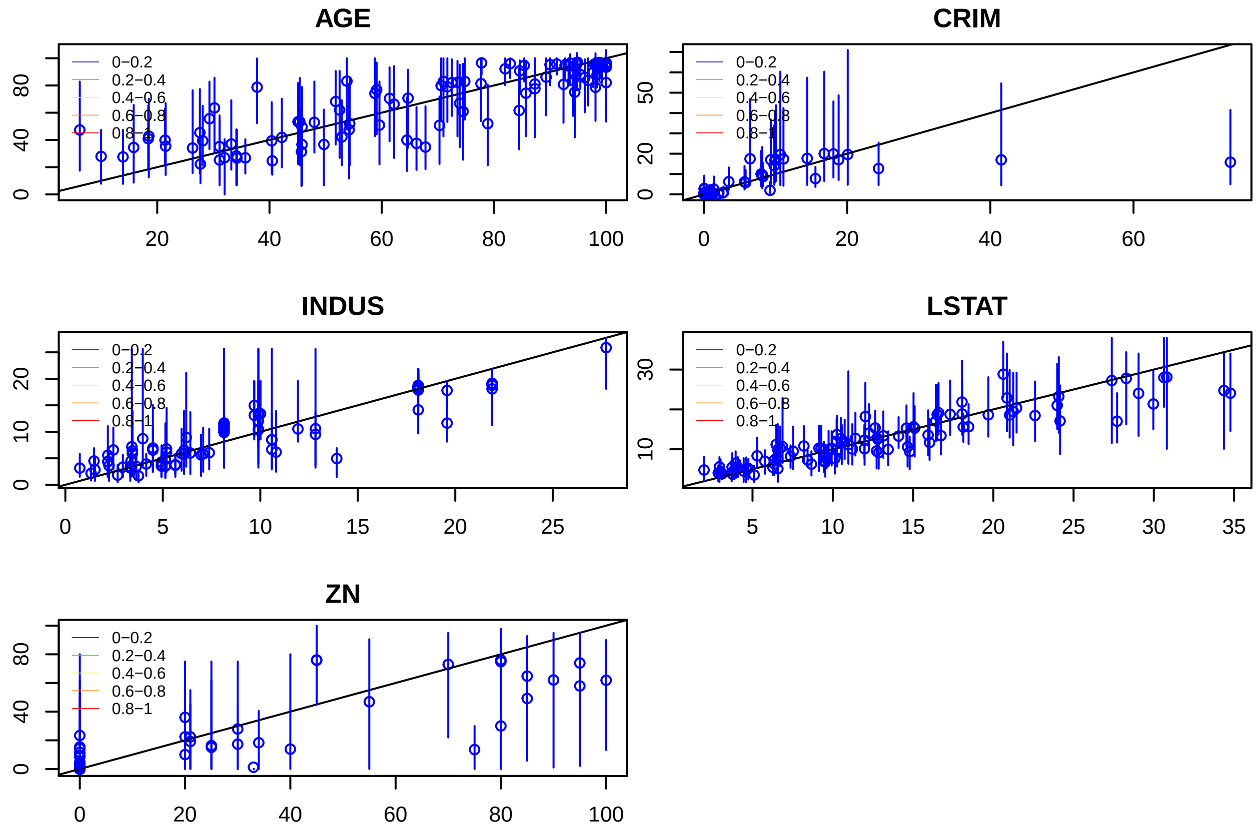
Ajustement des forêts aléatoires :



Ajustement de la predictive mean matching :



Ajustement de l'ACP :



On observe que, quelque soit la méthode d'imputation, la plupart des intervalles de confiance des variables INDUS, AGE et LSTAT coupent la 1ere bissectrice des méthodes : ces variables sont donc correctement imputées sur leurs données manquantes. Par contre, l'imputation des données manquantes sur CRIM et ZN est moins satisfaisante : particulièrement sur les méthodes de regression stochastique, de forêts aléatoires et PMM.

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 compléter de façon adéquate les variables CRIM et ZN (cf. overimputation).

Si on compare les méthodes d'imputation en terme de rapidité machine, la méthode APC est celle qui demande plus de temps, suivi de près par les forêts aléatoires.

La qualité de la régression linéaire sur le jeu de données complété est très proche quelques soient les méthodes d'imputation utilisées : les R^2 ajustés sont quasiment les mêmes (0.78) et les distributions des résidus oscillent autour de 0 (cf.**annexe 4.2**).

Résumé des performances des méthodes d'imputation

Critères	Reg.stochastique	Random forest	PMM	ACP
Diag.1 : distribution	3/6	5/6	5/6	3/6
Diag.2 : convergence	OK	OK	OK	OK

Critères	Reg.stochastique	Random forest	PMM	ACP
Diag.3 : ajustement	3/5	4/5	4/5	5/5
Temps machine (sec)	4.37	33.24	5.25	48.76
Modelisation de MEDV	R^2 ajusté = 0.78 résidus proches 0	R^2 ajusté = 0.78 résidus proches 0	R^2 ajusté = 0.78 résidus proches 0	R^2 ajusté = 0.78 résidus proches 0

D'après ces éléments de performance, les imputations par PMM donnent les meilleurs résultats pour notre cas.

Sur le dataset étudié, nous obtenons des résultats proches pour la régression linéaire de MEDV, avec ou sans imputation des données manquantes. Cela peut s'expliquer par :

- la petite dimension du jeu de données.

- Un faible pourcentage des données manquantes.
- le modèle de regression linéaire choisi qui n'inclut pas 2 des variables avec données manquantes (INDUS et AGE).

Annexes

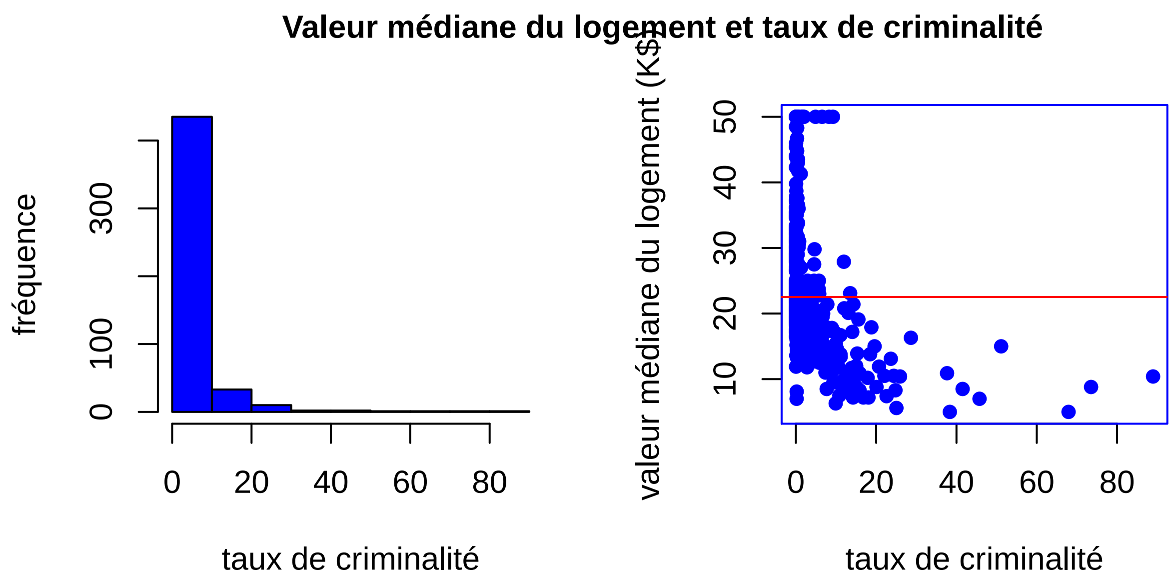
Annexe 1.1 : structure du jeu de données d'origine

```
##       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
##
##      LSTAT             MEDV
##  Min.   : 1.730   Min.   : 5.00
##  1st Qu.: 7.125   1st Qu.:17.02
```

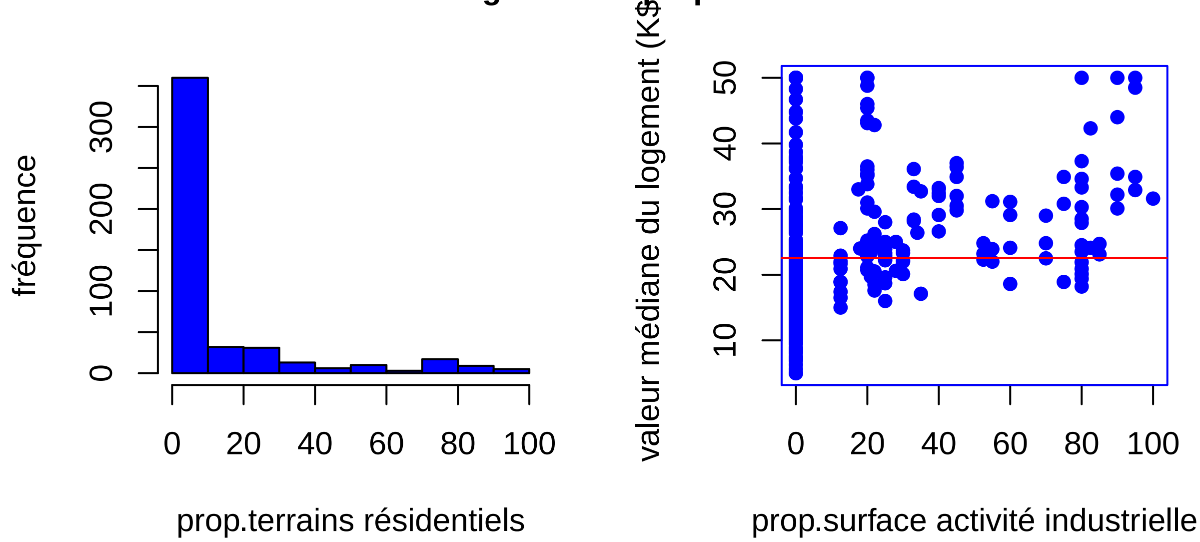
```
##   Median :11.430   Median :21.20
##   Mean   :12.715   Mean   :22.53
##   3rd Qu.:16.955   3rd Qu.:25.00
##   Max.   :37.970   Max.   :50.00
##   NA's   :20
```

Annexe 1.2 : graphiques exploratoires du jeu de données

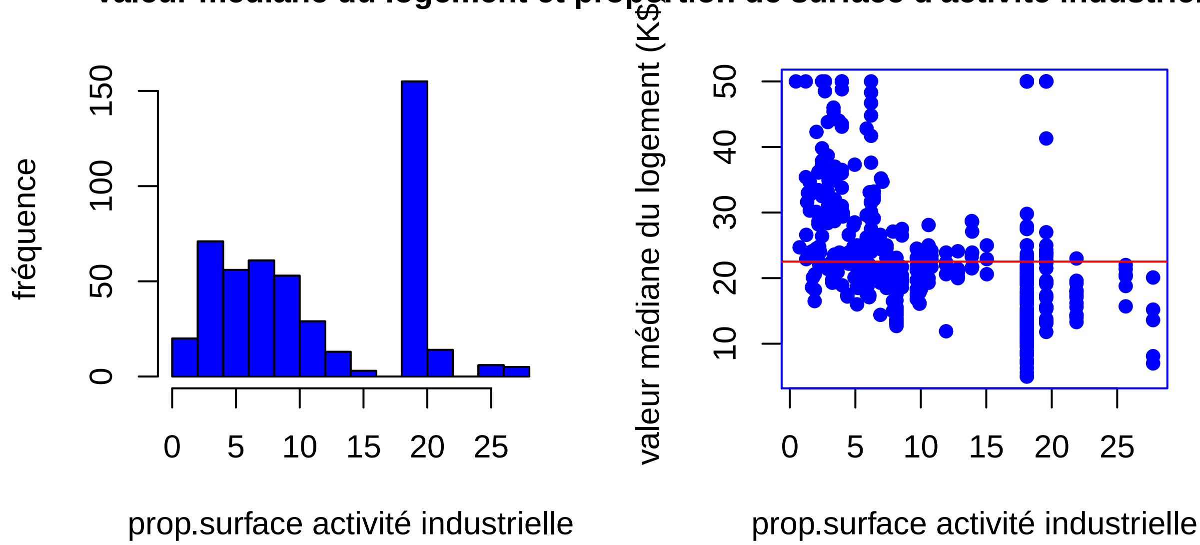
Les répresentations graphiques donnent, pour chaque variable explicative, la distribution sous forme d'histogramme et le nuage de points de la variable vs la valeur médiane du logement (en rouge la valeur moyenne = 22.5K$).

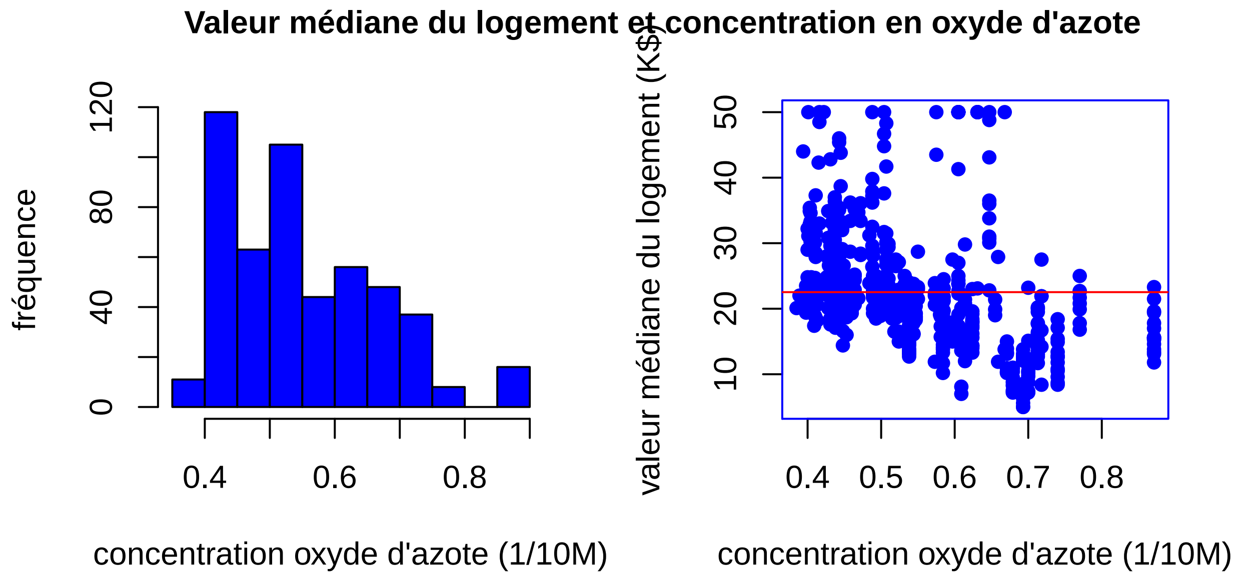


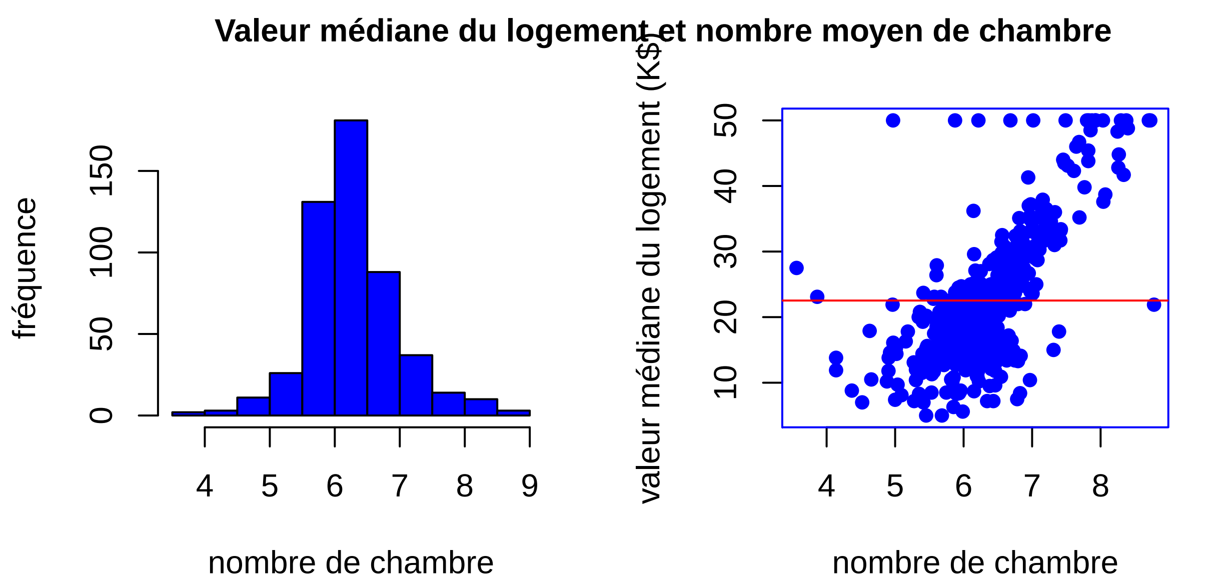
Valeur médiane du logement et proportion de terrains résidentiels

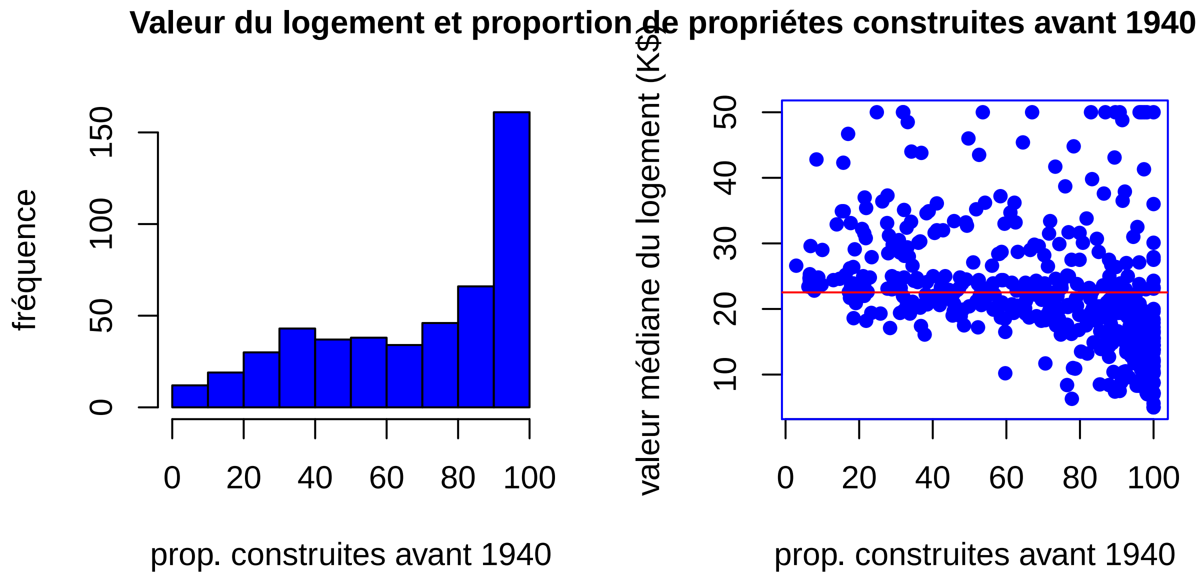


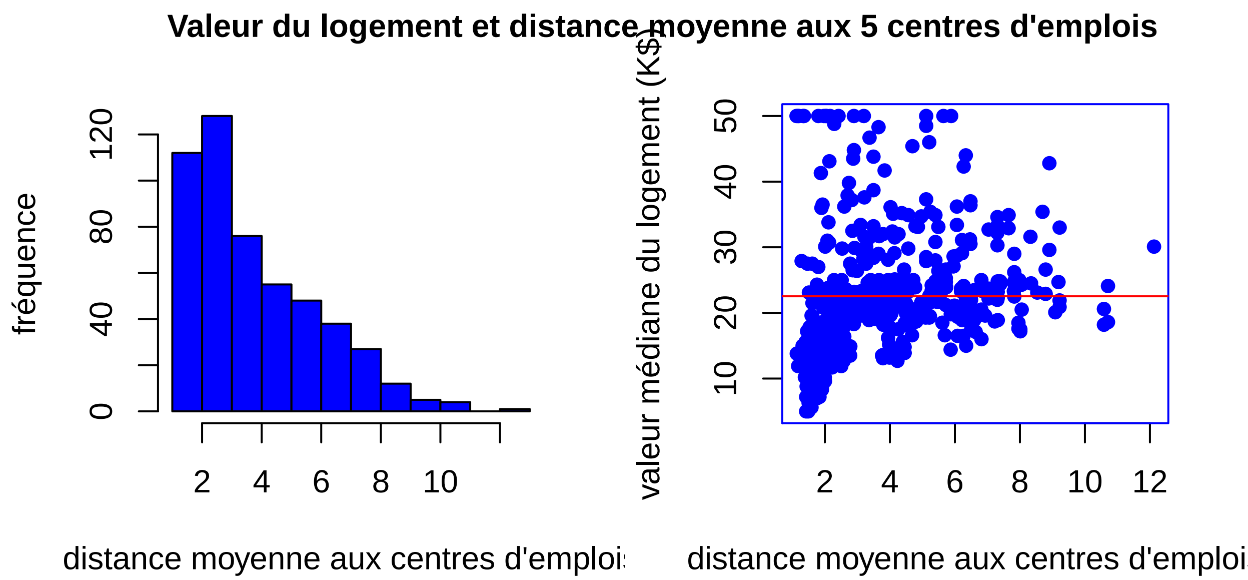
Valeur médiane du logement et proportion de surface d'activité industrielle

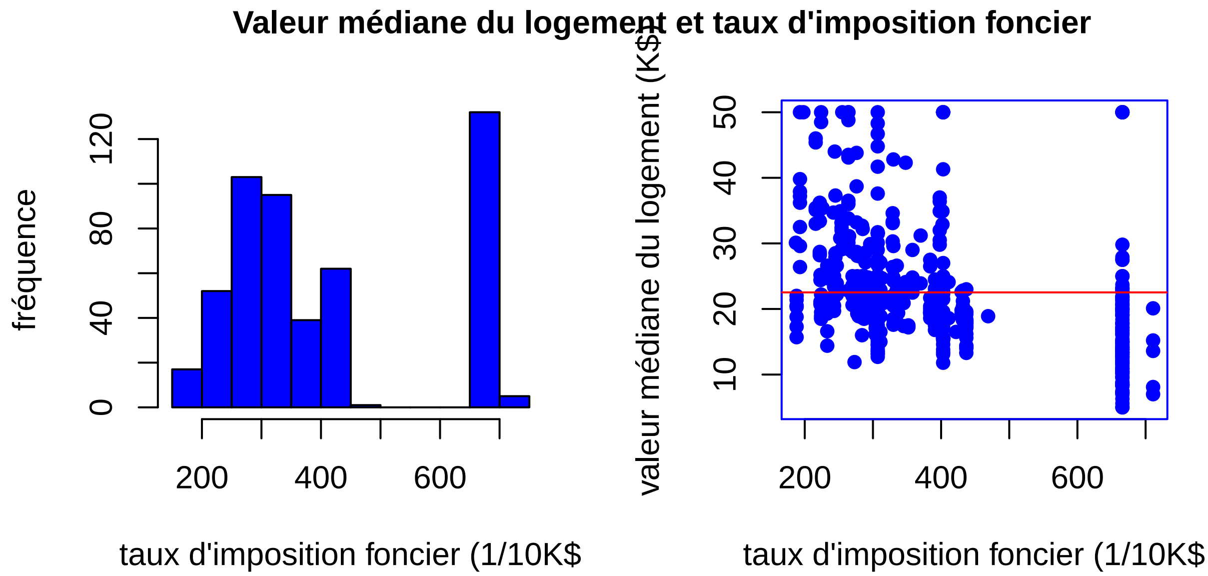


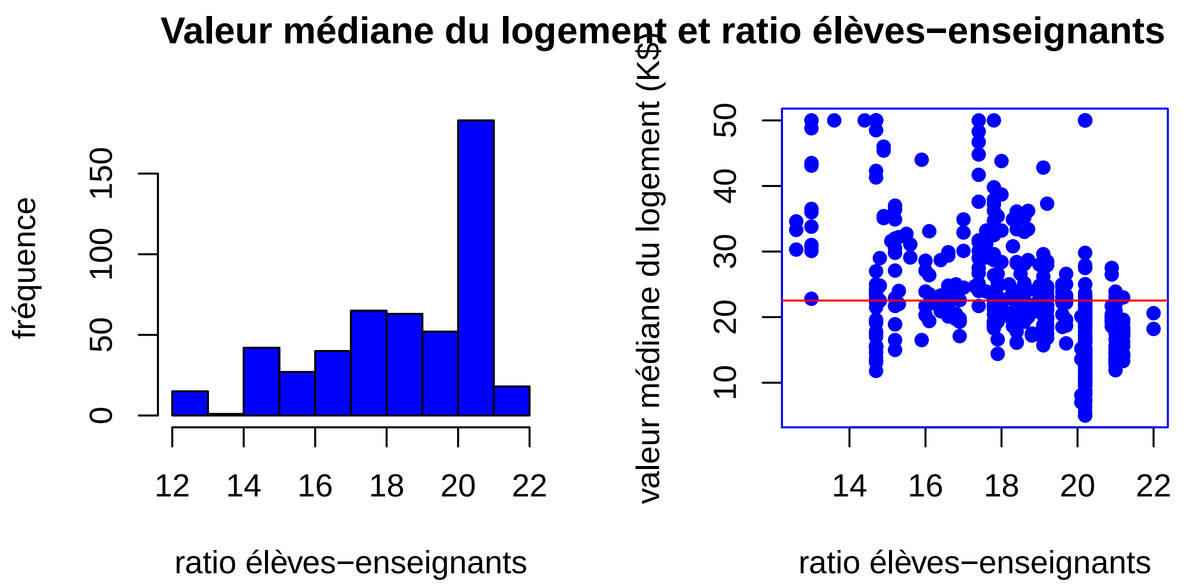

Valeur médiane du logement et concentration en oxyde d'azote
fréquence
0
40
80
120
0.4
0.6
0.8
concentration oxyde d'azote (1/10M)
valeur médiane du logement (K$)
10
20
30
40
50
0.4
0.5
0.6
0.7
0.8
concentration oxyde d'azote (1/10M)

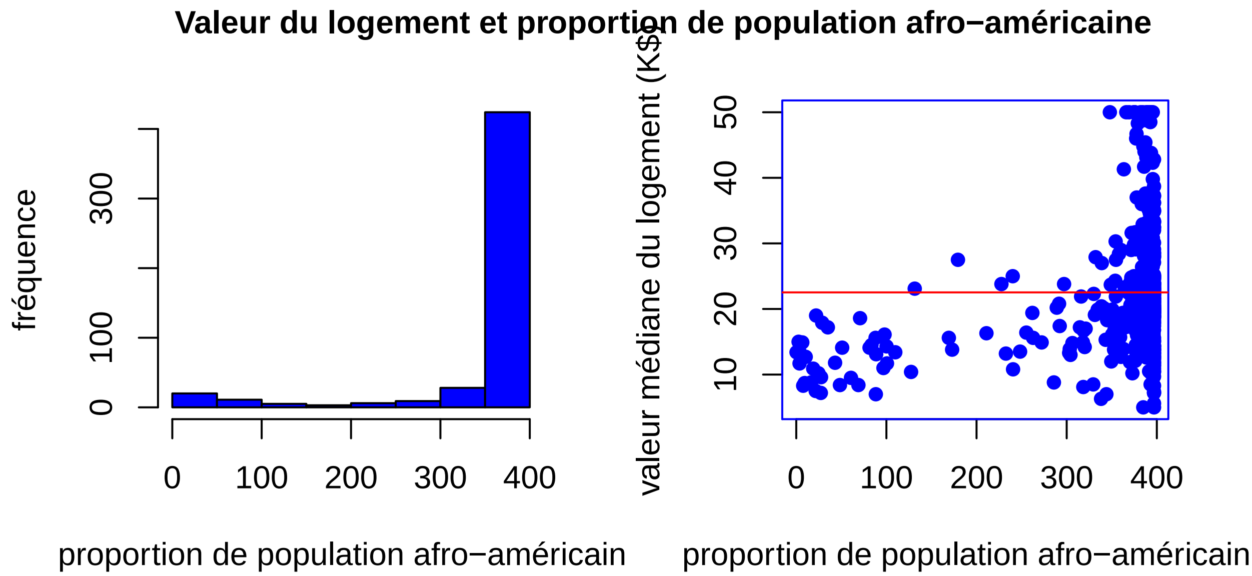

Valeur médiane du logement et nombre moyen de chambre
fréquence
0
50
100
150
4
5
6
7
8
9
nombre de chambre
valeur médiane du logement (K$)
10
20
30
40
50
4
5
6
7
8
nombre de chambre

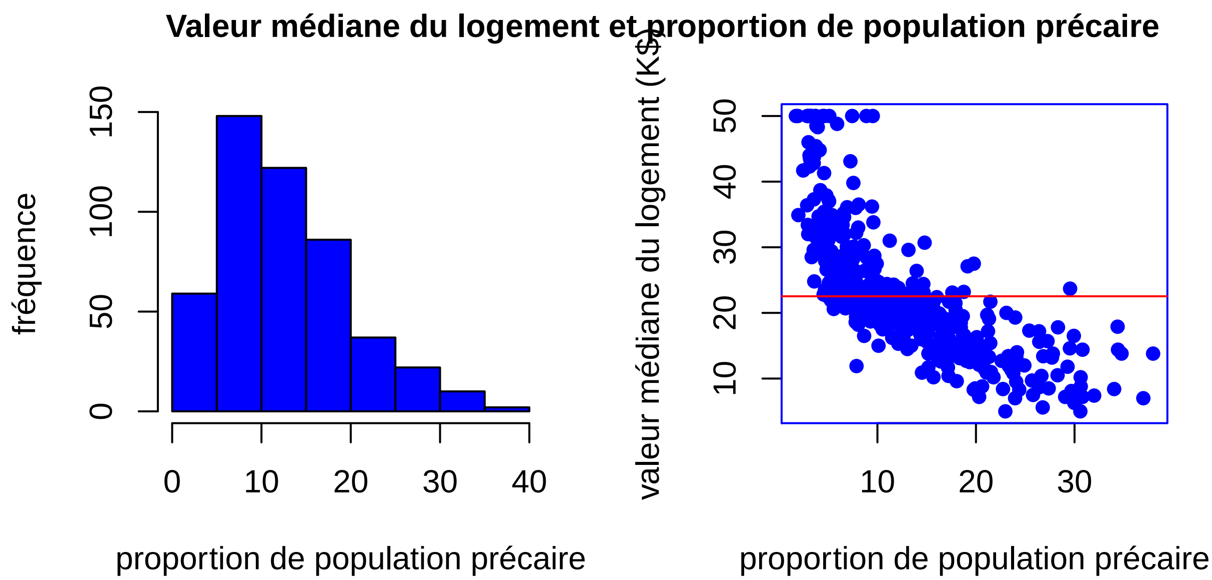

Valeur du logement et proportion de propriétes construites avant 1940
fréquence
prop. construites avant 1940
valeur médiane du logement (K$)
prop. construites avant 1940

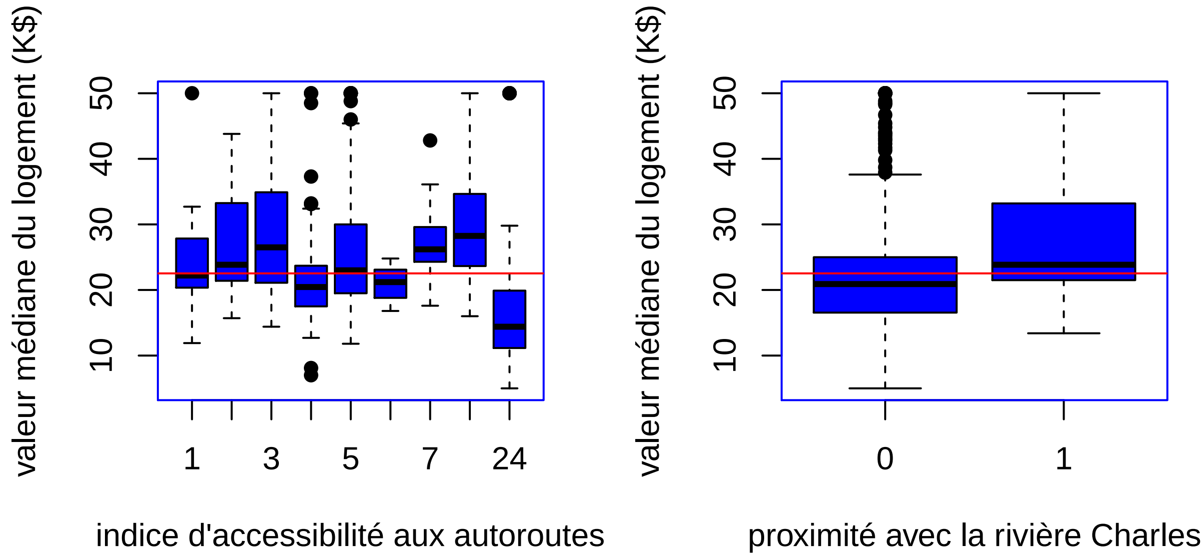

Valeur du logement et distance moyenne aux 5 centres d'emplois
fréquence
distance moyenne aux centres d'emploi
valeur médiane du logement (K$)
distance moyenne aux centres d'emploi


Valeur médiane du logement et taux d'imposition foncier
fréquence
0
40
80
120
200
400
600
taux d'imposition foncier (1/10K$
valeur médiane du logement (K$)
10
20
30
40
50
200
400
600
taux d'imposition foncier (1/10K$


Valeur médiane du logement et ratio élèves−enseignants
fréquence
0
50
100
150
12
14
16
18
20
22
ratio élèves−enseignants
valeur médiane du logement (K$)
10
20
30
40
50
14
16
18
20
22
ratio élèves−enseignants


Valeur du logement et proportion de population afro−américaine
fréquence
0
100
300
0
100
200
300
400
proportion de population afro−américain
valeur médiane du logement (K$)
10
20
30
40
50
0
100
200
300
400
proportion de population afro−américain


Valeur médiane du logement et proportion de population précaire
fréquence
0
50
100
150
0
10
20
30
40
proportion de population précaire
valeur médiane du logement (K$)
10
20
30
40
50
10
20
30
proportion de population précaire



Annexe 1.3 : 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
CRIM	1	-0.19	0.39	-0.05	0.42	-0.23	0.34	-0.37	0.61	0.56	0.27	-0.39	0.46
ZN	-0.19	1	-0.52	-0.03	-0.52	0.34	-0.57	0.65	-0.3	-0.31	-0.42	0.17	-0.42
INDUS	0.39	-0.52	1	0.05	0.76	-0.4	0.64	-0.7	0.59	0.73	0.4	-0.34	0.6
CHAS	-0.05	-0.03	0.05	1	0.08	0.1	0.07	-0.1	0.01	-0.03	-0.1	0.07	-0.04
NOX	0.42	-0.52	0.76	0.08	1	-0.32	0.73	-0.77	0.63	0.68	0.21	-0.38	0.59
RM	-0.23	0.34	-0.4	0.1	-0.32	1	-0.25	0.22	-0.24	-0.32	-0.39	0.12	-0.64
AGE	0.34	-0.57	0.64	0.07	0.73	-0.25	1	-0.75	0.44	0.5	0.26	-0.28	0.6
DIS	-0.37	0.65	-0.7	-0.1	-0.77	0.22	-0.75	1	-0.48	-0.53	-0.23	0.29	-0.51
RAD	0.61	-0.3	0.59	0.01	0.63	-0.24	0.44	-0.48	1	0.9	0.44	-0.44	0.51
TAX	0.56	-0.31	0.73	-0.03	0.68	-0.32	0.5	-0.53	0.9	1	0.45	-0.44	0.57
PTRATIO	0.27	-0.42	0.4	-0.1	0.21	-0.39	0.26	-0.23	0.44	0.45	1	-0.18	0.4
B	-0.39	0.17	-0.34	0.07	-0.38	0.12	-0.28	0.29	-0.44	-0.44	-0.18	1	-0.38
LSTAT	0.46	-0.42	0.6	-0.04	0.59	-0.64	0.6	-0.51	0.51	0.57	0.4	-0.38	1
MEDV	-0.4	0.41	-0.51	0.17	-0.46	0.72	-0.41	0.28	-0.42	-0.51	-0.54	0.35	-0.74

Annexe 1.4 : Jeu de données incomplet - regression lineaire 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
```

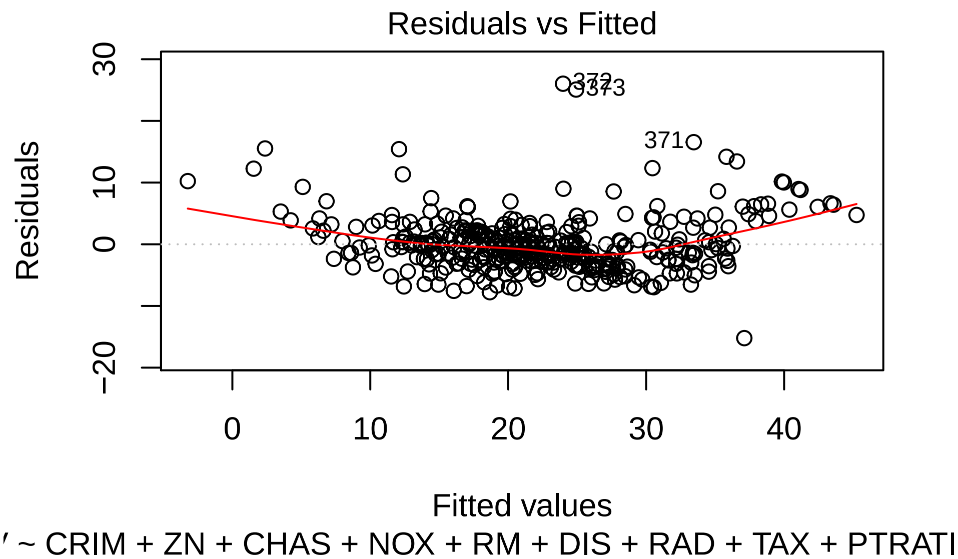
```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.680059   5.681290   5.752 1.81e-08 ***
## CRIM         -0.097594   0.032457  -3.007 0.002815 **
## ZN            0.048905   0.014398   3.397 0.000754 ***
## INDUS         0.030379   0.065933   0.461 0.645237
## 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 1.5 : Jeu de données incomplet -regression linéaire MEDV sur variables choisies par stepAIC

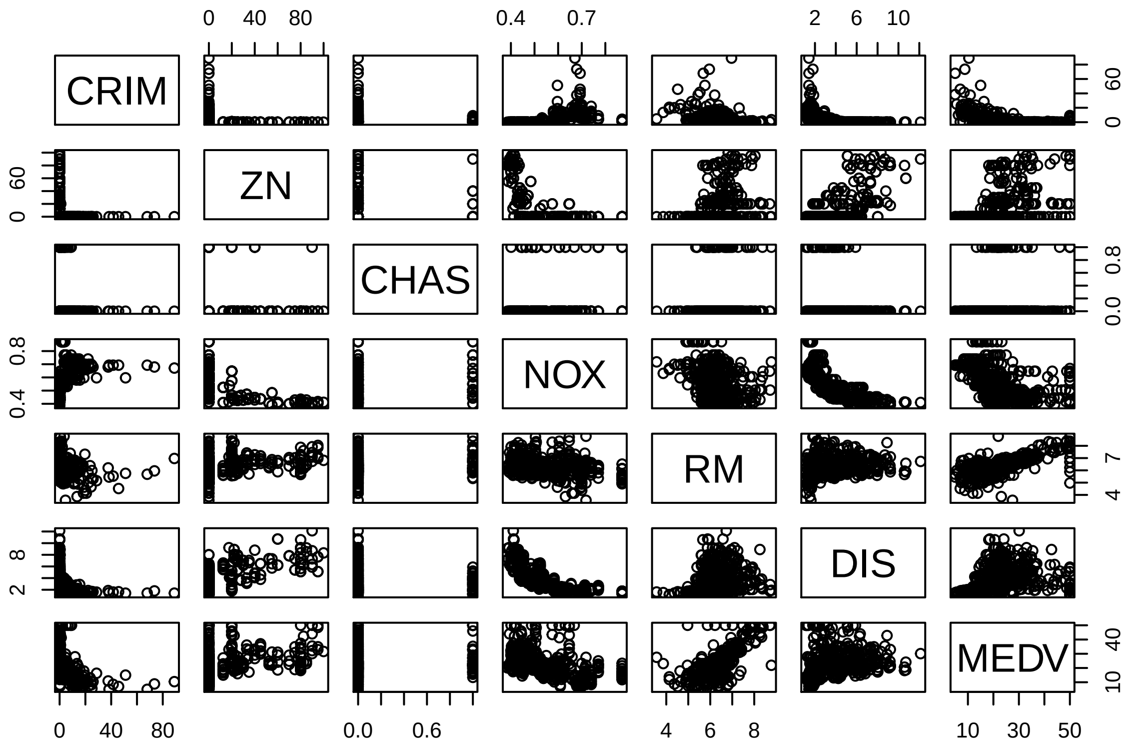
```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = dHB2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.214  -2.552  -0.503   1.768  26.027
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (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
```

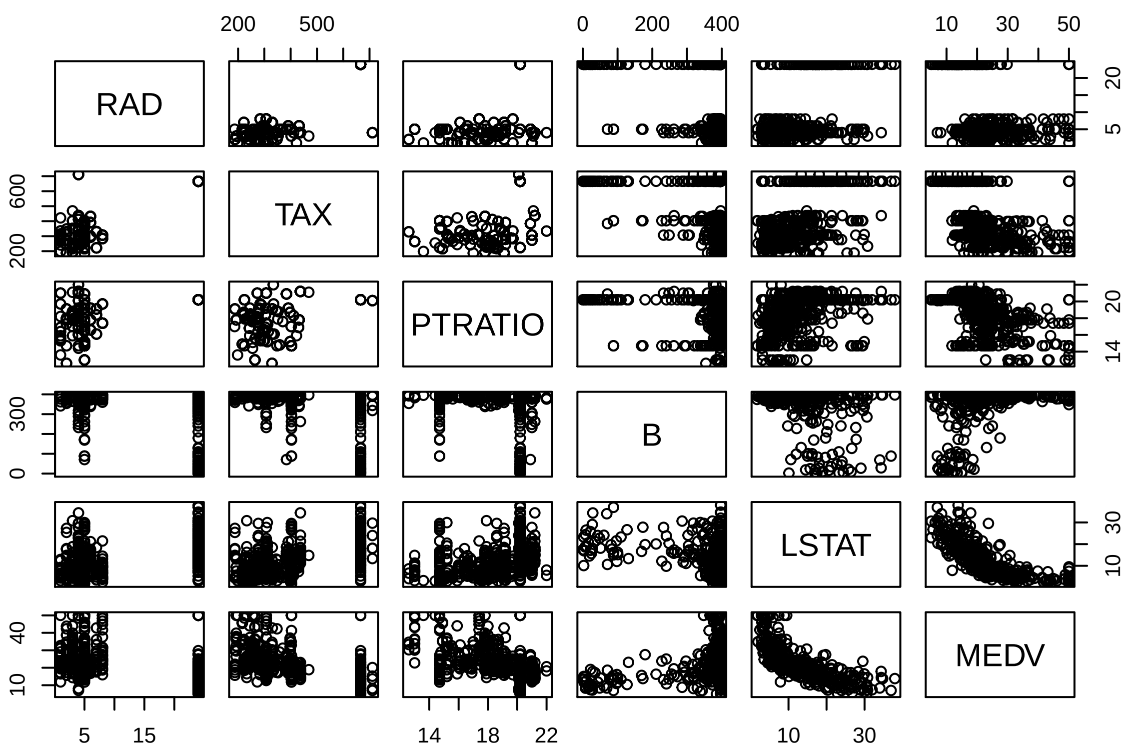
```
##
## Residual standard error: 4.481 on 382 degrees of freedom
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.7598
## F-statistic:   114 on 11 and 382 DF,  p-value: < 2.2e-16
```

Annexe 1.6 : Graphique des résidus du modèle



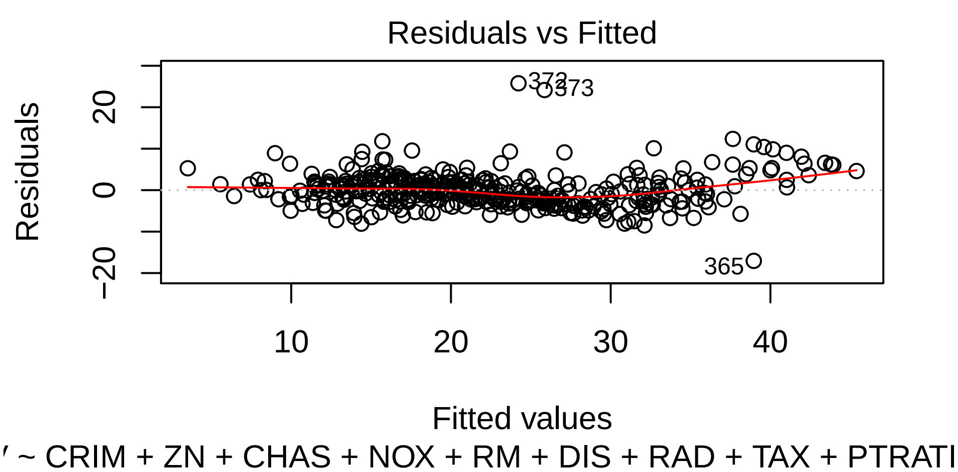
Annexe 1.7 : pairs plot



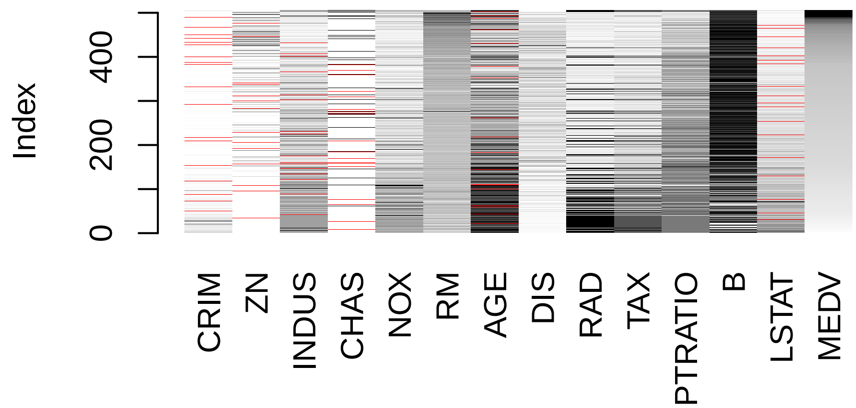


Annexe 1.8 : Jeu de données incomplet - regression linéaire apres sélection de variables et incluant relation ordre 2 avec LSTAT

```
##
## 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'algorithm MICE a besoin d'avoir une méthode univariée d'imputation pour chaque variable incomplète.

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

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

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

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 predicteurs. 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 répresente une thématique différente, utile pour l'estimation de la valeur de la maison.

Etape 5 - Définir l'ordre d'imputation des variables (influe sur la convergence de l'algorithme). Par défaut, algorithme MICE impute 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.   :-9.45933   Min.   : -6.424   Min.   : 0.46   Min.   :-0.43113
##  1st Qu.: 0.07987   1st Qu.:  0.000   1st Qu.: 5.19   1st Qu.: 0.00000
##  Median : 0.26042   Median :  0.000   Median : 9.69   Median : 0.00000
##  Mean   : 3.64137   Mean   : 11.663   Mean   :11.12   Mean   : 0.06972
##  3rd Qu.: 3.75547   3rd Qu.: 17.231   3rd Qu.:18.10   3rd Qu.: 0.00000
##  Max.   :88.97620   Max.   :100.000   Max.   :27.74   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   : 68.68   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :120.26   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.   :-7.331   Min.   : 5.00
##  1st Qu.: 7.037   1st Qu.:17.02
##  Median :11.395   Median :21.20
##  Mean   :12.639   Mean   :22.53
##  3rd Qu.:16.955   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
##
##      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

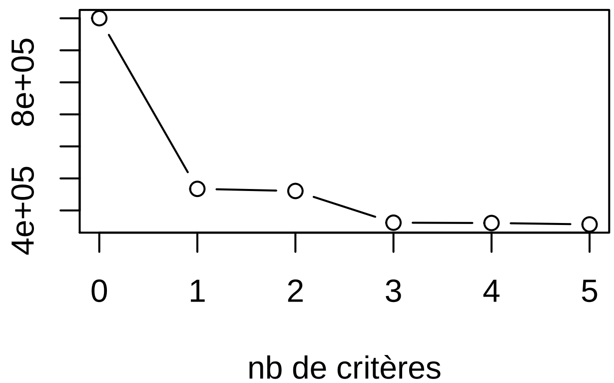
```
##       CRIM                ZN             INDUS            CHAS
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08057   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.62430   Mean   : 11.05   Mean   :11.15   Mean   :0.06719
##  3rd Qu.: 3.64742   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   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 : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.68   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
##       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.   : 1.730   Min.   : 5.00
## 1st Qu.: 7.125   1st Qu.:17.02
## Median :11.330   Median :21.20
## Mean   :12.697   Mean   :22.53
## 3rd Qu.:16.955   3rd Qu.:25.00
## Max.   :37.970   Max.   :50.00
```

Annexe 3.4 : Structure des données imputées par pmm

```
##      CRIM                ZN             INDUS           CHAS
## Min.   : 0.00632   Min.   :  0.0   Min.   : 0.46   Min.   :0.00000
## 1st Qu.: 0.08190   1st Qu.:  0.0   1st Qu.: 5.19   1st Qu.:0.00000
## Median : 0.25651   Median :  0.0   Median : 9.69   Median :0.00000
## Mean   : 3.60471   Mean   : 11.4   Mean   :11.11   Mean   :0.07312
## 3rd Qu.: 3.68939   3rd Qu.: 12.5   3rd Qu.:18.10   3rd Qu.:0.00000
## Max.   :88.97620   Max.   :100.0   Max.   :27.74   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.: 45.62   1st Qu.: 2.100
## Median :0.5380   Median :6.208   Median : 76.95   Median : 3.207
## Mean   :0.5547   Mean   :6.285   Mean   : 68.77   Mean   : 3.795
## 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   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.   : 1.730   Min.   : 5.00
## 1st Qu.: 7.125   1st Qu.:17.02
## Median :11.360   Median :21.20
## Mean   :12.635   Mean   :22.53
## 3rd Qu.:16.930   3rd Qu.:25.00
## Max.   :37.970   Max.   :50.00
```

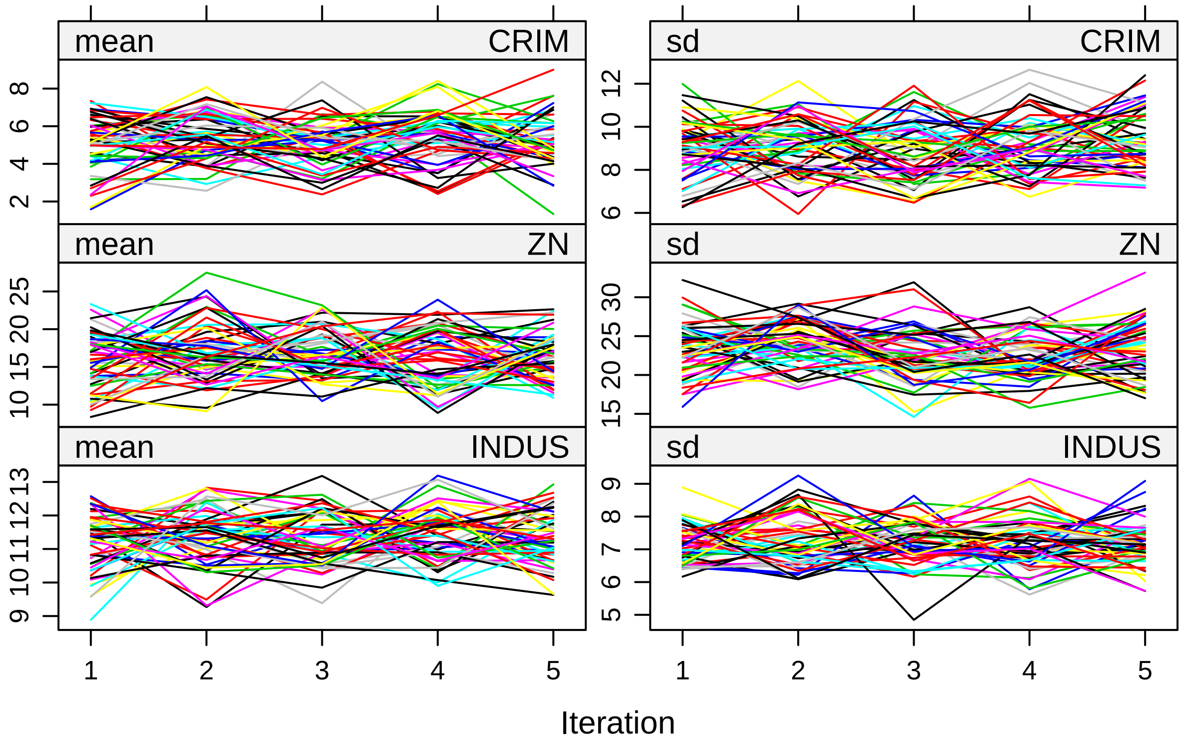
Annexe 3.5 : Imputation par ACP



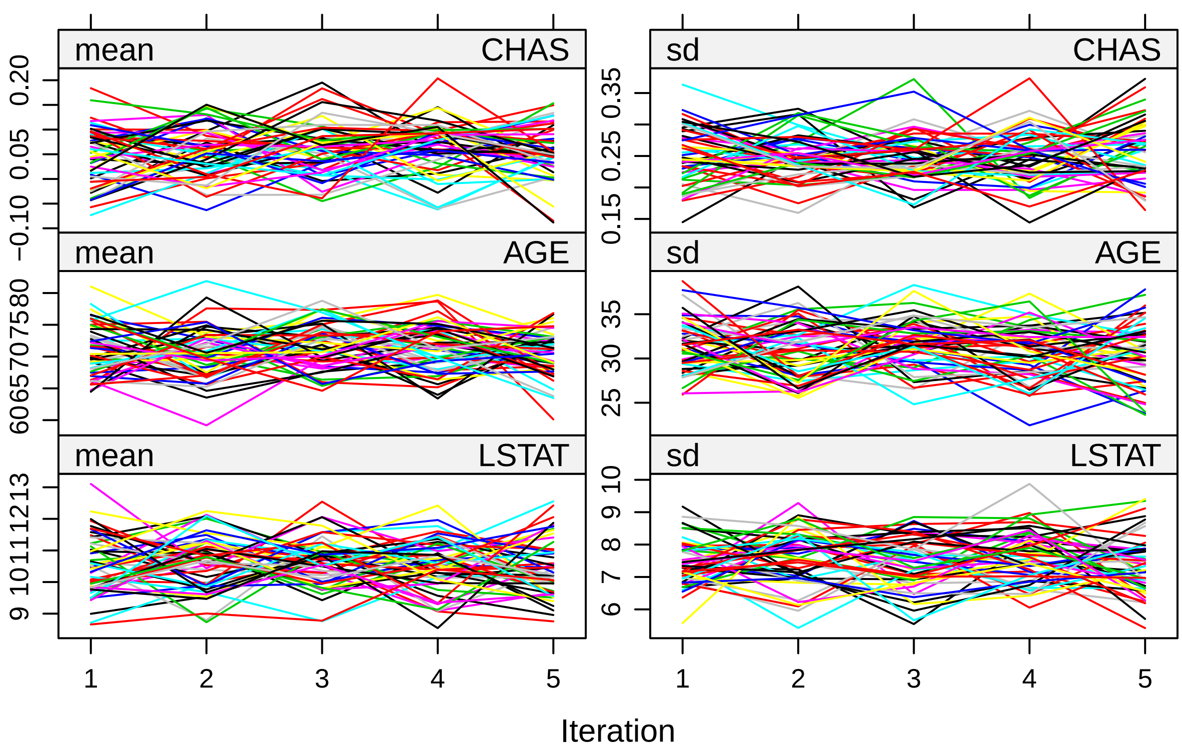
graphique des dimensions

Annexe 4.1 Diagnostic 2 : convergence des algorithmes.

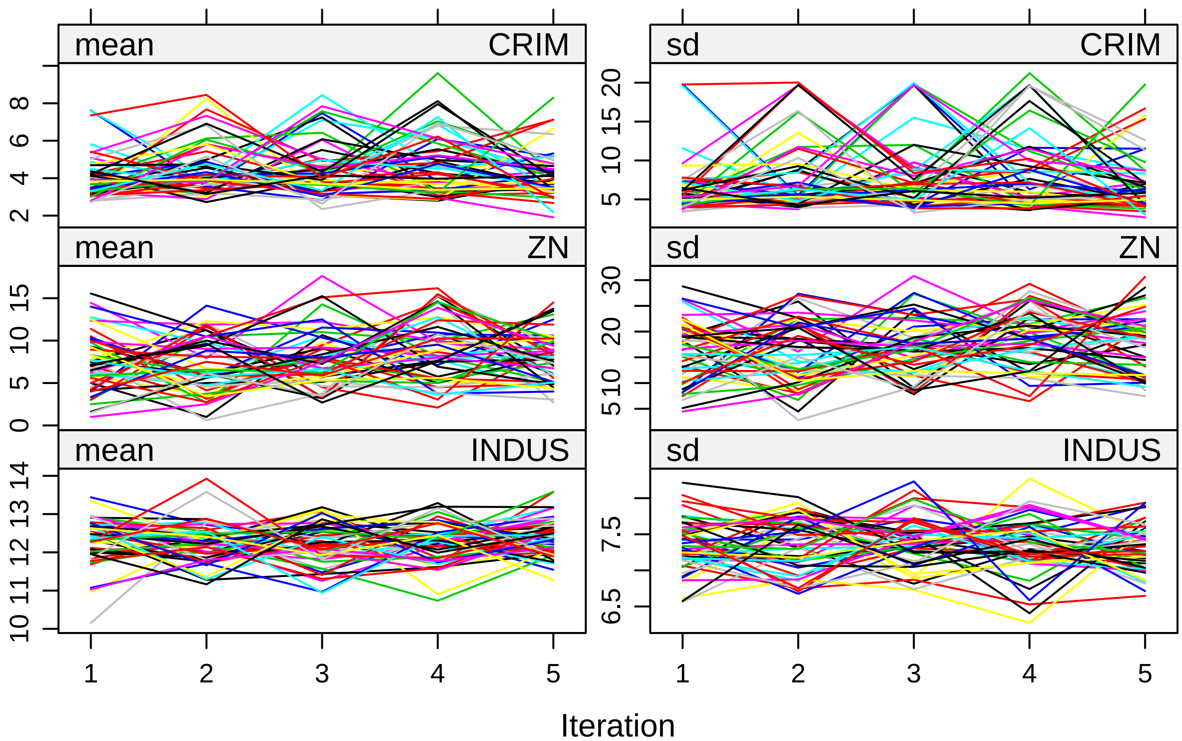
Régression Stochastique

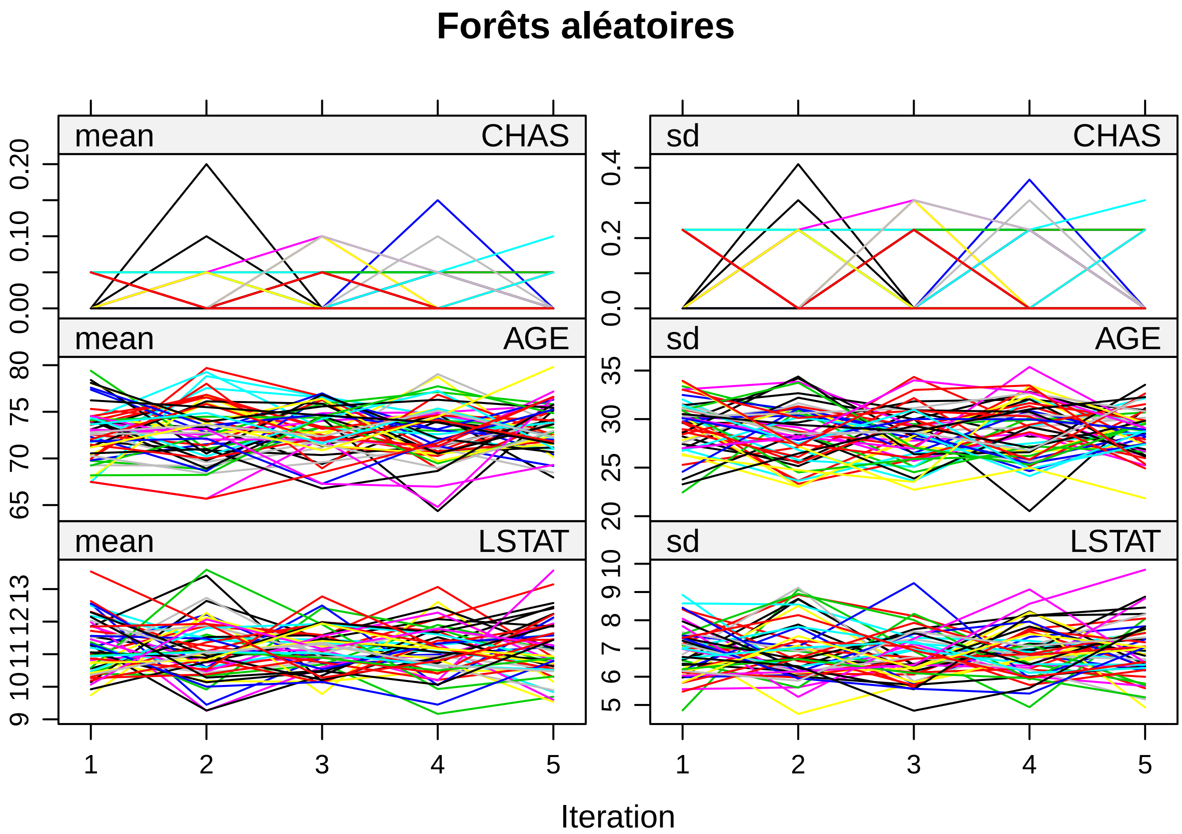


Régression Stochastique

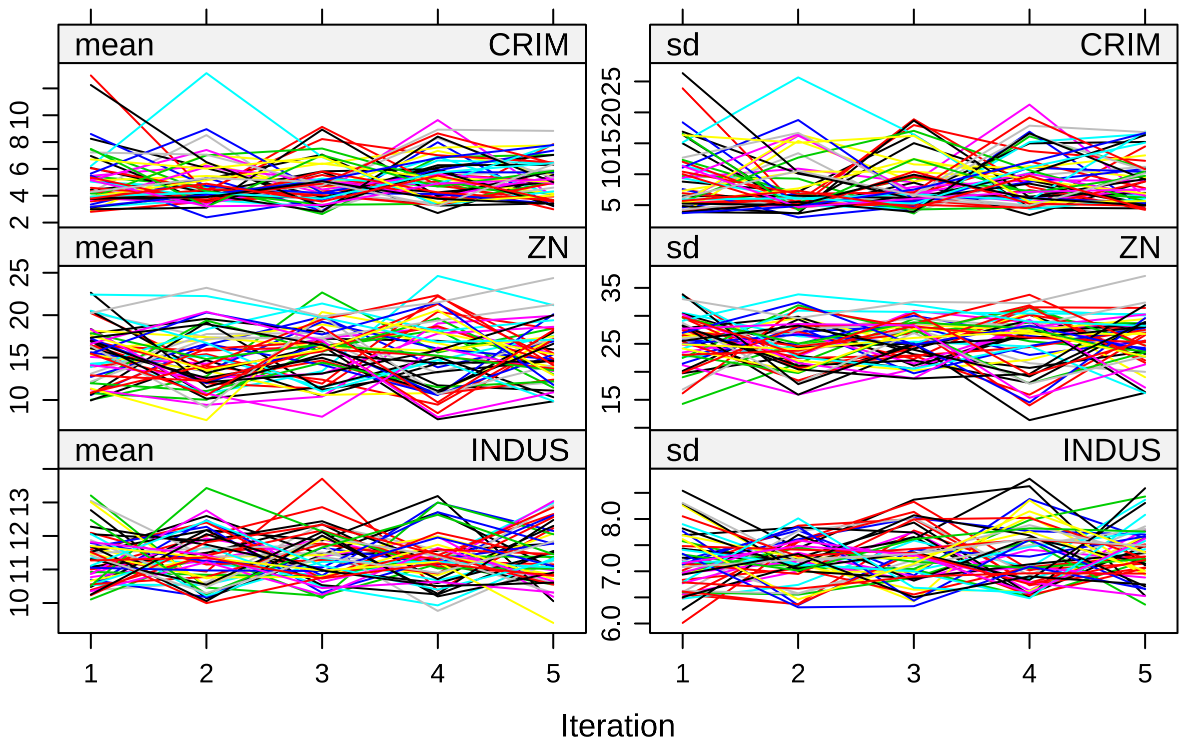


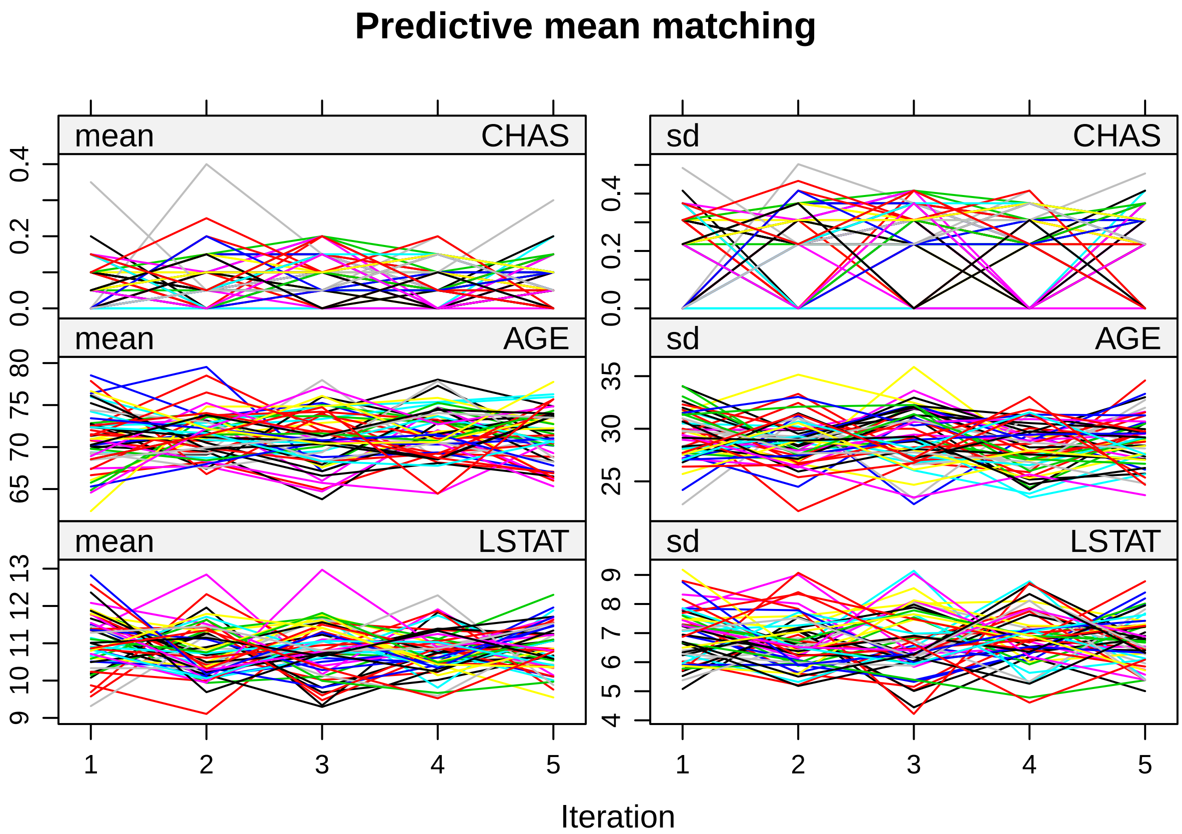
Forêts aléatoires




Forêts aléatoires
mean
CHAS
sd
CHAS
mean
AGE
sd
AGE
mean
LSTAT
sd
LSTAT
Iteration

Predictive mean matching





Predictive mean matching
mean
CHAS
sd
CHAS
mean
AGE
sd
AGE
mean
LSTAT
sd
LSTAT
Iteration

Annexe 4.2 : Comparaison des résidus