Journalisme des données : les revenus du patrimoine dans les communes de montagne

Sébastien Chaillou, Hélie Bazin, Kossivi Justin Ayivi

Mars 2022

Table des matières

1 Présentation des données **1**

2 Distribution des revenus du patrimoine **2**
2.1 Comparaison entre montagne et plaine 2
2.2 Comparaison par massif 5

3 Analyses Comparatives Montagne vs. Plaine **7**
3.1 Analyse de corrélation 7
3.2 Régression des revenus du patrimoine 11
3.3 Relation à la politique de taxe foncière 12

4 Conclusion **14**

1 Présentation des données

Pour la construction de la base de données, nous somme partis de la base Filosofi qui offre un descriptif assez complet de la répartition des revenus des communes françaises. On y trouve entre autres par commune :

1. nombre de ménages fiscaux et leur nombre d'habitants
2. médiane du niveau de vie
3. part des ménages fiscaux imposés
4. répartition du revenu (salaire, patrimoine, revenus d'activités, chomage, retraites...)
5. taux d'impositions

Il a ensuite fallu trier les différentes villes entre villes de montagne et de plaine. Pour cela, nous nous sommes basés sur les données de l'observatoire des territoires qui liste les communes classées en zone de montagne (zonage agriculture). Enfin, à l'aide du code géographique donné dans la base filosofi, nous avons pu répartir les communes en différents massifs : Alpes, Jura, Vosges, Pyrénnées, Massif Central, Réunion, Martinique, Corse.
Finalement, nous avons consulté les données INSEE de la taxe foncière (TF), qui nous est apparu comme un indicateur pertinent lorsqu'on traite de patrimoine, afin d'ajouter à notre base de données différents indicateurs par commune : Base nette TF, Base TF par habitant, taux TF, Produit TF.
En fusionnant donc trois bases de données différentes, nous avons pu construire une base de données complète de 34 000 communes. Cependant, les données étaient manquantes pour la majorité des communes : en effet, en dessous d'un certain nombre d'habitants, il devient difficile de récolter des données sur le revenu, notamment pour des questions d'anonymat. Nous avons donc dû mettre un filtre sur le nombre

de ménages par commune, et en observant le dataset, le seuil de 900 ménages s'est révélé comme étant le plus pertinent, avec quasiment aucune donnée manquante au-dessus, et quasiment systématiquement manquantes en-dessous. Nous obtenons finalement une base de données de 4840 communes.

2 Distribution des revenus du patrimoine

2.1 Comparaison entre montagne et plaine

Dans cette section, l'intérêt est porté sur les tests de comparaison de moyenne d'une part et d'analyse de la variance de deux échantillons indépendants d'autre part. Mais avant cela, il est présenté ci-dessus le diagramme en barre représentant la proportion des différents massifs et de la plaine.

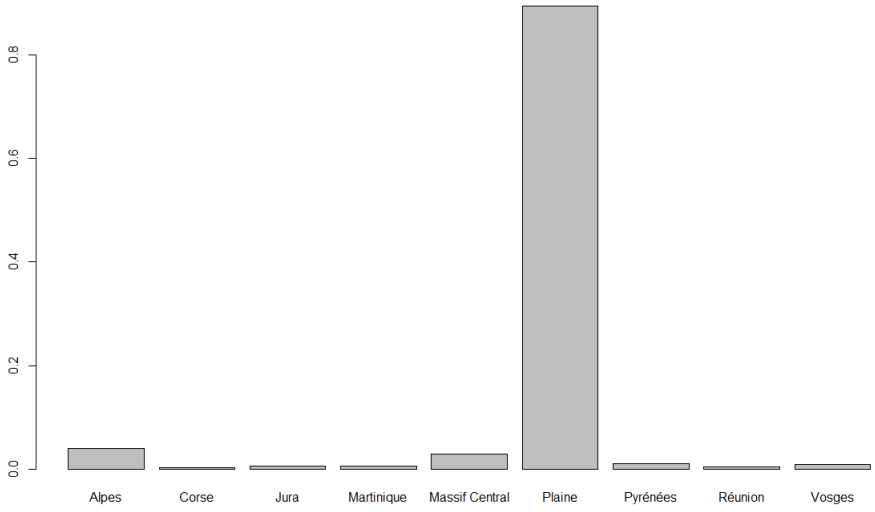

FIGURE 1 – Proportion des différentes zones Massif et Plaine

Il ressort de ce graphique que les ménages vivant en zone plaine sont majoritairement présent dans le jeu de données. A l'opposé les zones de montagne à l'image de la Corse, le Jura et la Réunion se partagent des proportions relativement très faibles. Ce point important justifiera par la suite, la subdivision en des zones de montagne et non montagne avant la mise en application des différents tests.

Étant donné la variable d'intérêt qui est le revenu de patrimoine, nous avons représenté ci-contre le revenu de patrimoine en fonction de la zone (C = montagne, NC = non montagne)

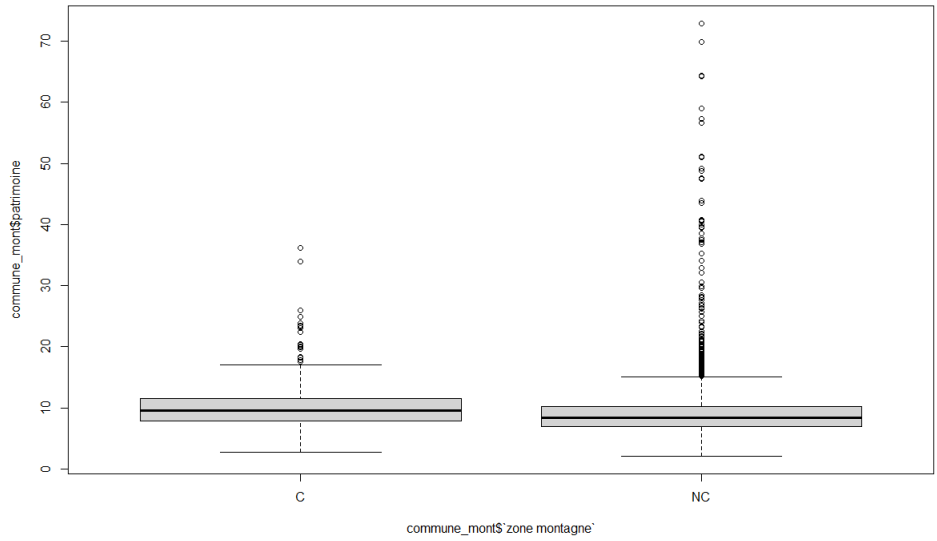

FIGURE 2 – Boite à moustaches des zones C et NC

En observant de prêt le diagramme à moustaches, on peut formuler une hypothèse d'égalité des moyennes dans les deux zones. Nous proposons de tester cela en appliquant un test d'égalité des moyennes. Le test de Welch est donc le plus adapté. Toutefois, ce test requiert deux hyothèses essentielles, celle que les données soient normalement distribuées et que la taille soit suffisamment grande($n > 20$). En supposant que ces derniers soient vérifiés (on pourrait tester l'hypothèse de normalité), on arrive au résultat suivant :

```
    Welch Two Sample t-test

data:  Massif and Plaine
t = 5.1467, df = 700.85, p-value = 3.446e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5771713 1.2891142
sample estimates:
mean of x mean of y
10.063796  9.130654
```

La p-valeur très faible nous amène à conclure qu'il subsiste une différence significative dans la moyenne des revenus de patrimoine au sein des zones. Néanmoins nous pourrions remettre en cause ce résultat suite à l'hypothèse de normalité admise. La comparaison des histogrammes de ces deux zones est obtenu ci-après :

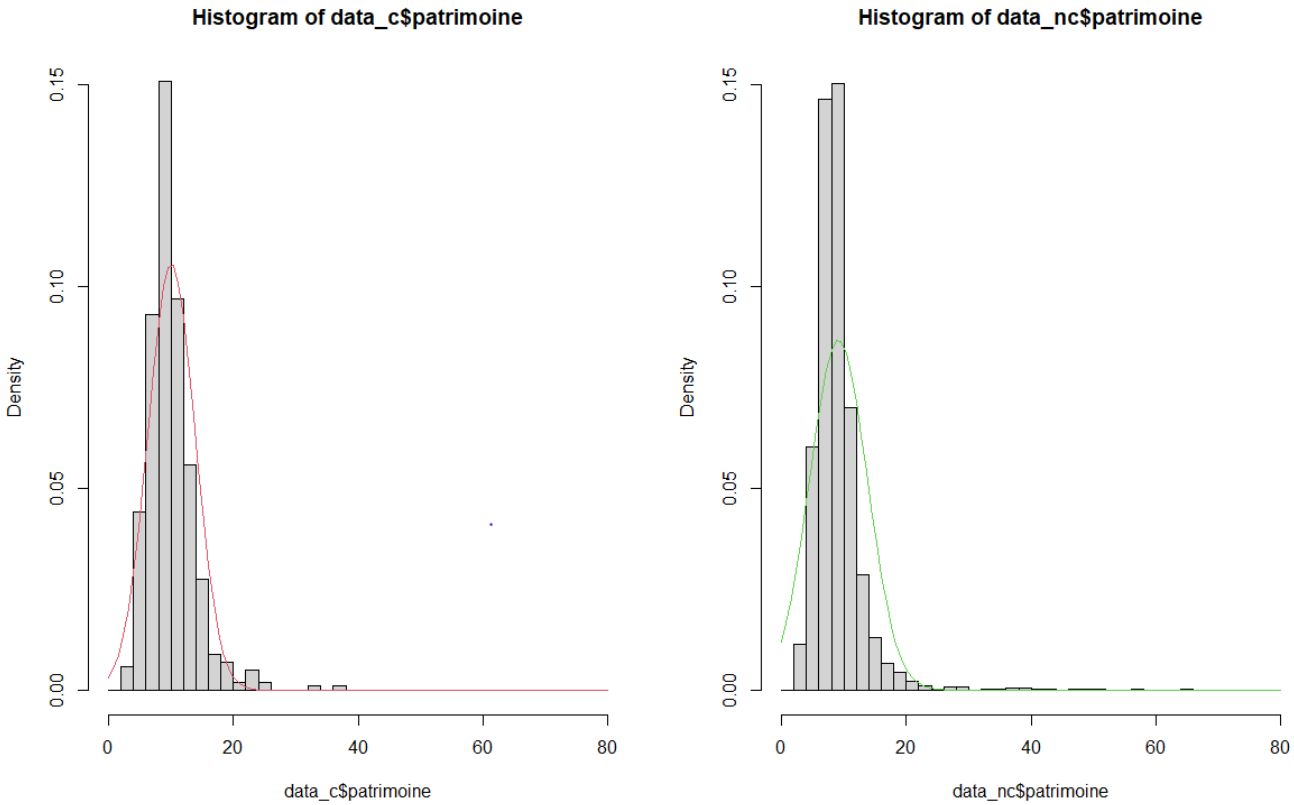

FIGURE 3 – Histogramme des revenus du patrimoine par zone C et NC avec fitting loi normale

Il suit donc que la loi normale n'est pas un meilleur ajustement de nos données notamment à cause de la hauteur des pics. La déduction juste graphique demeure insuffisante, nous proposons le test de Shapiro-Wilk pour tester l'adéquation à une loi normale. Les résultats sont présentés ci-après :

```
Shapiro-Wilk normality test

data:  data_c$patrimoine
W = 0.87021, p-value < 2.2e-16
```

```
Shapiro-Wilk normality test

data:  data_nc$patrimoine
W = 0.63213, p-value < 2.2e-16
```

La p-valeur très faible, ici également, indique le revenu de patrimoine dans les deux zones n'est vraisemblablement pas normalement distribué. Comme l'hypothèse de normalité n'est pas vérifiée, nous optons pour le test non paramétrique Wilcoxon Mann-Whitney au lieu du test de Student pour tester l'égalité des moyennes dans les deux zones. Dans ce cas, la statistique d'intérêt est plutôt la médiane que la moyenne.

```
    Wilcoxon rank sum test with continuity correction

data:  Massif and Plaine
W = 1350304, p-value = 2.918e-16
alternative hypothesis: true location shift is not equal to 0
```

En définitive, nous rejetons l'hypothèse que les revenus de patrimoine des zones C et NC sont statistiquement égaux (si on prend au hasard une commune de C et une de NC, il est statistiquement plus probable que celle de C ait un pourcentage des revenus du patrimoine plus élevé que celle de NC que l'inverse). En d'autres termes, ces diffférents tests rendent compte d'une présence d'un rôle de zone dans

les revenus de patrimoine.

2.2 Comparaison par massif

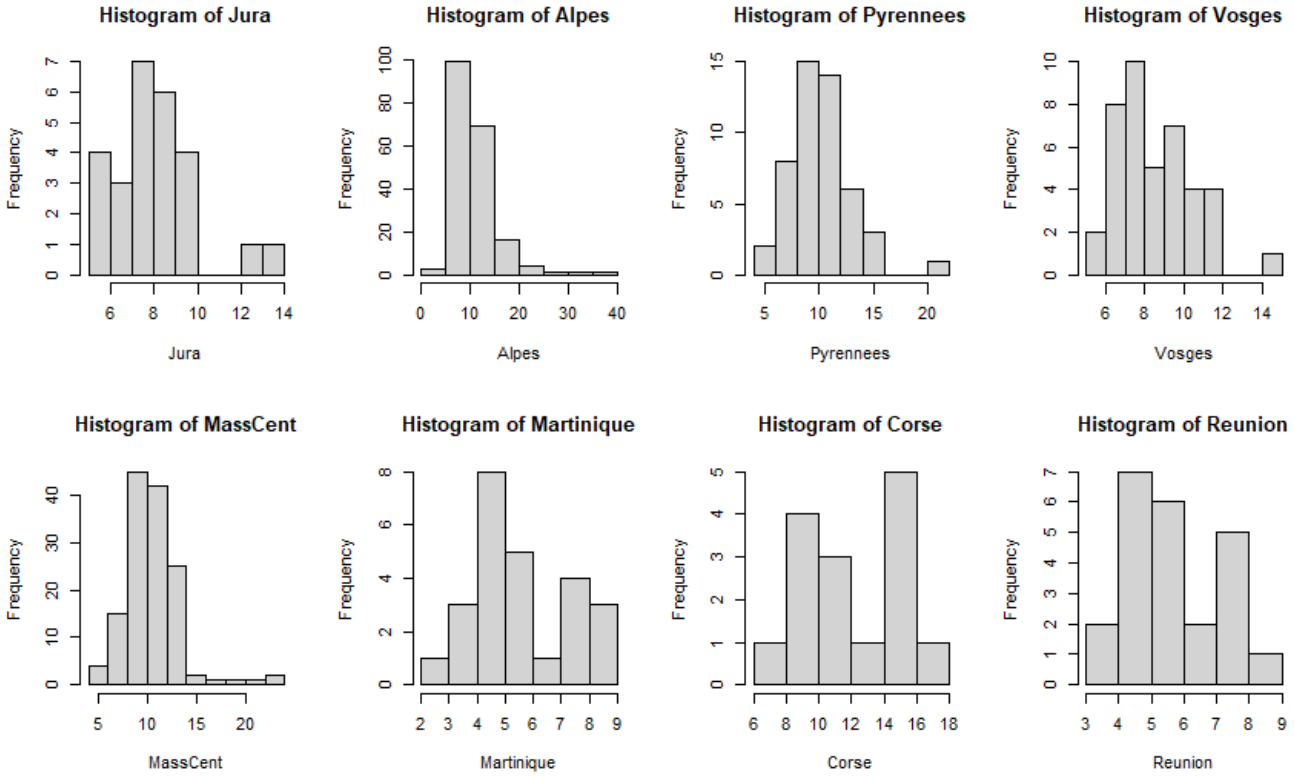

FIGURE 4 – Distribution du revenu du patrimoine par massif

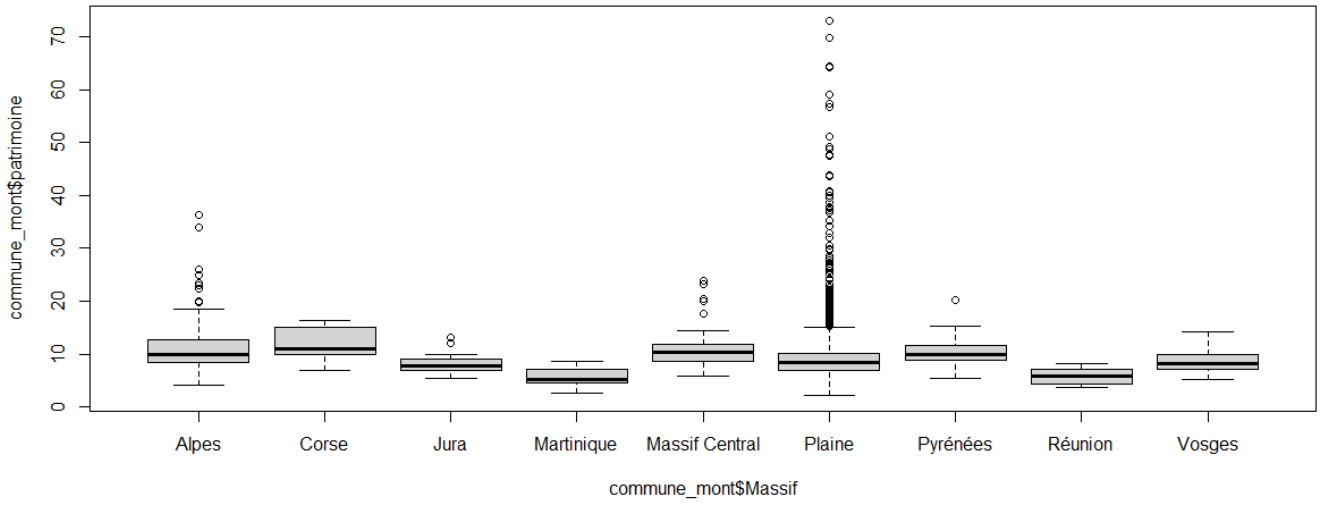

FIGURE 5 – Répartition du revenu du patrimoine par massif

Partant de cette analyse, nous pouvons appuyer les résultats en effectuant une analyse de variance au sein des différentes zones. Nous mettons ainsi un focus sur les différent massifs français. Le principe de l'anova est de trouver des différences de moyenne significatives entre les groupes. Toutefois, le test de l'anova ne nous donne pas assez d'informations sur les groupes qui sont très différents en terme de moyenne, nous ajoutons ainsi une fonction TukeyHSD [1] qui reporte les grandes différences de moyenne obtenues.

```
  Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = data_c$patrimoine ~ data_c$Massif)

$`data_c$Massif`
                              diff        lwr        upr     p adj
Corse-Alpes                0.9403780 -1.8383618  3.7191178 0.9697382
Jura-Alpes                -3.0455194 -5.2109539 -0.8800850 0.0005862
Martinique-Alpes          -5.5382887 -7.7415825 -3.3349948 0.0000000
Massif Central-Alpes      -0.5364336 -1.6910823  0.6182151 0.8506378
Pyrénées-Alpes            -0.7160846 -2.3738591  0.9416900 0.8931266
Réunion-Alpes             -5.4262887 -7.7128682 -3.1397091 0.0000000
Vosges-Alpes              -2.5726301 -4.3548545 -0.7904058 0.0003609
Jura-Corse                -3.9858974 -7.3477670 -0.6240279 0.0080831
Martinique-Corse          -6.4786667 -9.8650459 -3.0922874 0.0000003
Massif Central-Corse      -1.4768116 -4.2957253  1.3421021 0.7535301
Pyrénées-Corse            -1.6564626 -4.7160830  1.4031579 0.7207723
Réunion-Corse             -6.3666667 -9.8078157 -2.9255176 0.0000008
Vosges-Corse              -3.5130081 -6.6418073 -0.3842089 0.0156143
Martinique-Jura           -2.4927692 -5.3971239  0.4115854 0.1538130
```

1. Cette méthode a été inventé par John Tukey et lui doit son nom

```
Massif Central-Jura        2.5090858  0.2923347  4.7258370 0.0142336
Pyrénées-Jura              2.3294349 -0.1863133  4.8451830 0.0927639
Réunion-Jura              -2.3807692 -5.3488020  0.5872635 0.2237633
Vosges-Jura                0.4728893 -2.1265521  3.0723307 0.9993322
Massif Central-Martinique  5.0018551  2.7481064  7.2556038 0.0000000
Pyrénées-Martinique        4.8222041  2.2737955  7.3706126 0.0000004
Réunion-Martinique         0.1120000 -2.8837663  3.1077663 1.0000000
Vosges-Martinique          2.9656585  0.3345954  5.5967217 0.0149347
Pyrénées-Massif Central   -0.1796510 -1.9039178  1.5446158 0.9999843
Réunion-Massif Central    -4.8898551 -7.2250907 -2.5546194 0.0000000
Vosges-Massif Central     -2.0361965 -3.8804316 -0.1919614 0.0188448
Réunion-Pyrénées          -4.7102041 -7.3309537 -2.0894545 0.0000020
Vosges-Pyrénées           -1.8565455 -4.0511311  0.3380400 0.1675928
Vosges-Réunion             2.8536585  0.1524670  5.5548501 0.0299018
```

Chaque ligne du tableau de sortie inclut la différence entre les moyennes de deux groupes (diff) ainsi que les bornes inférieure et supérieure de l'intervalle de confiance (lwr et upr) pour la différence. En parcourant le tableau, nous voyons que la comparaison Massif Central-Martinique présente la plus grande différence, qui est de 5.0018 avec un intervalle de confiance de (2.7481064, 7.2556038). Ce qui nous amène à conclure qu'au sein des zones composant la zone de montagne, certaines se démarquent en termes de revenus de patrimoine. Par ailleurs, plusieurs intervalles de confiance (lwr, upr) ne contiennent pas 0, ce qui indique que la différence de moyenne au sein des groupes dans la zone de montagne est significative.

3 Analyses Comparatives Montagne vs. Plaine

Nous nous proposons dans cette partie de faire des comparaisons entre montagne et plaine en amenant d'autres variables que le % des revenus du patrimoine pour tenter de mettre le doigt sur de potentielles disparités remarquables.

3.1 Analyse de corrélation

Nous avons donc tracé deux matrices de corrélation entre les différents paramètres du modèle pour les communes de montagne et celles de plaine. Voici les résultats obtenus :

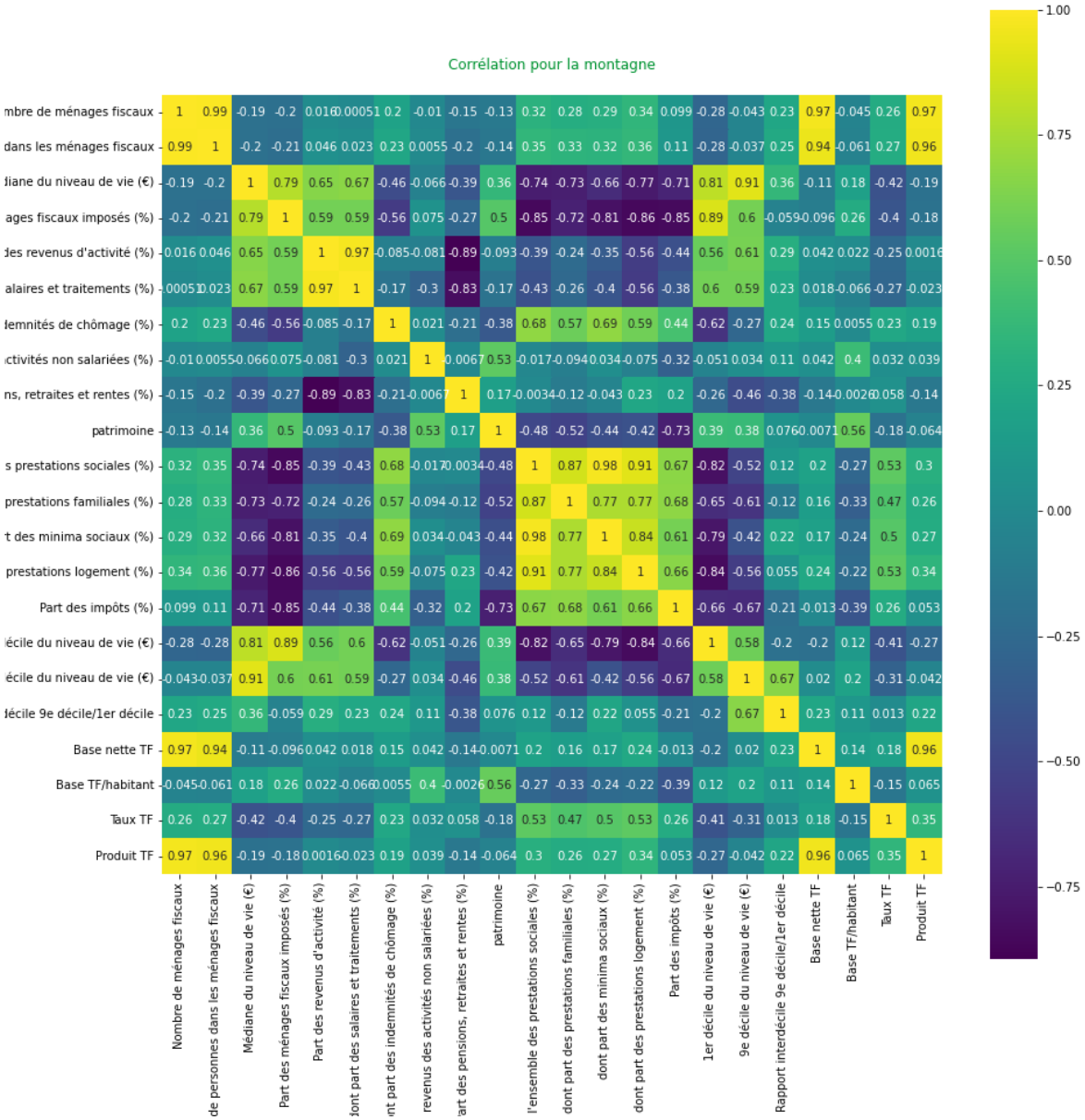

FIGURE 6 – Matrice de corrélation communes de montagne

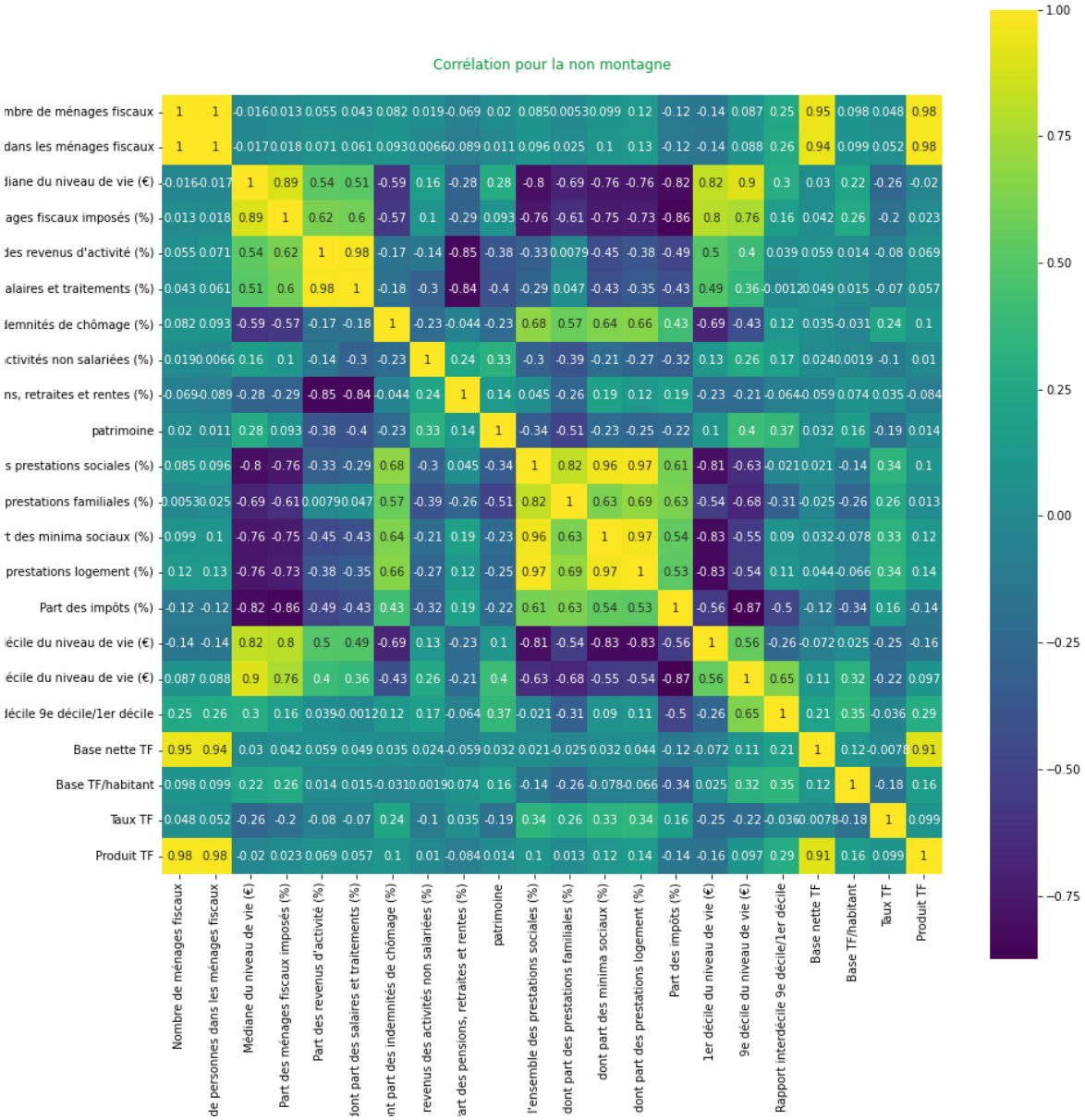

FIGURE 7 – Matrice de corrélation communes de plaine

Nous avons ensuite comparé les écarts entre plaine et montagne pour les corrélations entre revenus du patrimoine et les autres paramètres. Voici les résultats avec le plus de différence :

Paramètres	Base TF/habitant	part de ménages imposés	part des revenus d'activités	part des salaires	part des minimas sociaux	part des impots
Montagne	0.56	0.50	-0.01	-0.17	-0.44	-0.73
Plaine	0.16	0.093	-0.38	-0.4	-0.23	-0.22

Les écarts de corrélation peuvent se voir sur les graphiques. Nous regardons en particulier le logarithme de la base de la taxe foncière par habitant en fonction du pourcentage de revenus du patrimoine. Si la tendance est plus marquée en montagne, on constate que l'écart de corrélation est aussi dû à une répartition plus linéaire au niveau de la montagne. Ainsi, en régressant les deux quantités, le R^2 pour la montagne est 0.341 alors qu'il est de 0.043 pour la plaine.

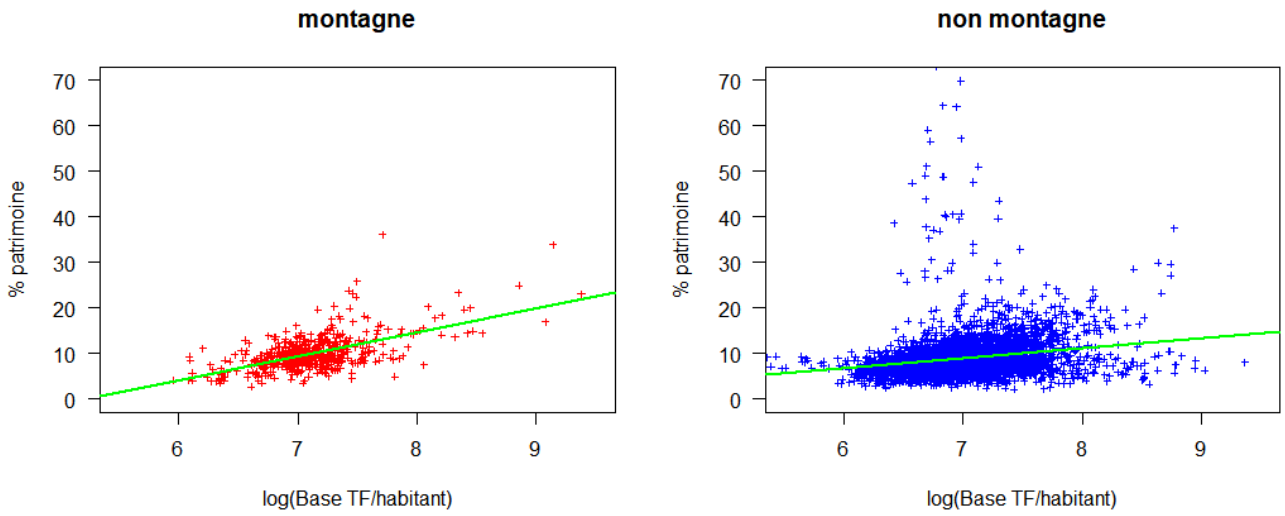

FIGURE 8 – patrimoine ~ base TF/habitant

Pour mettre en avant les différences relatives au nombre d'habitants où les écarts de corrélation sont moins importants, nous avons également tracé les graphiques du logarithme du nombre de ménages en fonction du pourcentage des revenus du patrimoine. Les droites de régression montrent une tendance similaire mais plus marquée pour les communes de montagne :

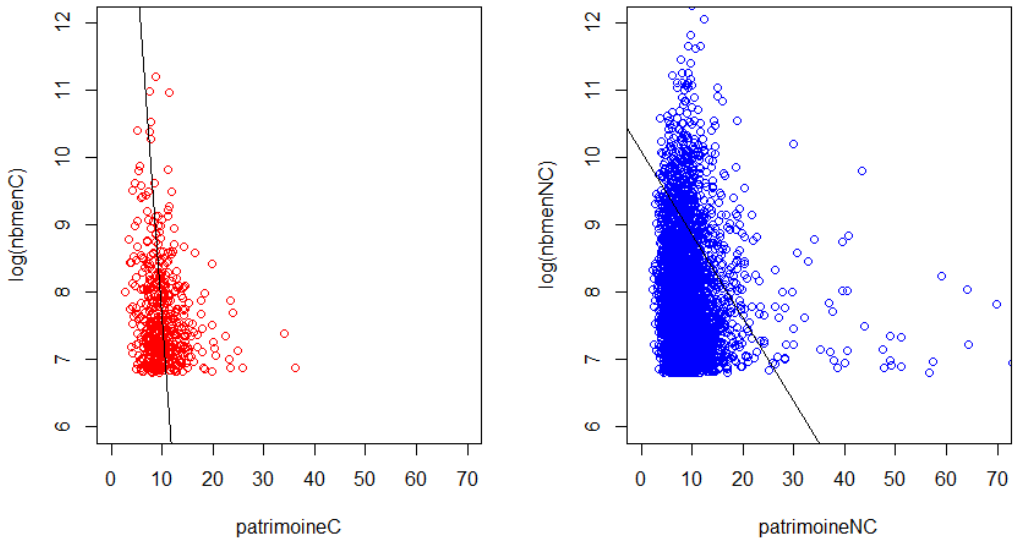

FIGURE 9 – patrimoine ~ nombre de ménages fiscaux

On peut donc tirer plusieurs hypothèses de ces études : à part de revenus du patrimoine égale en communes de montagne, il y a moins d'habitants et plus de personnes imposées qu'en plaine. De plus, pour la montagne, les communes avec des personnes plus riches (associées avec plus de revenus de patrimoine) ont tendance à avoir moins de personnes plus pauvres (ayant une plus grande porportion de minima sociaux). Enfin, à part de revenus du patrimoine égale, la commune de montagne paie plus d'impôts que les plaines. Tout cela tend à montrer qu'à revenu du patrimoine égal, la population des communes de montagne est plus riche, plus imposable et plus socialement homogène que les communes de plaine.

3.2 Régression des revenus du patrimoine

L'étape suivante dans notre étude a été de tenter d'expliquer la part de revenus du patrimoine en fonction des autres éléments de notre base de données. Nous avons donc régréssé les revenus du patrimoine en fonction du nombre de ménages, de la médiane du niveau de vie, de la part des ménages imposés, de la base taxe foncière et des taux taxe foncière. Voici les résultats obtenus :

```
Call:
lm(formula = patrimoineC ~ nbmenC + medianeC + menimpotsC + tfC +
    tauxtfC)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5802 -1.6941 -0.2654  1.3477 21.8151

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.552e+00  8.902e-01   1.744   0.0818 .
nbmenC      -2.760e-05  2.172e-05  -1.271   0.2045
medianeC    -2.993e-05  4.503e-05  -0.665   0.5066
menimpotsC   1.241e-01  1.631e-02   7.609 1.36e-13 ***
tfC          1.852e-03  1.383e-04  13.392  < 2e-16 ***
tauxtfC      2.571e-02  1.890e-02   1.360   0.1744
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.821 on 505 degrees of freedom
Multiple R-squared:  0.449,     Adjusted R-squared:  0.4435
F-statistic:  82.3 on 5 and 505 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = patrimoineNC ~ nbmenNC + medianeNC + menimpotsNC +
    tfNC + tauxtfNC)

Residuals:
    Min      1Q  Median      3Q     Max
-31.536  -1.845  -0.403   1.112  48.539

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.270e+00  5.510e-01  -4.119 3.88e-05 ***
nbmenNC      8.803e-06  3.288e-06   2.677  0.00746 **
medianeNC    1.289e-03  4.207e-05  30.644  < 2e-16 ***
menimpotsNC -3.170e-01  1.215e-02 -26.093  < 2e-16 ***
tfNC         8.178e-04  8.483e-05   9.641  < 2e-16 ***
tauxtfNC    -5.192e-02  9.562e-03  -5.430 5.95e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.06 on 4323 degrees of freedom
Multiple R-squared:  0.2209,    Adjusted R-squared:   0.22
F-statistic: 245.1 on 5 and 4323 DF,  p-value: < 2.2e-16
```

Le premier tableau donne les résultats obtenus pour les communes de montagne et le second pour les communes de plaine. On note une certaine hétérogénéité des résultats entre montagne et plaine. Il est

intéréssant de noter un impact bien plus important de la base taxe foncière habitant pour les communes de montagne, un coefficient de signe opposé pour la part de ménages imposés ainsi qu'un R^2 supérieur pour la montagne, qui tend à montrer un meilleur fit de ce modèle pour les communes de montagne que celles de plaine.

3.3 Relation à la politique de taxe foncière

Les communes votent pour leur taux de taxe foncière tous les ans. Ce taux est appliqué à la moitié de la valeur locative du bien réelle ou estimée et est donc parfait pour avoir une idée de la valeur patrimoniale immobilière au sein des communes. Ce taux est très disparate selon les communes et traduit les différentes approches des municipalités quant à leur revenus. Ici, nous souhaitons savoir si ces décisions politiques sont différentes entre les communes de plaine et de montagne.

Nous avons d'abord cherché à avoir la meilleur loi pour décrire les taux de la taxe foncière. Voici le graphe de Cullen et Frey pour la montagne et la plaine :

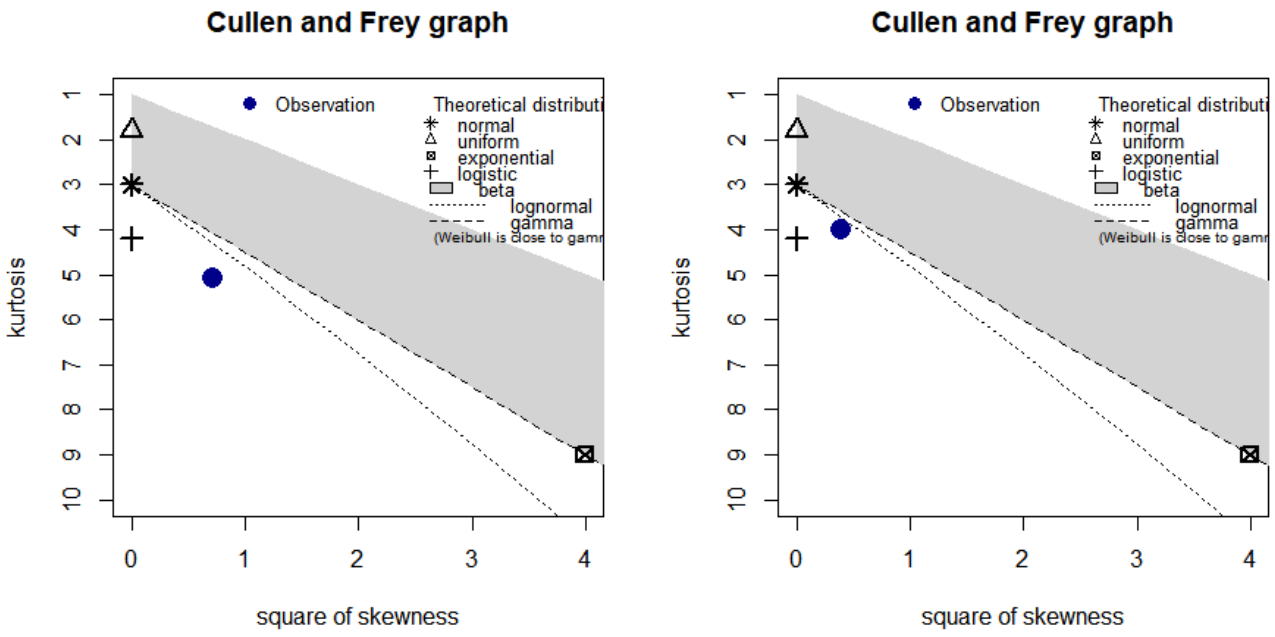

FIGURE 10 – Montagne (gauche) vs Plaine (droite)

Nous pouvons faire le constat que la loi lognormale semble la plus proche dans les deux cas, suivie de la loi gamma. Mais lorsque nous regardons les QQplots, la loi gamma semble plus appropriée notamment pour les valeurs extrêmes.

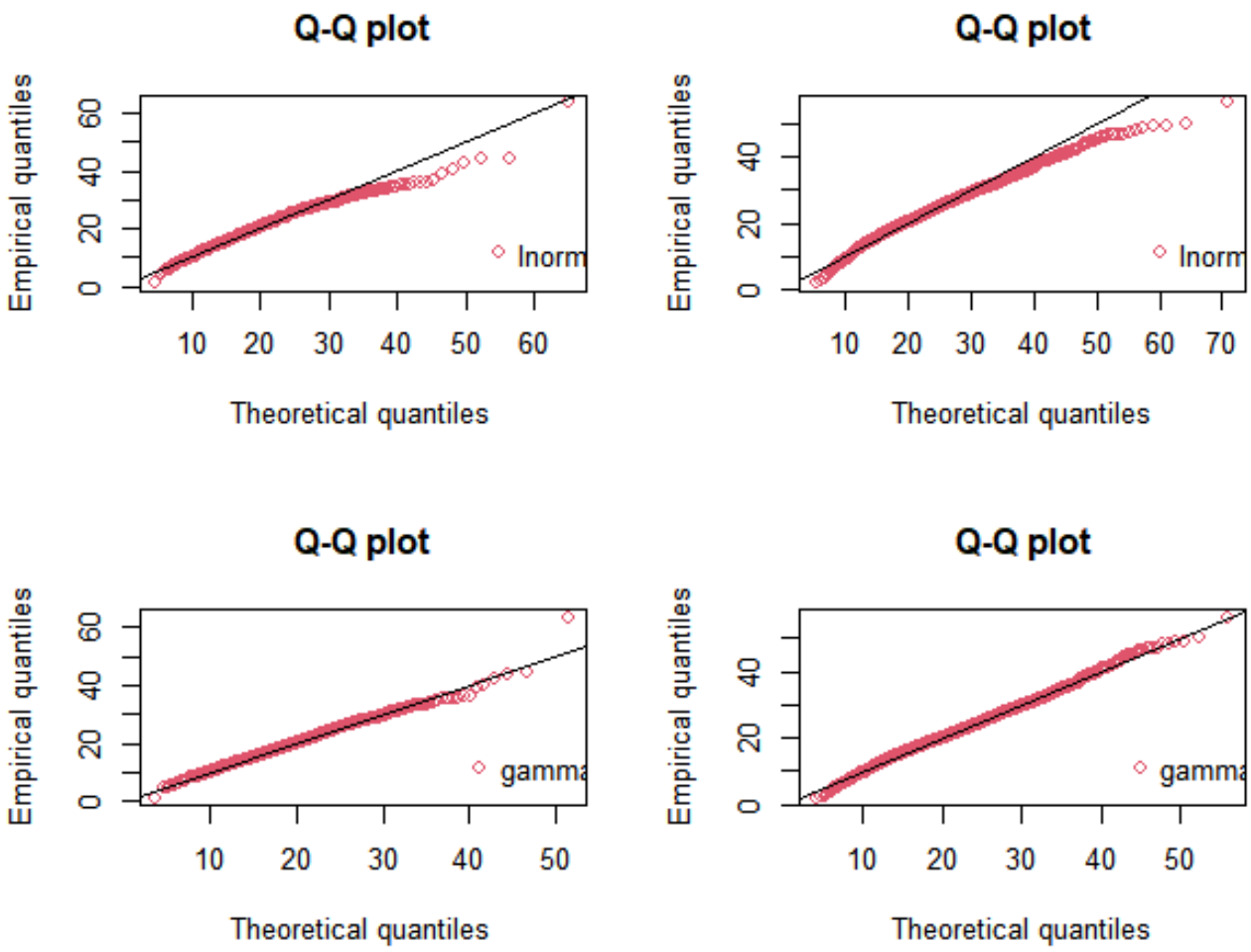

FIGURE 11 – Montagne (gauche) vs Plaine (droite)

Le critère AIC est aussi moins élevé pour la loi gamma que pour la loi lognormale (3471 contre 3506 pour la montagne). En utilisant le module scipy de Python, nous avons réalisé des tests de Kolmogorov-Smirnov vis à vis des lois gamma estimées. Pour la montagne, la p-valeur est de 0.989 et pour la plaine, elle est de 0.010. Le fait qu'il y plus de 8 fois plus de points pour la plaine accentue les différences avec les lois de référence mais dans les deux cas, ces valeurs propres restent assez élevées étant donné la taille des échantillons. Cependant, un tel indicateur n'est pas approprié pour comparer la qualité de l'adéquation entre la montagne et la plaine.

Nous avons ensuite tracé les deux lois gamma obtenues sur un graphe contenant l'histogramme du taux de la taxe foncière pour les communes de montagne.

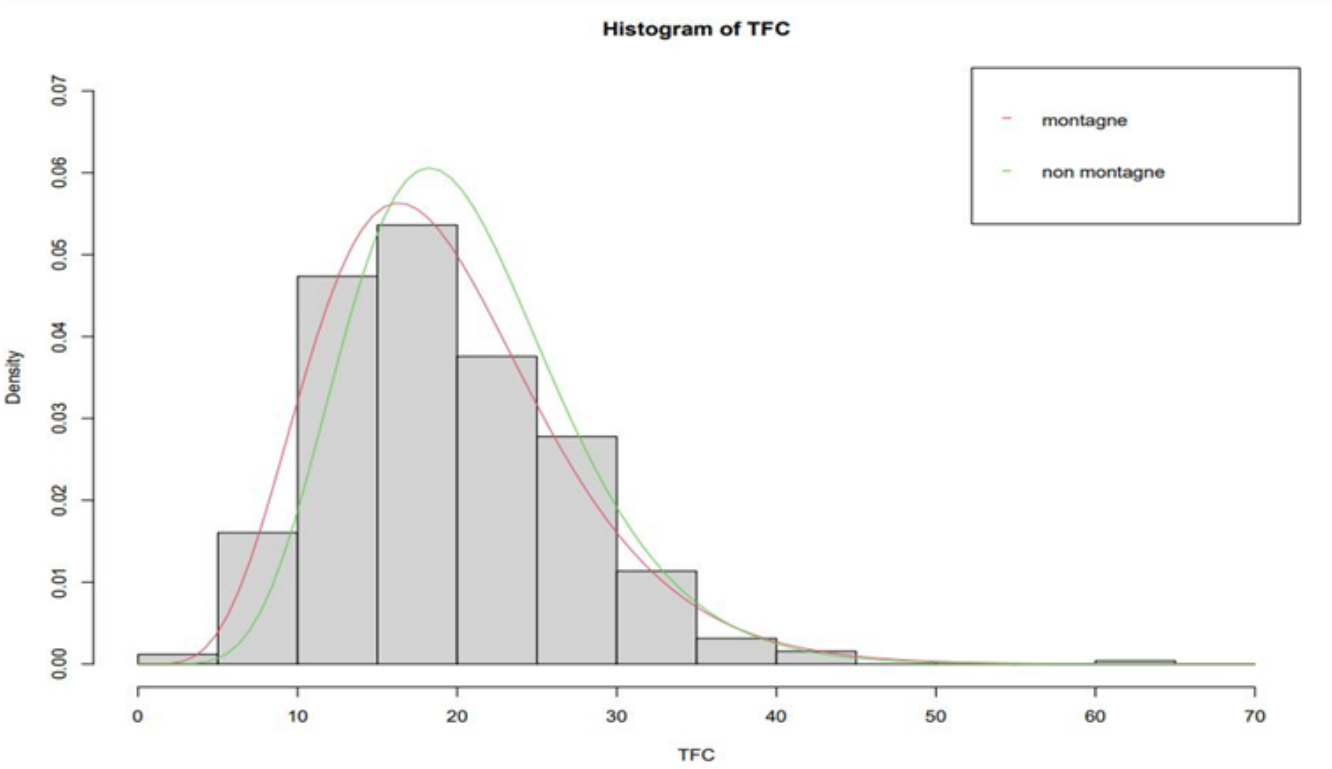

FIGURE 12 – Lois du taux de la taxe foncière

Les lois estimées semblent bien indiquer une différence entre les deux catégories. Sur R, la fonction fitdist nous donne aussi les estimations des paramètres shape et rate de la loi gamma ainsi que les écarts-types de ces estimations.

	montagne	non montagne
shape	6.469 ± 0.395	8.852 ± 0.187
rate	0.335 ± 0.021	0.430 ± 0.009

En multipliant par 2 les écarts-types nous obtenons facilement des intervalles de confiance d'au moins 95% qui sont pour le shape de [5.679 ; 7.258] pour la montagne et [8.478 ; 9.225] pour la plaine, et pour le rate de [0.293 ; 0.378] pour la montagne et [0.411 ; 0.449] pour la plaine. Ces intervalles ne se coupent pas ce qui nous permet de conclure que les lois sont bien différentes. Ainsi, le taux de la taxe foncière a tendance à être moins fort dans les communes de montagne avec une distribution plus étendue. Cela pourrait notamment traduire des politiques plus favorables à l'implantation immobilière pour les communes de montagne.

4 Conclusion

Nous avons pu mettre en évidence des différences relatives aux revenus du patrimoine en France entre les communes de montagne et de plaine. Les communes de montagne ont notamment tendance à avoir un pourcentage des revenus de patrimoine plus élevé ce qui peut être associé à une plus grande richesse. Le rapport à la richesse, notamment à travers le patrimoine immobilier, est plus saillant en montagne quand on compare à la base de la taxe foncière.

Références

B. L. Welch *The generalization of "Student's" problem when several different population variances are involved* Biometrika, 1947.

S. S. Shapiro, M. B. Wilk *An analysis of variance test for normality (complete samples)* Biometrika, 1965.

F. Wilcoxon *Individual comparisons by ranking methods* Biometrics Bulletin, 1945.

H. B. Mann, D. R. Whitney *On a test of whether one of two random variables is stochastically larger than the other* Ann. Math. Stat., 1947.

J. Tukey *Comparing Individual Means in the Analysis of Variance* Biometrics, 1949.

A. Cullen, H. Frey *Probabilistic Techniques in Exposure Assessment* Plenum Publishing Co., 1999.

A. Kolmogorov *Sulla determinazione empirica di una legge di distribuzione* G. Ist. Ital. Attuari., 1933.

N. Smirnov *Table for Estimating the Goodness of Fit of Empirical Distributions* Ann. Math. Statist., 1948.