# Rapport technique :

Recherche de liens et de régularités dans les données

## 1 : Tests entre les variables individuelles

La recherche de liens entre toutes les données a commencé par une série de tests statistiques :   
Régression linéaire, corrélation de Spearman, Chi2 d’indépendance, et tests d’égalité de moyennes et de variances. Pour tous ces tests, j’ai choisi 5% comme seuil de significativité, et étant donné le nombre de tests, j’ai utilisé la correction de Benjamini-Hochberg et dans certains cas celle de Bonferroni pour corriger les p-valeurs.

Les conclusions étant qu’il existe beaucoup de liens entre les variables, mais le nombre de variables rend cela difficile à analyser. Nous avons donc cherché à réduire la dimension.

Pour la suite, nous n’utiliserons pas les individus ayant une chaudière ou une cuisinière, dans les anciens et les nouveaux appareils, car cela représente des appareils trop différents et pas de chauffe. Nous avons aussi renommé les poêles de masse et les poêles hydrauliques en simples poêles.

## 2 : Première réduction de dimension au niveau de l’individu

Étant donné que nous avons des variables sont quantitatives ainsi que des qualitatives, nous avons procédé à une analyse factorielle des données mixtes (AFDM). Mais, il fût difficile de résumer l’information, même en sélectionnant minutieusement les variables à inclure ou en essayant de faire plusieurs AFDM selon des groupes (ex : territoire, type de combustible…). Nous avons donc choisi de changer d’échelle.

## 3 : Passage au niveau communal

Étant donné les premiers résultats de la réduction de dimension au niveau de l’individu, et étant donné que l’une des hypothèses est que le comportement des bénéficiaires change en fonction de la proximité à la ressource et des caractéristiques socio-économique du lieu de vie, nous avons continué l’analyse mais au niveau communal cette fois ci.

Pour cela, nous avons transformé les variables individuelles pertinentes, en les regroupant par communes : en utilisant la moyenne communale pour les variables qualitatives, et en créant une variable représentant le taux par communes de chaque modalité des variables quantitatives (par exemple, le taux de primes majorées par commune).

Nous avons d’abord choisi de ne pas traiter les communes présentant moins de 10 dossiers, car leur petit effectif engendre une grande volatilité des taux. Puis, nous avons sélectionné les variables représentant des comportements de consommation du bois, mais aussi celles représentant la capacité budgétaire des demandeurs, ainsi que leurs motivations :  
"Type.Ancien.appareil", "Usage.ancien.matériel", "Nouveau.matériel", "Type.combustible.nouveau.materiel", "Usage.nouveau.matériel", "majoration", "cout\_total\_TTC, et "Motivation changement appareil".  
Nous n’utilisons pas les variables représentant les modalités de type "autre", "pas de réponse, "NA" car elles présentent peu d’individus, ni une des modalités des variables devenues binaires (combustible, majoration) car elles sont déductibles de l’autre modalité, on évite donc un doublon.

Afin d’illustrer la proximité à la ressource et les caractéristiques socio-économique des communes, nous avons ajouté des variables communales provenant de SIDDT ou des bases de données de l’INSEE, telles que le taux de forêt par commune, la médiane de niveau de vie, la densité de population, ou encore l’évolution du nombre de logements entre 2015 et 2020, mais aussi le nombre de dossiers, le nombre d’appareils « anciens » a changé fin 2022, et le nombre d’appareils changés en 2023, et ceci aussi ramené au nombre d’habitants.

La recherche de communes « outliers » n’a pas donné de résultat consistant, sur les variables issues des dossiers comme les variables supplémentaires.

Il existe des variables liées entre elles parmi les variables supplémentaires, mais nous pouvons toutes les garder car elles ne rentrent pas dans le calcul de l’ACP, et donc du clustering. Cependant, nous devrons prendre en compte ces corrélations dans l’interprétation des futurs résultats.

## 4 : ACP communale

Les 10 premières dimensions de l’ACP représentant plus de 95% de la variance, nous avons choisi de garder ces dix dimensions pour la suite.

En utilisant l’algorithme des kmeans et le clustering ascendant hiérarchique sur les 10 premières dimensions, le choix de séparer en deux groupes semble le plus optimal, déjà car cela simplifie l’interprétation. De plus, la séparation entre les deux groupes se fait principalement sur la première dimension de l'ACP, or c'est sur cette dimension que la plupart des variables sont le mieux représentées.

Nous avons ensuite cherché à comparer ces deux groupes sur les différentes variables à l’aide de tests statistiques. Etant donné que seules 2 des variables sont normalement distribuées, nous avons utilisé le test non paramétrique d'égalité des distributions de Wilcoxon, puis, afin de nous assurer de la véracité des résultats, nous avons réalisé des tests paramétriques d'égalité d'espérances, via ANOVA. Puisque nous réalisons une série de tests, nous avons utilisés la correction de p-valeur de Benjamini-Hochberg, et nous avons choisi le seuil 5% pour rejeter ou non l’hypothèse nulle.

Les hypothèses de différences de comportements entre utilisateurs étant pour les utilisateurs du bois bûche, nous avons ensuite réalisé la même analyse mais uniquement avec les personnes se chauffant à la bûche. Les résultats furent assez similaires.

## 5 : Deuxième réduction de dimension a niveau de l’individu

Nous avons choisi de recommencer l’AFMD au niveau de l’individu, mais cette fois-ci en utilisant principalement les résultats permettant d’exploiter plus de deux dimensions, afin de garder au moins 80 % de variance expliquée. Le but étant de différencier les individus selon leurs comportements quant à l’utilisation du bois comme combustible, tout en prenant en compte leurs caractéristiques socio-professionnelles, c’est sur ces variables que nous avons basé la réduction de dimension : "Type.Ancien.appareil", "Usage.ancien.matériel", "Nouveau.matériel", "Usage.nouveau.matériel", "Année.installation", "Durée.séchage.bois", "Approvisionnement", "Stockage.bois", "Ramonage.installation", "cout\_total\_TTC", "Revenus", "Nb.pers.ménage", « Sit.pro.conjoint », « Sit.pro.demandeur ».

Le type de combustible des anciens appareils étant principalement la bûche, nous ne l’avons pas pris en compte. Pour les nouveaux appareils, seules les poêles peuvent ne pas être à la bûche, nous ajoutons la distinction du type de combustible directement dans la variable "Nouveau.matériel".  
Les individus ne présentant pas de réponse aux variables "Nouveau.matériel", et "Type.combustible.nouveau.materiel" mais uniquement pour les poêles, ont été supprimés.

Pour que les situations professionnelles des demandeurs et celles des conjoints ne représente plus qu’une seule variable, car nous ne connaissons pas le « chef.fe de famille », nous avons procédé à la transformation suivante : Nous avons créé une variable binaire pour chaque modalité, qui sera d’un si elle correspond à au moins une personne du ménage.  
Nous avons transformé les variables où plusieurs réponses pouvaient être cochées : "Ramonage.installation", "Stockage.bois" , "Approvisionnement" : en variables binaires : une réponse par modalités, en enlevant les « Pas de réponse ». Ce n’est pas aberrant de les inclure tel quel dans l’AFDM, car l’algorithme de réduction de dimension procède aussi à cette transformation des modalités en binaire, mais l’aurait fait sur l’ensemble des combinaisons de réponses réalisés.

Nous avons gardé quelques variables supplémentaires que nous n’avons pas inclues dans la réduction de dimensions car elles représentent des doublons ou qui ne sont juste pas pertinentes : "Territoire", "montant\_aide", "cluster\_com", "type\_commune", "Age" (met trop en avant les retraités), "majoration", "Freq.utilisation.période.chauffe".

Nous n’avons pas utilisé les motivations du changement d’appareils car elles n’apportent pas d’information supplémentaires.

Nous avons réalisé plusieurs réductions de dimension en séparant le jeu de données selon le temps, le territoire, ou même le type de combustible, en prenant que les variables relatives à l’usages du bois, et en réalisant un analyse factorielle multiple (donc qu’avec les variables qualitatives), mais nous n’avons pas obtenu de résultats complétement différents ou pertinents.

Nous avons gardé toutes les dimensions de l’AFDM afin de conserver 100% de la variance pour la suite.

En utilisant l’algorithme des kmeans et le clustering ascendant hiérarchique, nous avons opté pour une séparation des individus en deux clusters : ils se séparent très bien sur les deux premières dimensions, c’est ce que suggère les dendrogrammes du clustering ascendant hiérarchique, et c’est ce nombre de cluster qui maximise le critère de Calinski et Harabasz. De plus, les deux groupes se séparent selon des caractèristiques que l’on voyait apparaitre aussi en modifiant le nombre de clusters, et en sélectionnant différemment les individus et / ou les variables.

Afin de comprendre les caractéristiques de ces clusters, nous avons calculé, en plus des graphiques, des « odds-ratios » : Les chances d’un individu d’appartenir au groupe 2 quand il présente une caractéristique, contre les chances s’il ne présente pas la caractéristique. Un OR est significatif, c’est-à-dire que la modalité amène à être dans le groupe 2 ou à ne pas être dans le groupe 2, si la borne inférieure de son intervalle de confiance est supérieure à 1 dans le premier cas, ou si la borne supérieure de son intervalle de confiance est inférieure à 1 pour le deuxième. Et afin que les résultats soient symétriques, nous avons calculé le logarithme de l’OR et de ses intervalles de confiance. On compare donc non plus à 1 mais à 0.