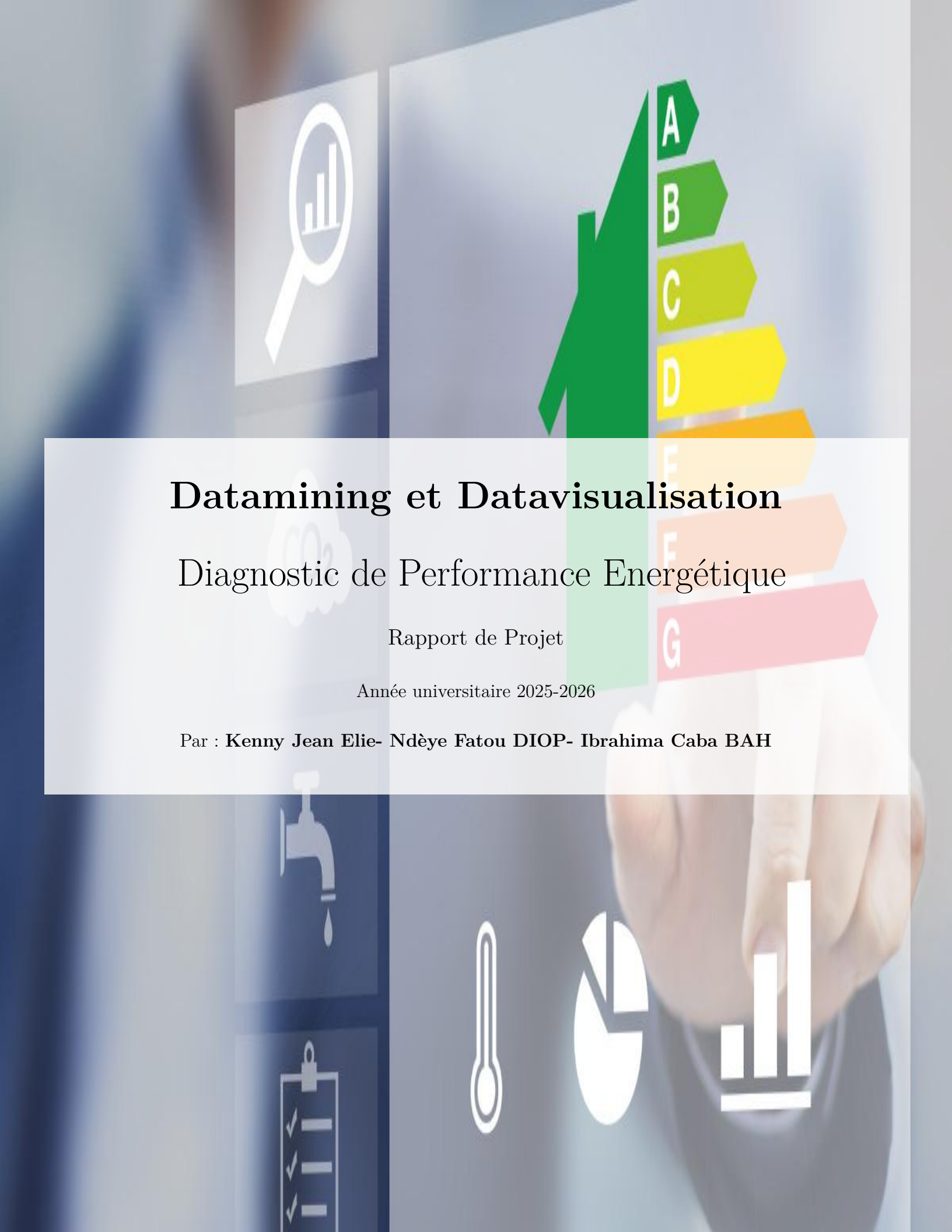

Datamining et Datavisualisation

Diagnostic de Performance Energétique

Rapport de Projet

Année universitaire 2025-2026

Par : **Kenny Jean Elie- Ndèye Fatou DIOP- Ibrahima Caba BAH**

Table des matières

Introduction

Face aux enjeux climatiques contemporains et à la nécessité impérieuse de réussir la transition énergétique, le secteur du bâtiment se retrouve au cœur des politiques publiques. En France, le secteur résidentiel est l'un des plus gros consommateurs d'énergie finale et l'un des principaux émetteurs de gaz à effet de serre. Dans ce contexte, l'exploitation des Diagnostics de Performance Énergétique (DPE) ne représente plus seulement une obligation légale, mais devient une source de données stratégique pour comprendre et agir sur l'efficacité thermique de notre parc immobilier.

Le présent projet s'inscrit dans cette démarche d'intelligence des données (Data Mining). Nous disposons d'un jeu de données riche et multidimensionnel incluant des variables techniques (surface, année de construction), économiques (coûts annuels d'énergie) et environnementales (émissions de GES). Cependant, la complexité de ces informations et les fortes corrélations entre elles rendent leur interprétation directe difficile. L'enjeu de notre étude est donc de répondre à la problématique suivante : Comment structurer et segmenter les données énergétiques pour identifier des profils types de logements et hiérarchiser les facteurs déterminants de leur consommation ?

Pour répondre à cette question, nous avons adopté une méthodologie rigoureuse articulée autour de trois axes majeurs. Tout d'abord, nous mettons en œuvre une Analyse en Composantes Principales (ACP) afin de réduire la dimensionnalité du problème, d'éliminer les redondances d'informations et de dégager les axes structurants de la consommation. Ensuite, sur la base de ces axes, nous réalisons une Classification (HCPC) pour regrouper les logements en classes homogènes. Enfin, nous procédons à une caractérisation détaillée de ces groupes afin d'isoler les spécificités des « passoires thermiques » par rapport aux logements performants.

Ce rapport détaille chacune de ces étapes, en commençant par l'analyse fondamentale des structures de corrélation via l'ACP, pivot central de notre exploration statistique.

0.1 Traitement de la base de donnés

La phase de prétraitement a été déterminante pour réconcilier les caractéristiques techniques des Diagnostics de Performance Énergétique (DPE) avec les relevés de consommation réelle. Ce processus s'est articulé autour de trois axes :
- **Standardisation et appariement**: Le défi principal résidait dans l'absence d'identifiant unique commun aux deux sources. Nous avons donc construit une adresse standardisée en procédant à une normalisation syntaxique rigoureuse : passage en majuscules, suppression de la ponctuation et correction des espaces. Ce traitement garantit la fiabilité de la jointure en éliminant les disparités de saisie textuelle entre les bases.
- **Filtrage temporel et gestion de l'unicité** :Afin d'assurer la pertinence de l'étude, nous avons restreint l'échantillon aux données de l'année 2024. Les dates de réception des DPE ont été converties pour isoler cette période, et un traitement spécifique a été appliqué aux codes postaux pour harmoniser leur format. Par ailleurs, une procédure de dédoublonnage a été effectuée sur la base DPE pour ne conserver qu'un seul enregistrement par adresse, assurant ainsi l'unicité des logements analysés.
- **Fusion et constitution de la base finale**: Le processus s'est conclu par une jointure interne (Inner Join) basée sur l'adresse standardisée. Ce choix méthodologique permet d'exclure les enregistrements incomplets et de ne retenir que les logements présents simultanément dans les deux référentiels.

Le résultat est une table consolidée et robuste, socle de notre analyse statistique. Elle permet d'étudier avec précision la corrélation entre la performance théorique des logements et leur consommation énergétique réelle pour l'année 2024.

1 Analyses en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) constitue la première étape de notre étude multidimensionnelle. Son objectif est double : d'une part, synthétiser l'information contenue dans notre jeu de données composé de variables quantitatives, et d'autre part, identifier les structures de corrélation sous-jacentes entre les indicateurs de consommation énergétique et les caractéristiques des bâtiments.

1.1 Analyse de la matrice de corrélation

Avant de procéder à la réduction dimensionnelle, l'examen de la matrice de corrélation est essentiel pour justifier l'emploi de l'ACP.

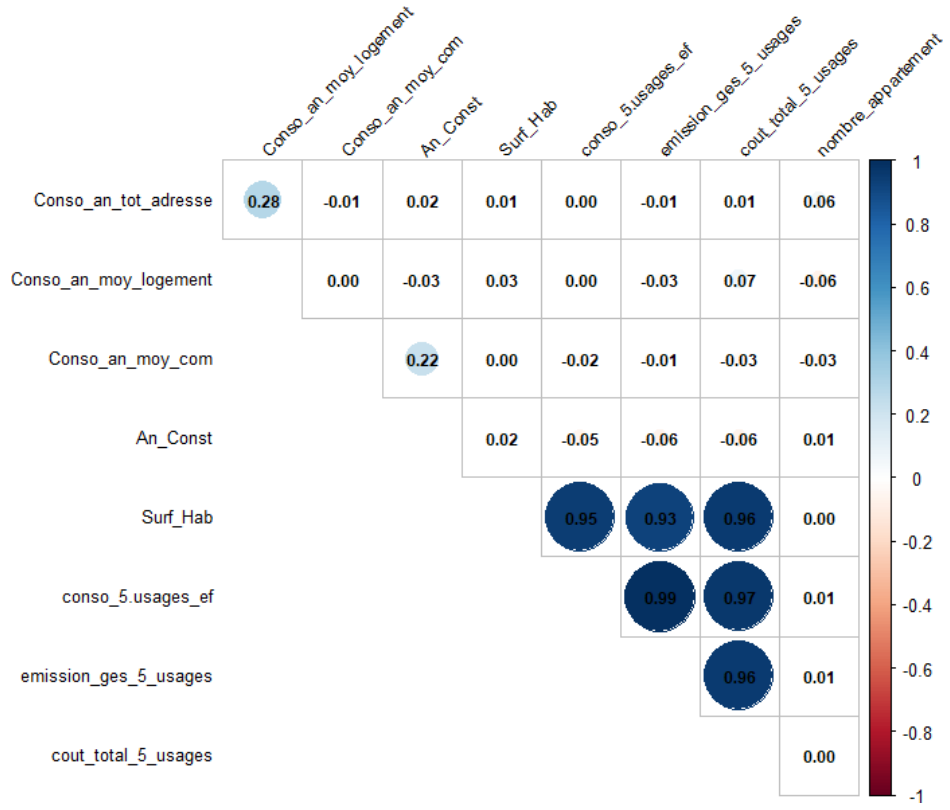

FIGURE 1 – Matrice de corrélation

- **Bloc de forte colinéarité** : Nous observons un groupe de variables extrêmement liées entre elles : la surface habitable (*surface_habitable_logement*), la consommation des 5 usages (*conso_5.usages_ef*), les émissions de gaz à effet de serre (*emission_ges_5_usages*) et le coût total (*cout_total_5_usages*). Les coefficients de corrélation entre ces variables sont tous supérieurs à 0,93, frôlant même 0,98 pour le couple coût/consommation. Cela indique une redondance d'information : ces variables décrivent essentiellement le même phénomène de "volume énergétique" lié à la taille du logement.
- **Corrélations modérées et faibles** : La variable C.totale présente une corrélation modérée avec *C.moyenne/adr* (0,30). En revanche, l'année de construction et le nombre d'appartements semblent être des variables indépendantes des indicateurs de consommation brute, avec des coefficients proches de zéro.

1.2 Étude des valeurs propres et choix de la dimensionnalité

L'analyse de l'inertie (les valeurs propres) nous permet de décider du nombre de composantes principales à conserver pour ne pas perdre trop d'information tout en simplifiant le modèle. D'après les résultats :
- Le critère de Kaiser (valeurs propres > 1) nous suggère de retenir les 4 premiers axes.
- L'inertie cumulée montre que les trois premiers axes capturent déjà plus de 70 % de la variance totale du jeu de données. Pour ce rapport, nous nous concentrerons principalement sur le premier plan factoriel (Dim 1 & 2) et le troisième axe pour l'interprétation.

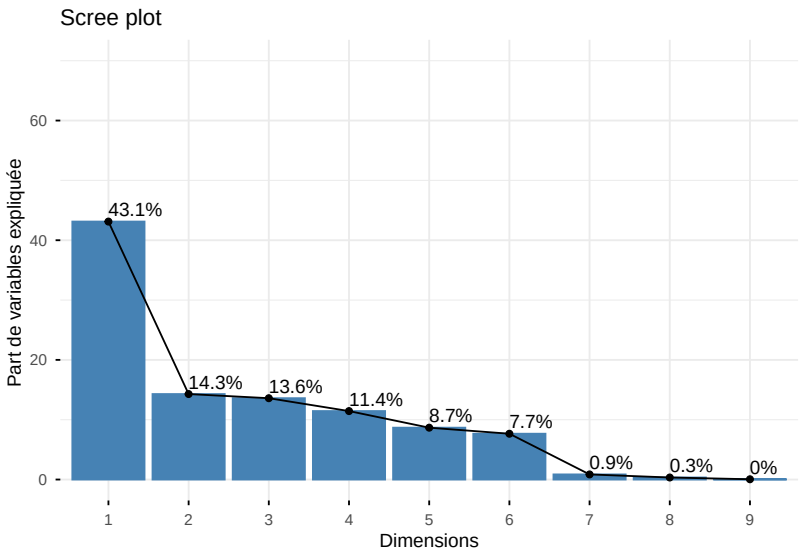

FIGURE 2 – Représentation des inerties

1.3 Tableau des coordonnées des variables:

TABLE 1: Coordonnées des variables sur les axes 1 à 3

	F1	F2	F3
Conso_an_tot_adresse	0.002	0.792	0.084
Conso_an_moy_logement	0.025	0.803	0.044
Conso_an_moy_com	-0.028	-0.067	0.776
An_Const	-0.056	-0.066	0.777
Surf_Hab	0.972	0.007	0.069
conso_5.usages_ef	0.993	-0.022	0.001
emission_ges_5_usages	0.985	-0.044	-0.002
cout_total_5_usages	0.987	0.033	-0.002
nombre_appartement	0.005	-0.009	-0.060

1.4 Etude des contributions des variables:

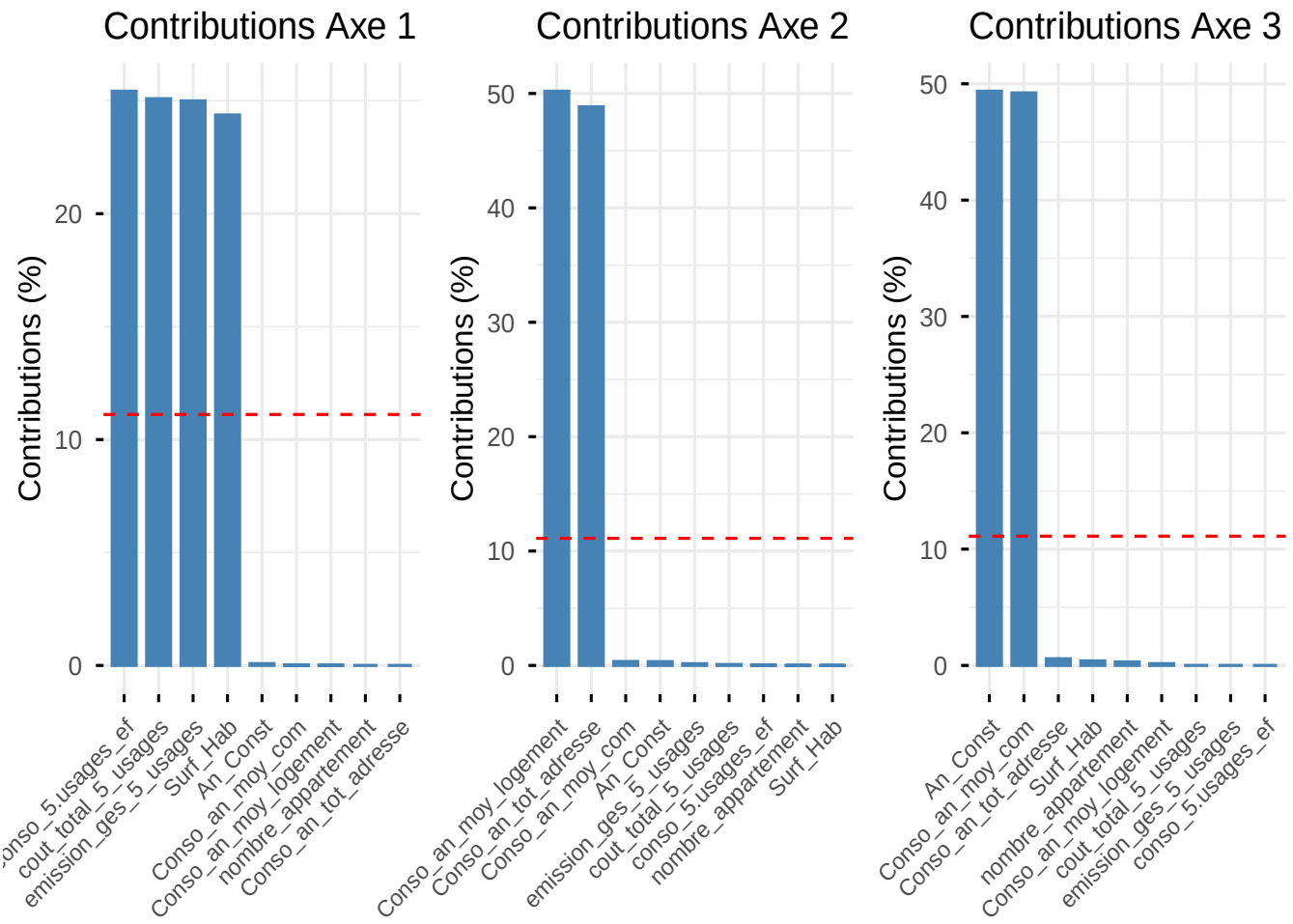

FIGURE 3 – Etude des contributions des variables

Interprétation des axes factoriels:
Axe 1 : Le facteur de "Gabarit et Consommation" (43,1 % d'inertie)
Cet axe est le plus discriminant. Il est presque exclusivement construit par les variables de volume :
- *Variables dominantes* : conso_5.usages_ef (coord: 0,99), cout_total_5_usages (0,99) et surface_habitable_logement *(0,97).*
- Signification* : Cet axe oppose les grands ensembles ou logements vastes, fortement énergivores et coûteux, aux petites unités d'habitation économes. Sa contribution à la variance est massive.
Axe 2 : Le facteur de "Consommation Spécifique" (14,5 % d'inertie) Ce deuxième axe apporte une information complémentaire non liée à la surface :
- *Variables dominantes*: C.totale (coord: 0,80) et C.moyenne/adr (0,80).
- *Signification* : Il représente l'intensité de la consommation par adresse, indépendamment du gabarit physique du bâtiment.
Axe 3 : Le facteur "Temporel et Communal" (13,4 % d'inertie) Ce troisième axe introduit une dimension temporelle et géographique :
- *Variables dominantes* : annee_construction (coord: 0,80) et C.moyenne/com (0,80).
- *Signification* : Cet axe permet d'isoler l'effet de l'ancienneté du bâtiment et les disparités géographiques (moyennes par commune).

1.5 Représentation des variables avec la qualité de représentation

Le cercle des corrélations (Plan F1-F2) illustre parfaitement ces regroupements:
- Les vecteurs représentant la surface, le coût et la consommation de GES sont très longs et superposés sur la partie droite de l'axe 1, indiquant une excellente qualité de représentation ($Cos^2 > 0{,}94$ pour la plupart).

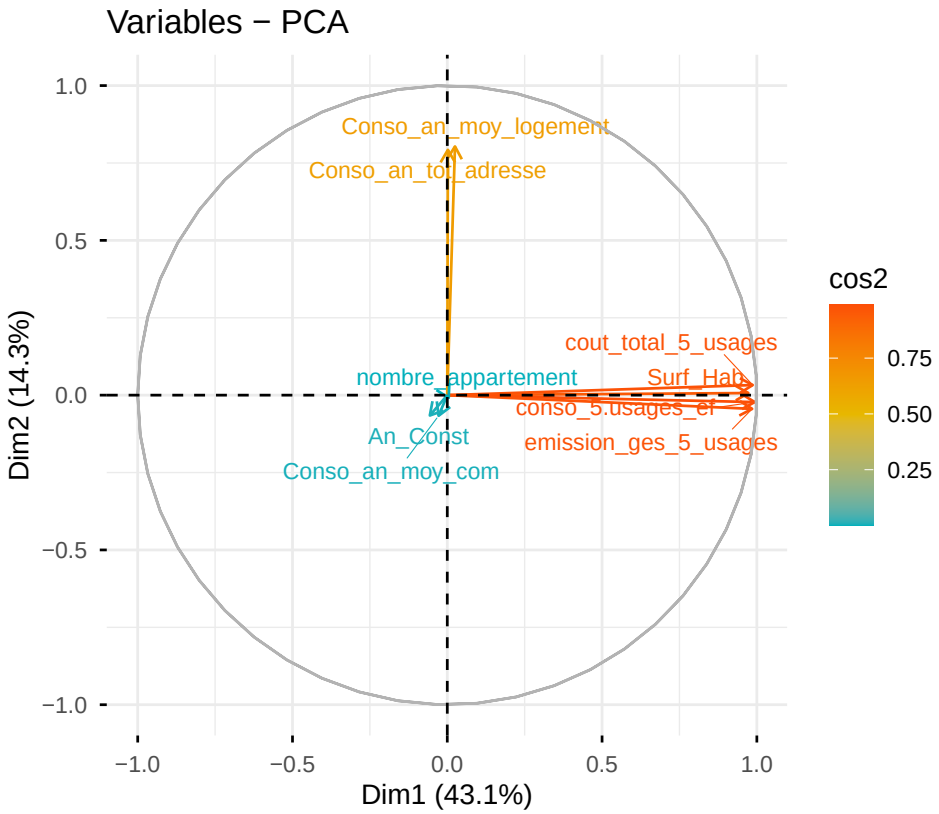

FIGURE 4 – Cercle des corrélations

- Les variables *C.totale* et *C.moyenne/adr* pointent verticalement vers le haut sur l'axe 2.
-Le centre du cercle contient les variables *nombre_appartement* ou *annee_construction*, signifiant qu'elles ne sont pas expliquées par ce premier plan mais nécessitent les axes suivants (notamment l'axe 3) pour être interprétées.

1.6 Conclusion de l'ACP:

L'ACP a permis de réduire la complexité du jeu de données initial tout en conservant l'essentiel de l'information. Les trois premiers axes factoriels identifiés offrent une interprétation claire des dimensions sous-jacentes : le gabarit et la consommation énergétique globale, l'intensité de consommation spécifique, et les effets temporels et géographiques. Ces résultats serviront de base pour les analyses ultérieures, notamment la classification des bâtiments selon leurs profils énergétiques.

2 Analyse des correspondances Factorielles

L'Analyse des Correspondances Factorielles (ACF) est une méthode statistique descriptive permettant d'étudier les relations entre deux variables qualitatives présentées sous la forme d'un tableau de contingence. Elle vise à résumer l'information contenue dans ce tableau en un nombre réduit de dimensions, tout en conservant les principales structures et oppositions présentes dans les données.

Dans le cadre de notre étude, l'AFC constitue un outil particulièrement pertinent pour analyser les liens entre les Modalités observées des variables étiquettes DPE et les étiquettes GES, (de gaz a effet de serre) afin d'étudier les correspondances entre les classes de performance énergétique et les niveaux d'émissions de gaz à effet de serre et mettre en évidence des profils ou des associations spécifiques.

2.1 Test de Chi2

Resultat test
Chi2 = 14907.6
df = 36
P_value = 0

Le test du Chi² met en évidence une association très forte entre les étiquettes DPE et les étiquettes GES. Avec une statistique de Chi² de **14907.6** pour **36 degrés de liberté** et une **p-value égale à 0**, on rejette clairement l'hypothèse d'indépendance. Autrement dit, la répartition des classes GES n'est pas due au hasard : elle varie nettement selon la classe énergétique du logement.

2.2 Tableau de contingence

Table 3 – Tableau de contingence

DPE	Étiquettes GES							Total
	A	B	C	D	E	F	G	Total
A	28	0	0	0	0	0	0	**28**
B	61	102	0	0	0	0	0	**163**
C	256	132	1664	0	0	0	0	**2052**
D	96	322	332	1716	0	0	0	**2466**
E	2	275	60	272	977	0	0	**1586**
F	1	26	123	20	111	324	0	**605**
G	0	0	93	3	7	55	98	**256**
Total	**444**	**857**	**2272**	**2011**	**1095**	**379**	**98**	**7156**

Ce tableau de contingence permet de visualiser comment les deux variables se répartissent conjointement. On observe déjà des concentrations fortes dans certaines combinaisons, laissant supposer une dépendance entre les modalités.

2.3 Qualité globale de l'ACF

Il est désormais pertinent d'examiner plus finement la structure de leur association. L'étude des inerties va nous permettre d'identifier quelles dimensions résument le mieux l'information contenue dans le tableau, et de déterminer combien d'axes sont réellement nécessaires pour interpréter les correspondances observées.

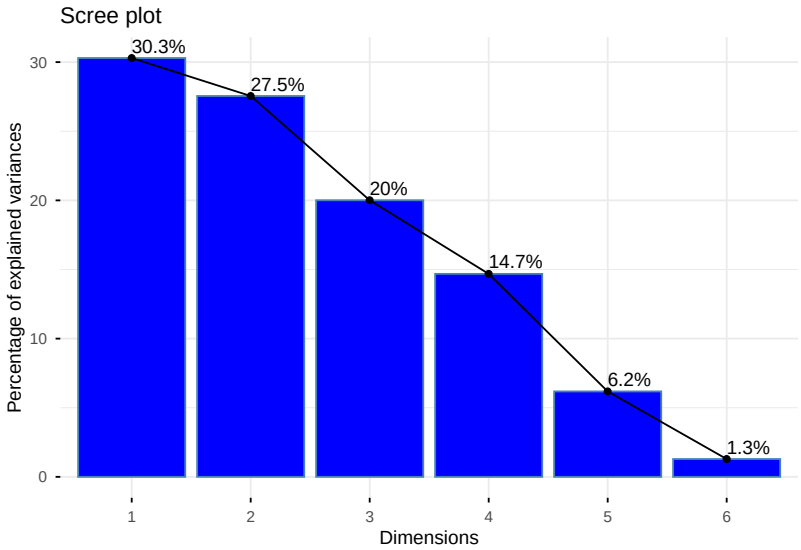

Figure 5 – Etude de la dimensionnalité

L'inertie montre que les quatre premiers axes concentrent l'essentiel de l'information, cumulant 93 % de la variance totale. Ils décrivent donc très bien la structure du tableau de contingence. Toutefois, nous choisissons de nous concentrer uniquement sur les deux premiers axes, qui résument déjà 58 % de l'inertie et révèlent les oppositions les plus structurantes. Les résultats détaillés des **dimensions 3 et 4**, ainsi que leurs représentations graphiques, sont présentés en **annexe** pour approfondir l'analyse.

2.3.1 Etude des contributions à l'axe 1

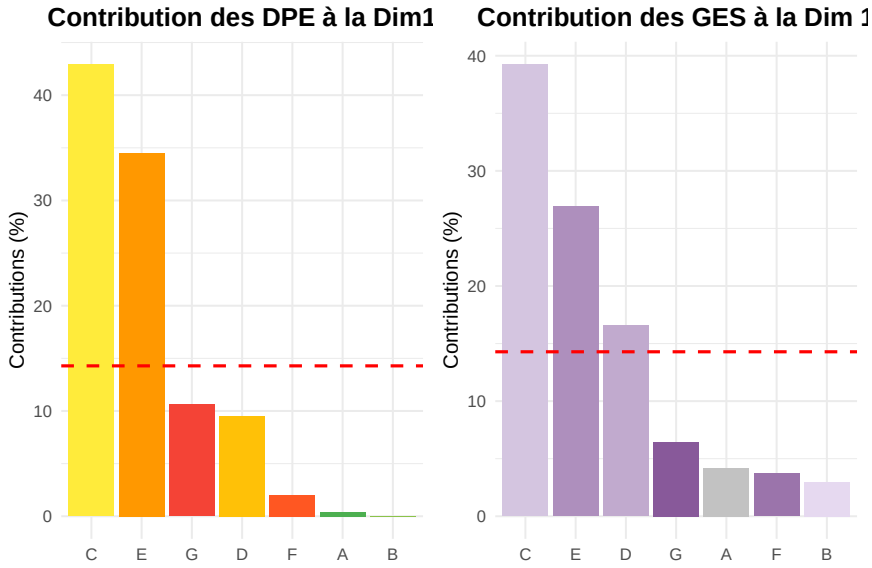

FIGURE 6 – Etude des contributions des modalités à l'axe 1

Ce graphique des contributions à l'axe 1 met en évidence une structuration très nette autour des modalités C et E, tant du côté des étiquettes DPE que GES. Ces deux classes dominent largement la construction de l'axe, avec des contributions supérieures à 30 %, ce qui indique qu'elles jouent un rôle central dans l'opposition représentée. À l'inverse, les classes A, B et F apparaissent comme marginales, avec des contributions très faibles, elles interviennent peu dans la dynamique principale captée par cet axe. Aussi peut-etre car A et B, ont des contributions très faibles à l'axe 1. Cela peut s'expliquer par leur faible fréquence dans le tableau de contingence, ce qui limite leur poids dans la construction des dimensions. À l'inverse, les modalités C et E, sont beaucoup plus représentées.

L'axe 1 reflète donc une tension entre des niveaux intermédiaires de performance énergétique et d'émissions, plutôt que les extrêmes.

2.3.2 Etude des contributions à l'axe 2

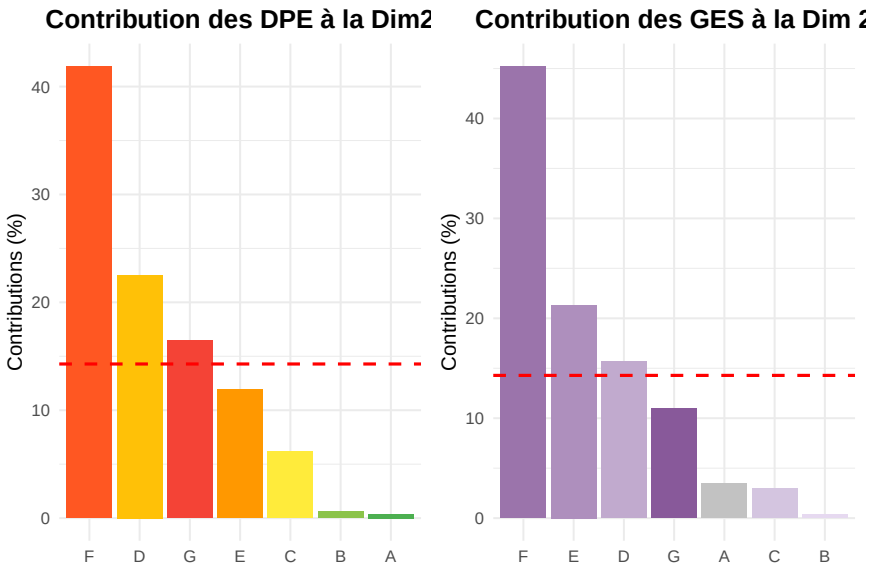

FIGURE 7 – Etude des contributions des modalités à l'axe 2

L'axe 2 est principalement structuré par les modalités F, D et G, aussi bien du côté des étiquettes DPE que GES. Ces classes, qui correspondent à des niveaux de performance énergétique et d'émissions plus élevés, contribuent fortement à la construction de cet axe. À l'inverse, les classes A, B et C sont peu représentées, ce qui peut s'expliquer par leur faible fréquence dans le tableau de contingence. Cet axe semble donc refléter une opposition entre les logements les plus polluants et les autres, en mettant en lumière les profils les plus énergivores et les plus émetteurs.

2.4 Synthèse finale

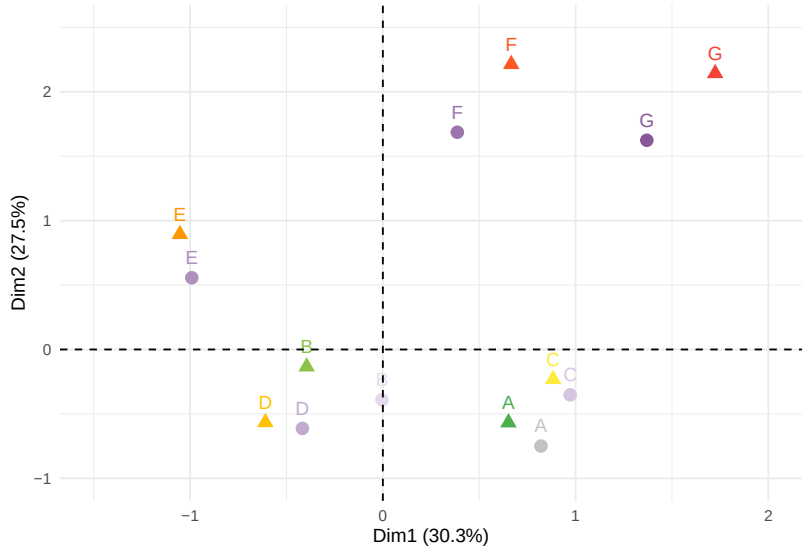

FIGURE 8 – Biplot des étiquettes DPE et GES

Les points en triangle représentent les étiquettes DPE, tandis que les points en cercle correspondent aux étiquettes GES. L'axe 1 (30.3 % de l'inertie) oppose des modalités très marquées à d'autres plus diffuses. L'axe 2 (27.5 %) met en évidence des nuances plus fines de comportement ou de positionnement. Certaines étiquettes, comme DPE-F ou GES-C, se démarquent nettement du centre, indiquant une forte contribution à la structuration de l'espace ; signe de profils atypiques ou de comportements singuliers. La proximité de certaines étiquettes rouges et bleues dans un même quadrant suggère des convergences ponctuelles entre les deux familles, bien que ces rapprochements restent minoritaires face à la séparation globale. Ce contraste illustre la diversité des profils et confirme l'intérêt de l'AFC pour révéler les logiques sous-jacentes aux données.

2.5 Conclusion de l'AFC

L'AFC met clairement en évidence une forte dépendance entre les étiquettes DPE et GES, confirmée par un Chi² très significatif. Les deux premiers axes, qui résument plus de la moitié de l'inertie totale, révèlent les oppositions majeures : l'axe 1 distingue surtout les modalités intermédiaires comme C et E, tandis que l'axe 2 isole les profils les plus énergivores et émetteurs, notamment F et G. Le biplot illustre ces contrastes et montre des regroupements cohérents entre classes énergétiques et niveaux d'émissions. Dans l'ensemble, l'AFC offre une lecture synthétique et structurée des correspondances entre performance énergétique et émissions de gaz à effet de serre.

3 Analyse en Composantes Multiples (ACM)

L'Analyse des Correspondances Factorielles (AFC) nous a permis de valider le lien direct entre la performance énergétique et les émissions de gaz à effet de serre. Cependant, pour répondre pleinement à notre problématique sur la différenciation des logements selon leur contexte de construction, une analyse bivariée ne suffit plus.

L'Analyse des Correspondances Multiples (ACM) est ici mobilisée pour traiter simultanément l'ensemble des dimensions qualitatives de notre étude. Elle va nous permettre d'observer comment les diagnostics techniques (DPE, GES) s'articulent avec les caractéristiques physiques comme la surface habitable et les périodes historiques de construction.

Alors que l'ACP se concentrait sur les volumes et les mesures quantitatives , l'ACM va révéler la structure latente de notre parc immobilier en regroupant les modalités qui définissent les profils types que nous analyserons par la suite dans la phase de clustering.

3.1 Etude des inerties

Représentation graphique des Inerties:

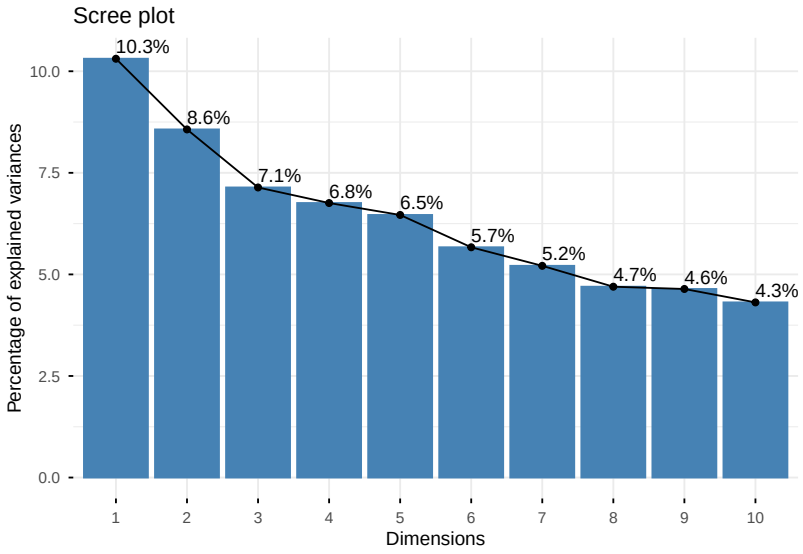

FIGURE 9 – Etude des inerties de l'ACM

Contrairement à l'ACP qui concentrait l'information sur 2-3 axes, l'ACM révèle une structure plus repartie avec 10 dimensions nécessaires pour capturer la complexité des données. Les deux premiers axes n'expliquent que 18,9% de l'inertie totale.

3.2 Etude des contributions des variables

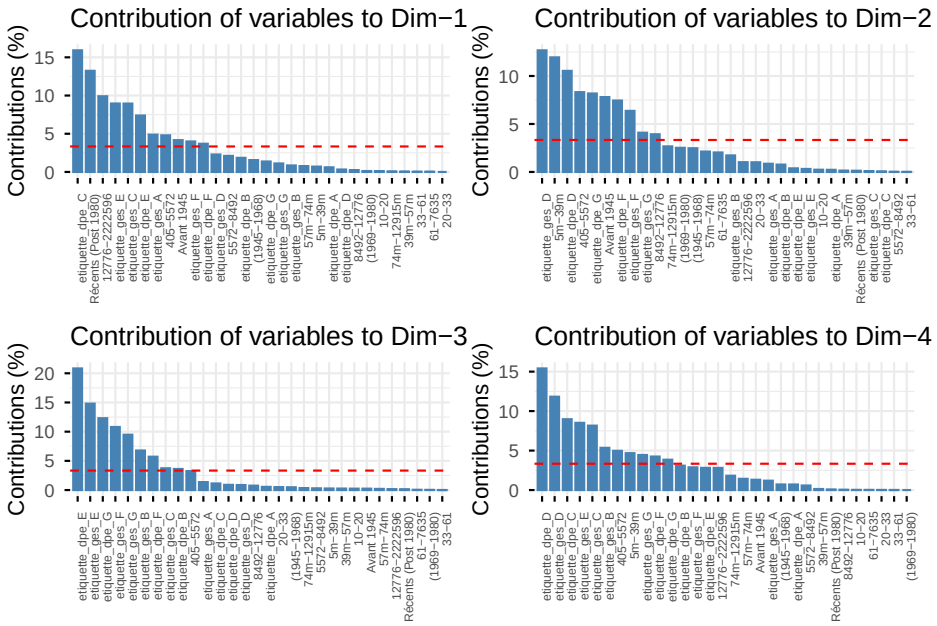

FIGURE 10 – Etude des contributions des variables aux axes de l'ACM

L'analyse des contributions des variables nous montre :

- **L'axe 1** est dominé par les logements Récents (Post-1980) et l'étiquette DPE C.On note également une contribution significative des classes intermédiaires à faibles (etiquette_ges_E, etiquette_dpe_E)
- **L'axe 2** est domminé par les logements de petites surfaces (5m-39m) , construits avant 1945 et l'étiquette centrale DPE/GES D
- **L'axe 3** est marqué par les mauvaises performances énergétiques. Il est dominé par l'étiquette E (qui a la plus forte contribution) ainsi que les étiquettes critiques G (etiquette_dpe_G, etiquette_ges_G)
- **L'axe 4** est principalement tiré par les étiquettes intermédiaires D (très forte contribution, etiquette_dpe_D) et C, représentant les logements "moyens" qui ne sont ni récents/vertueux, ni en situation de précarité énergétique extrême.

3.3 Représentation des variables

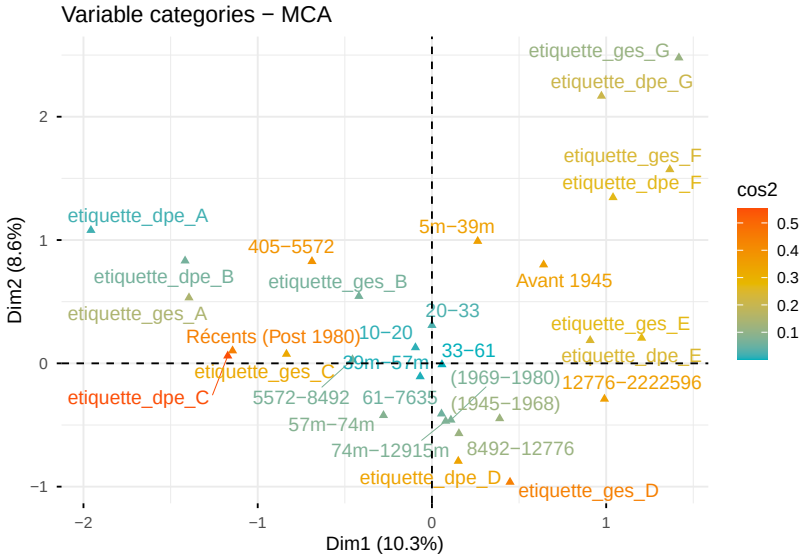

FIGURE 11 – Représentation des corrélations des variables de l'ACM

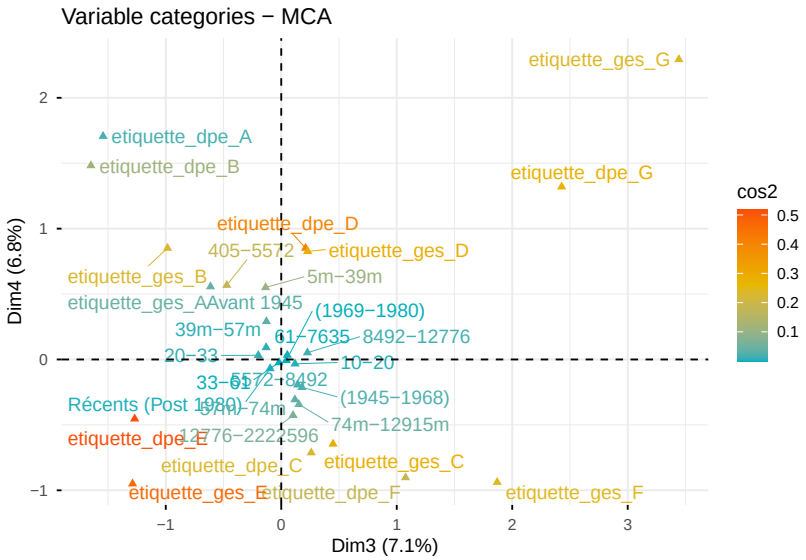

FIGURE 12 – Représentation des corrélations des variables de l'ACM

— **Analyse du premier plan (Dimensions 1 et 2)** l'axe 1 oppose nettement les logements "Recents (Post 1980)", situés à gauche et regroupés avec les bonnes étiquettes (A, B, C), aux logements construits "Avant 1945", situés à droite et associés aux étiquettes énergivores (F, G) pour le DPE et le GES Parallèlement, l'axe 2 sépare verticalement les "Passoires Critiques" (F, G en haut) des logements moyens (D , en bas).

— **Analyse du second plan (Dimensions 3 et 4)** l'axe 3 sépare les bons logements (A, B) de l'étiquette G (aussi bien en DPE qu'en GES) qui se retrouve toute seule à droite, bien à l'écart des autres. Parallèlement, l'axe 4 fait le tri dans le reste : il place les étiquettes G et D en haut du graphique, alors qu'il envoie les étiquettes (GES et DPE) F et E vers le bas, ce qui montre bien que ces groupes sont très différents et ne vont pas ensemble.

3.4 Nuage des Individus avec habillage

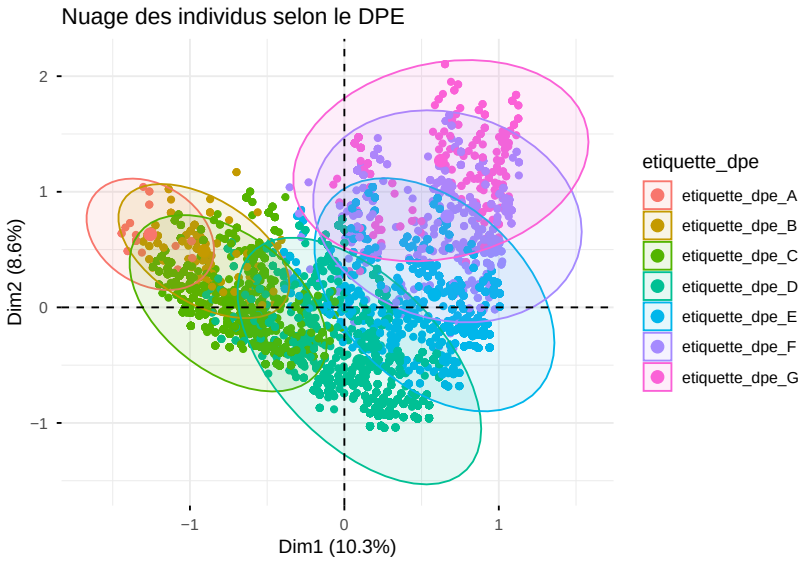

FIGURE 13 – Nuage des bâtiments selon le DPE

Ce graphique montre clairement comment les logements se séparent en fonction de leur étiquette DPE. On d'un côté, en bas à gauche, on a les logements économes (A, B, C) qui restent groupés. De l'autre, en haut à droite, on trouve les logements mal isolés (F et G) qui consomment beaucoup. Comme les groupes ne se mélangent pas, cela prouve que nos axes séparent bien les bons logements des mauvais, avec la note D qui se trouve logiquement au milieu.

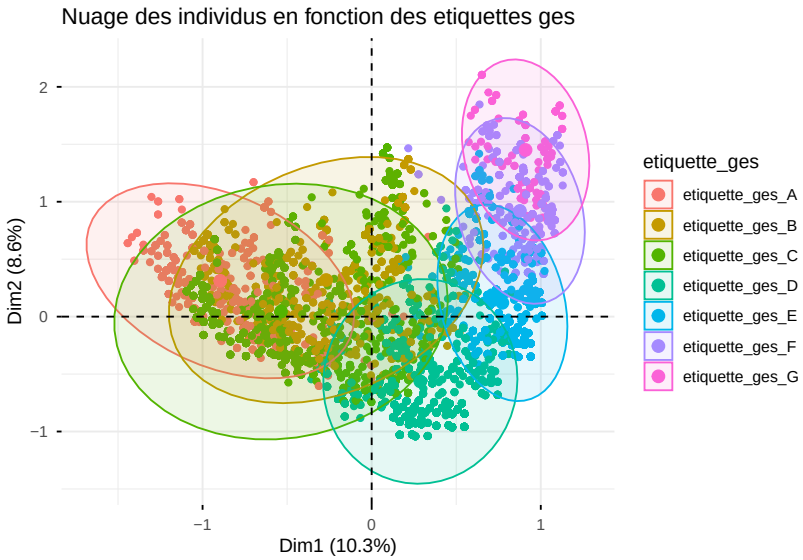

FIGURE 14 – Nuage des bâtiments selon les etiquettes GES

Ce graphique ressemble énormément au précédent, mais il concerne cette fois la pollution (les gaz à effet de serre). On voit que les logements qui polluent peu (classes A et B) sont rassemblés à gauche, alors que les logements qui polluent beaucoup (classes F et G) sont isolés tout en haut à droite. Le fait que ce dessin soit presque identique à celui du DPE prouve une chose simple : dans cette ville, les logements qui consomment le plus d'énergie sont aussi ceux qui ont le plus mauvais impact sur le climat.

3.5 Conclusion de l'ACM

En réponse à notre problématique, l'Analyse des Correspondances Multiples révèle que la diversité des constructions structure fortement la performance énergétique selon des logiques historiques précises. Au-delà de l'opposition attendue entre le neuf et l'ancien (Axe 1), l'étude met en lumière une fracture majeure au sein même de ce dernier : les axes secondaires isolent nettement les logements standards (étiquettes D) des "passoires thermiques" critiques (étiquettes G, période 1945-1968).

4 Classification non supervisée (Clustering)

Si l'ACM nous a permis de dégager les grandes dimensions structurantes de notre parc immobilier, elle ne permet pas encore de regrouper les logements de manière isolée. Afin de transformer ces tendances statistiques en profils concrets et opérationnels, nous allons maintenant procéder à une Classification Ascendante Hiérarchique (CAH) en nous basant sur les coordonnées factorielles précédemment extraites.

Dans un premier temps nous allons déterminer le nombre de groupe optimal que nous allons retenir à partir à de deux critères :

-**l'inertie intra-classe (Within cluster inertia)** qui mesure la cohésion au sein des clusters, et
-**le gain d'inertie inter-classe (Between inertia gain)** qui mesure la séparation entre les clusters.

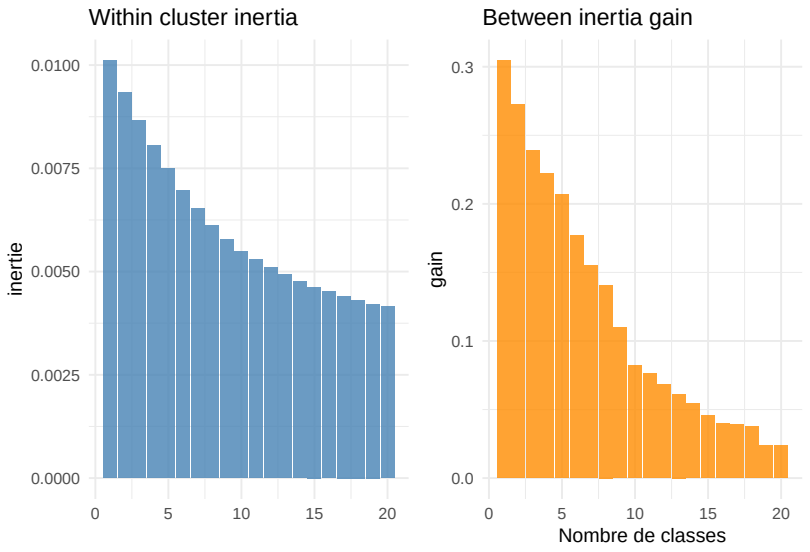

FIGURE 15 – Étude de l'inertie pour le choix du nombre de clusters

Le nombre optimal de clusters se situe à 3 car, au-delà de ce point, l'ajout de clusters n'apporte plus qu'un très faible gain d'inertie, comme l'illustre le graphique du gain d'inertie (Between inertia gain) à droite

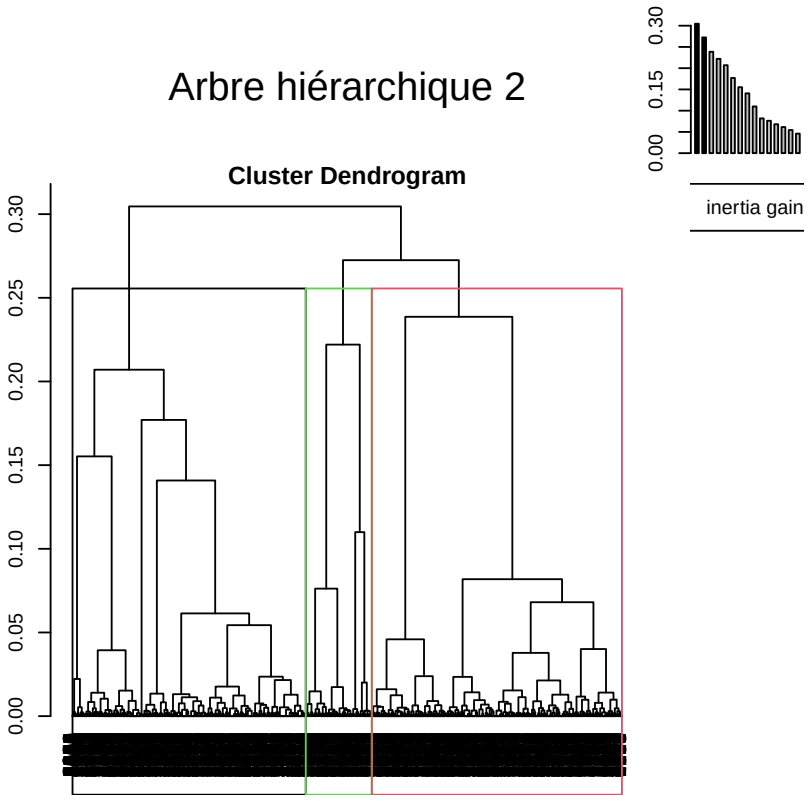

FIGURE 16 – Dendrogramme de la classification hiérarchique

L'arbre hiérarchique présenté ci-dessus permet de visualiser comment les logements s'assemblent en fonction des variables de l'étude. Il illustre la fusion progressive des logements en groupes homogènes. On y distingue nettement trois groupes majeurs, ce qui valide la cohérence de la typologie obtenue.

4.1 Paragons

Les parangons correspondent aux logements les plus proches du centre du cluster.Ce sont les logements les plus representatifs pour chaque cluster

	Conso Tot. (MWh)	Conso Moy. Log.	Conso Com.	DPE	GES	Année	Surface	Conso 5 us.	Emiss. GES	Coût Tot.	Nb Apparts
Cluster 1											
424	14.690	1.469	5.339	C	C	2016	67.7	4239.6	903.1	488.3	11
773	37.063	2.316	3.735	C	C	2013	61.8	5314.0	1150.1	432.2	18
1075	17.504	1.250	4.043	C	C	2017	59.0	4056.5	814.0	483.2	14
1466	47.477	2.064	4.095	C	C	2015	64.7	4965.3	1051.3	636.0	19
1747	25.830	1.614	5.339	C	C	2016	66.5	4193.2	880.8	645.2	16
Cluster 2											
93	303.404	1.744	2.631	D	D	1968	82.4	17812.7	3300.2	1486.0	320
286	78.914	1.338	2.631	D	D	1960	108.0	20278.7	3767.0	1675.0	66
360	101.560	1.319	2.631	D	D	1948	82.8	19058.0	3395.2	1710.0	260
555	63.830	1.303	2.631	D	D	1964	90.0	15208.0	2813.7	1276.0	200
743	30.679	1.805	2.855	D	D	1953	75.4	14856.3	3231.2	1811.2	74
Cluster 3											
3991	91.687	2.620	2.631	F	F	1900	37.0	12951.4	2853.4	1188.0	14
622	53.473	1.782	2.941	F	F	1929	37.7	13205.2	2733.6	1099.0	25
1631	39.976	1.599	2.631	F	F	1925	36.0	12988.4	2881.4	864.9	25
3334	13.177	0.659	2.631	F	F	1900	35.3	14555.5	3261.9	966.0	44
5556	53.827	2.153	2.631	F	F	1945	34.4	13459.2	2993.7	1525.0	50

Sur le grand tableau dans l'annexe , nous remarquons que : - **Le cluster 1** regroupe des logements récents (majoritairement **2017-2018**) et performants. Ces logements representent le standard actuel car ils bénéficient d'une isolation moderne et d'étiquettes énergétiques valorisantes (C). _ **Le cluster 2** regroupe des logements plus grands (entre 75m² et 108m²), construits apres la seconde Guerre mondiale et caractérisés par une performance moyenne (étiquettes D).

— **Le cluster 3** regroupe des logements datant de 1900 à 1929 caracterisés par de petites surfaces autour de (36 à 37 mètres) classées F tant pour le DPE que pour le GES.

4.2 Caractérisation des Classes

Table 5: Caractéristiques du cluster 1

	Cla/Mod	Mod/Cla	Global	p.value	v.test
annee_construction=Récents (Post 1980)	83.1	48.8	25.0	0.000000e+00	Inf
etiquette_ges=etiquette_ges_C	81.0	60.5	31.7	0.000000e+00	Inf
etiquette_dpe=etiquette_dpe_C	100.0	67.4	28.7	0.000000e+00	Inf
conso_5.usages_ef=405-5572	76.7	45.1	25.0	0.000000e+00	34.08
etiquette_ges=etiquette_ges_B	90.8	25.6	12.0	0.000000e+00	31.64
etiquette_ges=etiquette_ges_A	95.9	14.0	6.2	6.684885e-139	25.09

Le tableau ci-dessus représente les différentes caractéristiques du Cluster 1:
— 100% des logements avec une modalité etiquette_dpe_C, B et A sont dans ce premier cluster.
— On a également 83.1% des logements qui sont des logements Récents (Post 1980), et cette modalité représente 48.8% de la composition du cluster.
— De même, 76.7% des logements ayant une modalité 405-5572 (faible consommation) sont présents dans ce cluster.

Table 6: Caractéristiques du cluster 2

	Cla/Mod	Mod/Cla	Global	p.value	v.test
etiquette_ges=etiquette_ges_D	99.1	61.1	28.1	0.000000e+00	Inf
etiquette_dpe=etiquette_dpe_D	78.5	59.4	34.5	0.000000e+00	Inf
etiquette_dpe=etiquette_dpe_E	83.2	40.5	22.2	0.000000e+00	35.02

etiquette_ges=etiquette_ges_E	89.2	30.0	15.3	0.000000e+00	32.91
conso_5.usages_ef=12776-2222596	75.0	41.1	25.0	7.228784e-187	29.15
conso_5.usages_ef=8492-12776	62.7	34.4	25.0	4.122619e-63	16.77

Ce Tableau représente les différentes caractéristiques du Cluster 2:
— 99.1% des logements avec une modalité etiquette_ges_D sont dans ce deuxième cluster, ainsi que 83.2% des individus ayant l'étiquette E.
— On a également 75.0% des logements qui ont une consommation très élevée (12776-222596), et cette modalité représente 41.1% de la composition du cluster.
— De même, 61.8% des individus construits durant la période 1945-1968 sont présents dans ce cluster, souvent associés à de grands ensembles (61-7635 appartements).

Table 7: Caractéristiques du cluster 3

	Cla/Mod	Mod/Cla	Global	p.value	v.test
etiquette_ges=etiquette_ges_F	100.0	44.6	5.3	0.000000e+00	Inf
etiquette_dpe=etiquette_dpe_F	98.2	69.9	8.5	0.000000e+00	Inf
etiquette_dpe=etiquette_dpe_G	100.0	30.1	3.6	0.000000e+00	34.06
annee_construction=Avant 1945	27.5	58.1	25.1	1.290789e-107	22.04
etiquette_ges=etiquette_ges_G	100.0	11.5	1.4	1.224871e-93	20.53
surface_habitable_logement=5m-39m	26.4	55.6	25.0	3.578216e-93	20.48

Là on a une représentation des différentes caractéristiques du Cluster 3:
— 100% des logements avec une modalité etiquette_ges_F et etiquette_dpe_G sont dans ce troisième cluster, ainsi que 98.2% des individus ayant l'étiquette DPE F.
— On a également 58.1% des individus faisant partie du cluster qui datent de la période Avant 1945, et cette modalité caractérise la majorité du groupe.
— De même, 55.6% des individus du cluster possèdent une petite surface habitable (5m-39m), associée souvent à une précarité énergétique.

4.3 Conclusion du clustering

L'analyse a permis de segmenter l'ensemble du parc immobilier en 3 profils types très distincts, validés par les critères statistiques (le "coude" du graphique). Cette classification met en évidence une forte corrélation entre l'année de construction et la performance énergétique. Voici les trois typologies de logements:

Le Parc Moderne et Vertueux (Cluster 1) : Ce groupe représente le standard cible. Il s'agit de logements récents (majoritairement post-1980 voire très récents), bien isolés et économes, classés majoritairement A, B ou C.

Le Parc Intermédiaire d'Après-guerre (Cluster 2) : Ce sont de grands logements (souvent dans de grands ensembles) construits lors des Trente Glorieuses (1945-1968). Ils se caractérisent par une consommation élevée et une performance énergétique moyenne (étiquettes D et E).

Le Parc Ancien et Précaire (Cluster 3) : Ce groupe rassemble les logements critiques : de petites surfaces construites avant 1945. Ils cumulent ancienneté et très mauvaise isolation, correspondant aux passoires énergétiques (étiquettes F et G).

En résumé : L'étude confirme que la vétusté (l'âge) du bâtiment et la petite surface sont les deux facteurs principaux associés à une très mauvaise performance énergétique.

5 Conclusion Générale

La présente étude, structurée par une approche multidimensionnelle du parc immobilier, a permis de transformer des données brutes en un véritable outil de pilotage de la transition énergétique. L'articulation de nos analyses a

révélé une cohérence profonde dans la structure du bâti : si l'ACP a d'abord permis de hiérarchiser les variables numériques dominantes, l'AFC et l'ACM ont mis en lumière les relations de dépendance critiques entre les choix techniques (chauffage, isolation) et la performance finale. Plus précisément, nos analyses de correspondances ont prouvé que la classe DPE n'est pas une fatalité isolée, mais le résultat direct de signatures techniques historiques identifiables. Enfin, la Classification Ascendante Hiérarchique (CAH) a parachevé cette étude en segmentant le parc en groupes homogènes, permettant de passer d'un constat global à des préconisations ciblées. En isolant les "clusters" de passoires thermiques, ce rapport offre une base scientifique pour prioriser les stratégies de rénovation, optimisant ainsi les investissements vers les leviers d'isolation et de décarbonation les plus efficients. Cette expertise data-driven confirme que la maîtrise de la performance énergétique de demain repose sur une compréhension fine et typologique du parc d'aujourd'hui.

6 Annexe

6.1 Interprétation de l'axe 3

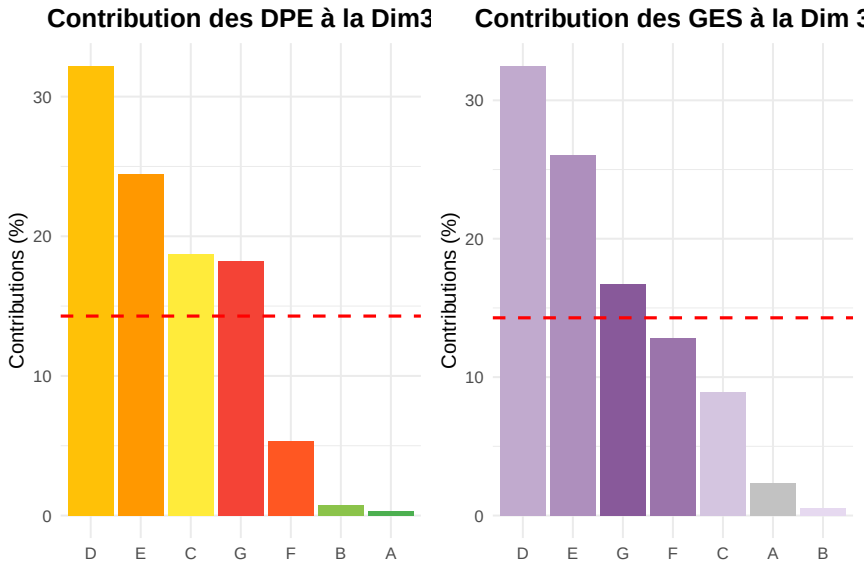

FIGURE 17 – Etude des contributions des modalités à l'axe 3

6.2 Interprétation de l'axe 4

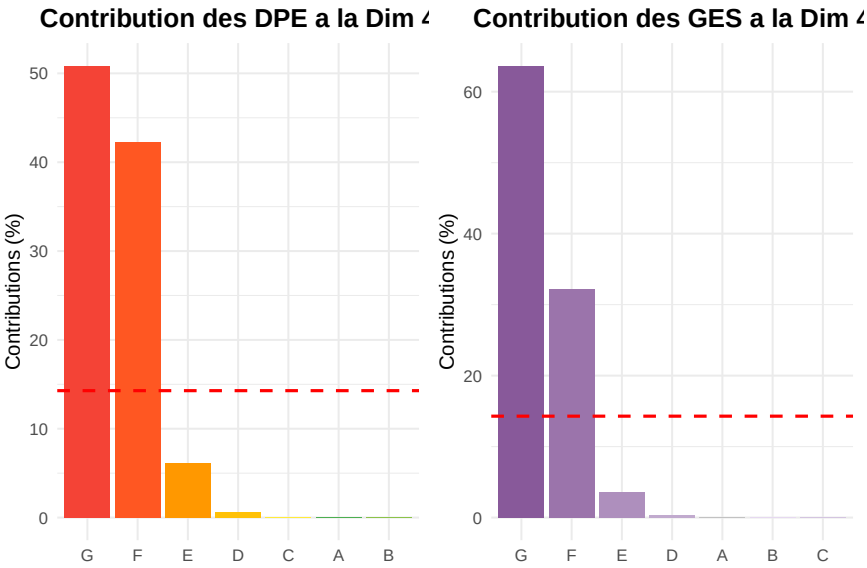

FIGURE 18 – Etude des contributions des modalités à l'axe 4

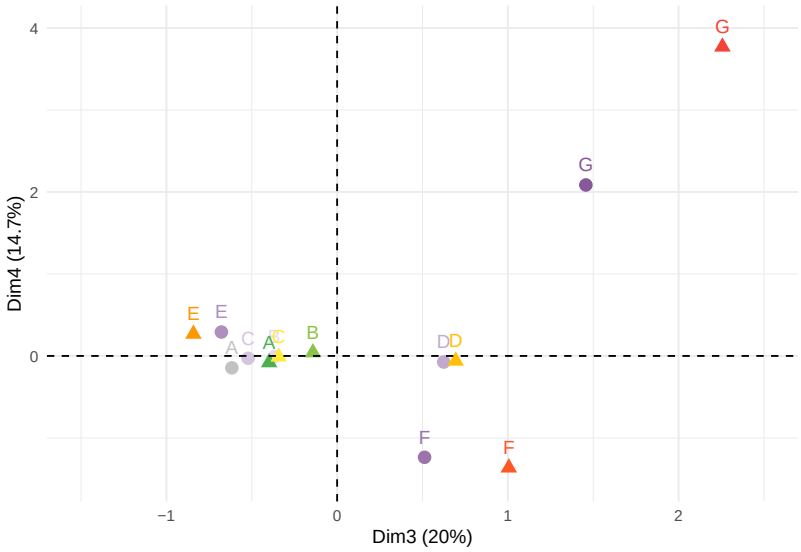

FIGURE 19 – Biplot des étiquettes DPE et GES