Classification et régionalisation

Application aux résultats des élections européennes de 2024 en France métropolitaine

ISSN 2743-8791 - Rzine.fr - Revue de méthodes pour les SHS

Claude Grasland  UMR 8504 Géographie-cités, Université Paris Cité

Date de publication : 13 août 2025

Résumé

La combinaison des méthodes de classification et de régionalisation est facilitée par le développement d'une nouvelle fonction du package `adespatial` qui permet de regrouper les unités spatiales les plus ressemblantes en respectant la contrainte de connexité. Mis au point par des écologues canadien, cette méthode de classification ascendante hiérarchique avec contrainte de contiguïté est beaucoup plus simple d'emploi et beaucoup plus efficace que les autres méthodes de régionalisation disponible actuellement dans le package Rgeoda. Elle s'appuie sur un corpus théorique d'analyse spatiale de la biodiversité des espaces animales ou végétales que l'on peut transposer à de nombreux problèmes géographiques. Nous prenons ici comme exemple l'analyse du résultat des élections européennes de 2024 en Frandce à trois niveaux d'agrégation : régions adminisratives , départements et circonscriptions législatives.

Keywords : Classification, régionnalisation

Table des matières

Liste des Figures

Liste des Tables

Introduction

Les géographes français qui sont confrontés à l'analyse d'un ensemble de variables décrivant un ensemble de lieux vont le plus souvent procéder à une analyse en deux étapes combinant analyse factorielle et classification ascendante hiérarchique (CAH). Si les variables sont hétérogènes (différentes unités de mesure) ils utiliseront une analyse en composantes principales sur le tableau des variables standardisées suivi d'une CAH en métrique euclidienne. Si les variables forment un tableau de contingence, il pourront appliquer les méthode précédentes sur un tableau de profils en ligne (standardisés ou non) ou bien opter pour le couplage entre analyse factorielle des correspondances et classification ascendante hiérarchique en métrique du chi-2. Ces approches qui s'inscrivent dans la tradition de l'analyse des données "*à la française*" ont été formalisées initialement par les travaux de Benzecri (1973), puis popularisés en géographie par l'ouvrage de Sanders (1989) et finalement mis à la disposition d'un large public grâce à l'excellent package FactomineR (Lê et al., 2008) et les publications de ses auteurs, notamment Husson et al. (2016). Les avantages du couplage entre les deux approches sont évidents puisque les méthodes factorielles permettent d'analyser d'abord les corrélations entre les colonnes du tableaux avant de procéder au regroupement des lignes à l'aide de la CAH (Husson et al., 2010).

Sans remettre en cause l'intérêt de ces approches, nous souhaiterions proposer ici une autre forme de couplage de méthodes statistiques associant **classification** et **régionalisation**, issue des travaux des écologues qui s'intéressent aux associations spatiales de plantes ou d'animaux et cherchent à en mesurer l'abondance, la spécialisation et la diffusion Legendre et De Cáceres (2013). Si le point de départ est le même (tableau croisant des lieux décrits par un ensemble de variables), les analyses vont ici surtout porter sur des tableaux homogènes décrivant soit l'abondance de différentes espèces (tableau de contingence), soit leur présence/absence (tableau disjonctif complet). Et surtout elles vont ajouter un élément supplémentaire sous la forme d'une matrice de proximité décrivant généralement la contiguïté des lieux sous la forme d'un **graphe de voisinage**. La procédure de régionalisation mise au point récemment avec la fonction `constr.hclust()`du package `adespatial` suit le schéma suivant (Figure 1) :

Il s'agit donc d'une méthode de classification ascendante hiérarchique comparable à celles réalisables en langage `R-Base` (fonction `hclust()`) ou dans `FactoMineR` (fonction `HCPC()`) mais avec deux différences importantes. D'une part, l'ajout de la contrainte de contiguïté limite les possibilités de

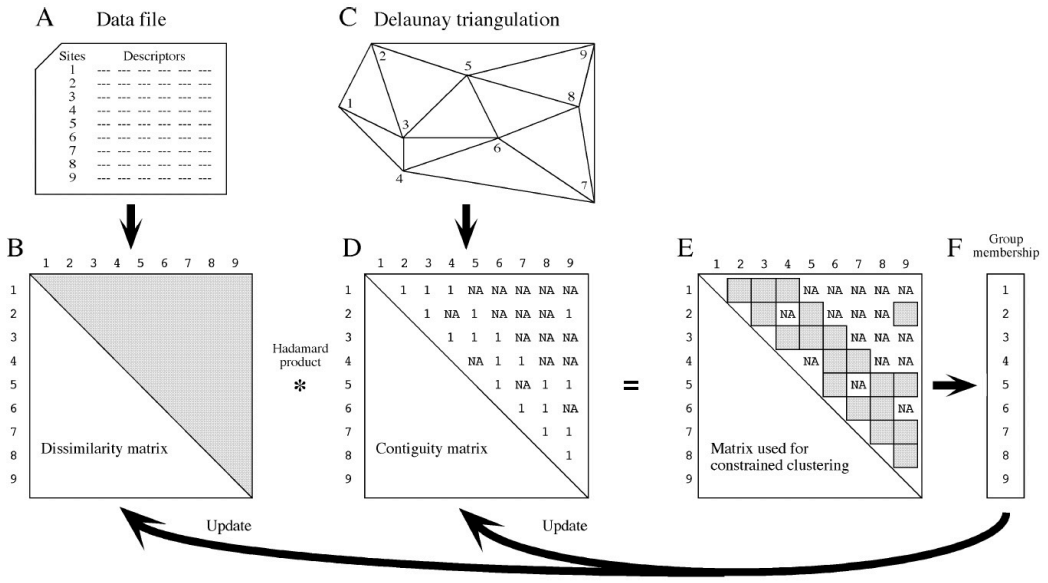


FIGURE 1 – Algorithme de régionalisation (Guénard & Legendre, 2022)

fusion des unités spatiales. D'autre part, il est possible d'utiliser un grand nombre de fonctions de dissimilarités en entrée, sans se limiter à celles qui sont privilégiées par les méthodes d'analyse factorielle à la française. Les écologues considèrent en effet que la distance euclidienne (normée ou non) et la distance du chi-2 ne sont pas toujours les plus pertinentes pour mesurer les ressemblances entre lieux, surtout si l'on considère leur caractère généralement non gaussien.

L'objectif du présent article est de discuter l'intérêt de cette procédure pour l'analyse géographique de tableaux de contingence dont les lignes sont des lieux et les colonnes des attributs dont la somme en ligne a un sens. Nous avons retenu comme exemple d'application les résultats des élections européennes de 2024 en examinant les résultats à trois niveaux d'agrégation : les régions administratives, les départements et les circonscriptions législatives (cf. partie suivante). Pour éviter une répétition, chaque niveau d'analyse sera dédié à un aspect différent du problème général de comparaison des approches de régionalisation et de classification :

- **L'échelle régionale** sera utilisé pour rappeler les principes de base des méthodes de classification et de régionalisation en insistant sur le rôle déterminant du choix de la matrice de dissimilarité utilisée en entrée.
- **L'échelle départementale** constituera le niveau privilégié de comparaison des résultats des méthodes de classification et de régionalisation afin de voir leurs apports respectifs à la compréhension du phénomène.
- **L'échelle des ciconcriptions** permettra d'examiner l'intérêt d'une approche multiscalaire et de souligner les difficultés de la régionalisation lorsque le phénomène change de nature en fonction du niveau d'agrégation.

Données utilisées

Pour faciliter la démonstration, les données brutes utilisées (**résultats des élections européennes de 2024**, fonds de carte et table de correspondance) ont été simplifiées (cf. Annexes). La mise en pratique présentée est réalisée à partir des données suivantes :

A. Résultats des élections européennes de 2024

0.0.0.1 Par régions

```
don_reg <- readRDS("data/net/don_regi.RDS")
```

regi	regi_nom	ins	vot	abs	bla	nul	exp
11	Île-de-France	7511355	3983305	3528050	38124	31292	3913889
24	Centre-Val de Loire	1847939	978080	869859	15842	17210	945028
27	Bourgogne-Franche-Comté	1997439	1115567	881872	17594	19391	1078582
28	Normandie	2420256	1294693	1125563	20612	18392	1255689
32	Hauts-de-France	4281428	2181828	2099600	31436	32891	2117501
44	Grand Est	3895241	2022271	1872970	28551	30431	1963289
52	Pays de la Loire	2871755	1527698	1344057	26204	25864	1475630
53	Bretagne	2591501	1479248	1112253	20821	23276	1435151
75	Nouvelle-Aquitaine	4525397	2517128	2008269	38236	51365	2427527
76	Occitanie	4403457	2490048	1913409	33532	39385	2417131

0.0.0.2 Par départements

```
don_dep <- readRDS("data/net/don_dept.RDS")
```

dept	dept_nom	regi	regi_nom	ins	vot	abs	bla
01	Ain	84	Auvergne-Rhône-Alpes	449217	244260	204957	3351
02	Aisne	32	Hauts-de-France	373728	189750	183978	2912
03	Allier	84	Auvergne-Rhône-Alpes	249428	138833	110595	2811
04	Alpes-de-Haute-Provence	93	Provence-Alpes-Côte d'Azur	129172	74445	54727	1026
05	Hautes-Alpes	93	Provence-Alpes-Côte d'Azur	115059	66931	48128	991
06	Alpes-Maritimes	93	Provence-Alpes-Côte d'Azur	789750	422112	367638	4134
07	Ardèche	84	Auvergne-Rhône-Alpes	259237	150502	108735	2408
08	Ardennes	44	Grand Est	186869	94450	92419	1350
09	Ariège	76	Occitanie	120494	69679	50815	1087
10	Aube	44	Grand Est	203935	109499	94436	1480

0.0.0.3 Par circonscriptions

```
don_cir <- readRDS("data/net/don_circ.RDS")
```

circ	dept	dept_nom	regi	regi_nom	ins	vot	abs	bla	nul	exp	vot1
75001	75	Paris	11	Île-de-France	85446	52564	32882	186	163	52215	1
75002	75	Paris	11	Île-de-France	75813	48127	27686	199	189	47739	0
75003	75	Paris	11	Île-de-France	73660	42912	30748	230	230	42452	0
75004	75	Paris	11	Île-de-France	73695	44104	29591	183	191	43730	0
75005	75	Paris	11	Île-de-France	80500	49242	31258	222	247	48773	0
75006	75	Paris	11	Île-de-France	81957	49464	32493	224	258	48982	0
75007	75	Paris	11	Île-de-France	82431	51466	30965	229	214	51023	0
75008	75	Paris	11	Île-de-France	84719	51007	33712	302	298	50407	1
75009	75	Paris	11	Île-de-France	71802	41153	30649	265	278	40610	0
75010	75	Paris	11	Île-de-France	71036	41371	29665	248	287	40836	2

0.0.0.4 Listes de candidats

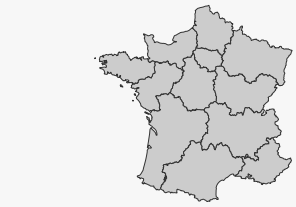
```
listes <- readRDS("data/net/don_listes.RDS")
```

tete_nom	tete_prenom	tete_sexe	tete_nais	typol	nom
DEHER-LESAINT	Léopold-Edouard	M	16/10/1947	LDIV	POUR UNE HUMANITE SOUVERAINE
PONGE	Philippe	M	21/08/1963	LDIV	POUR UNE DEMOCRATIE REELLE : DE
MARÉCHAL	Marion	F	10/12/1989	LREC	LA FRANCE FIERE, MENEE PAR MARI
AUBRY	Manon	F	22/12/1989	LFI	LA FRANCE INSOUMISE - UNION POP
BARDELLA	Jordan	M	13/09/1995	LRN	LA FRANCE REVIENT ! AVEC JORDAN
TOUSSAINT	Marie	F	27/05/1987	LVEC	EUROPE ÉCOLOGIE
AZERGUI	Nagib	M	11/11/1972	LDIV	FREE PALESTINE
THOUY	Hélène	F	23/12/1983	LDIV	PARTI ANIMALISTE - LES ANIMAUX C
TERRIEN	Olivier	M	01/11/1970	LEXG	PARTI REVOLUTIONNAIRE COMMUN
ZORN	Caroline	F	28/07/1980	LDIV	PARTI PIRATE

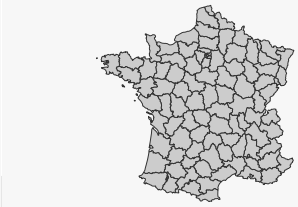
B. Fonds de carte des régions, départements et circonscriptions

```
map_regi <- readRDS("data/net/map_regi.RDS")
map_dept <- readRDS("data/net/map_dept.RDS")
map_circ <- readRDS("data/net/map_circ.RDS")
```

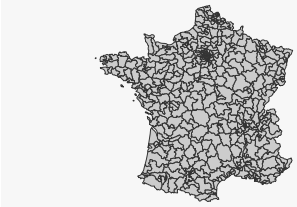
Régions



Départements



Circonscriptions



Packages nécessaires

Les packages nécessaires pour réaliser la chaîne de traitement présentée sont les suivants :

- `dplyr` pour la manipulation de données
- `ggplot2` pour la construction de graphique
- `ggrepel` pour la gestion des *labels* dans les graphiques `ggplot2`
- `sf` pour la manipulation de données géographiques vectorielles
- `mapsf` pour la construction de carte thématique
- `ineq` pour le calcul de l'indice de Gini
- `spdep` pour le calcul de matrices spatiales
- `adespatial` pour réaliser des classification à contrainte spatiale
- `cartogramR` pour construire des cartogramme (anamorphose)

Chargement des librairies :

```
# Packages utilitaires
library(dplyr)
```

```r
# Graphiques
library(ggplot2)
library(ggrepel)

# Manipulation données géographiques
library(sf)

# Cartographie
library(mapsf)
library(cartogramR)

# Statistique
library(ineq)
library(spdep)
library(adespatial)
```

1 Échelle régionale : principes de base

Afin de bien comprendre la différence entre classification et régionalisation et l'importance de la pondération, nous allons commencer par un exemple très simple portant sur la distribution des votes pour les deux principales listes d'extrême droite dans les 12 régions de France Métropolitaine.

```r
don_reg <- readRDS("data/net/don_regi.RDS")
```

On calcule le pourcentage de suffrages exprimés pour les listes conduites par Jordan Bardella (liste n°5 , RN) et Marion Maréchal (liste n°3, Reconquête) à l'échelle des 12 régions de France Métropolitaine (hors Corse).

```r
code_reg <- c("IDF", "CVDL", "BOFC", "NORM", "HDFR", "GEST",
              "PDLO", "BRET", "NAQU", "OCCI","AURA", "PACA")

don <- don_reg  |>
       mutate(Bardella = 100 * vot5 / exp,
              Marechal = 100 * vot3 / exp,
              regi_code = code_reg)  |>
       select(regi, regi_code, regi_nom, Bardella, Marechal) |>
       arrange(regi)
```

On obtient le tableau suivant :

Table 5 – Part des suffrages exprimés pour les listes Bardella et Marechal aux élections européennes de 2024 par région

regi	regi_code	regi_nom	Bardella	Marechal
11	IDF	Île-de-France	18.8	5.7
24	CVDL	Centre-Val de Loire	34.9	5.4
27	BOFC	Bourgogne-Franche-Comté	37.1	5.3
28	NORM	Normandie	35.3	4.6
32	HDFR	Hauts-de-France	42.4	4.6
44	GEST	Grand Est	38.3	5.5
52	PDLO	Pays de la Loire	27.6	4.7

regi	regi_code	regi_nom	Bardella	Marechal
53	BRET	Bretagne	25.6	4.2
75	NAQU	Nouvelle-Aquitaine	30.9	5.0
76	OCCI	Occitanie	33.7	5.5
84	AURA	Auvergne-Rhône-Alpes	30.9	5.6
93	PACA	Provence-Alpes-Côte d'Azur	38.6	7.7

1.1 Exploration des variables

1.1.1 Paramètres principaux

L'examen des paramètres statistiques des deux listes est effectué à l'intérieur des 12 régions étudiées en excluant la Corse et les DROM. Les valeurs sont donc légèrement différentes des résultats obtenus pour la France entière.

```
# Valeurs min
min <- apply(don[, 4:5], 2, min)

# Valeurs max
max <- apply(don[, 4:5], 2, max)

# Moyennes
moy <- apply(don[, 4:5], 2, mean)

# Écarts typess
ect <- apply(don[, 4:5], 2, sd)

# Variance
var <- ect^2

# coeff. variation (%)
cv <- 100 * ect / moy

# Tableau des paramètres calculés
tab <- cbind(min, max, moy, ect, var, cv)
row.names(tab) <- c("Bardella", "Marechal")

# Paramètres principaux des deux listes
print(round(tab,1))
```

```
          min  max  moy ect  var   cv
Bardella 18.8 42.4 32.9 6.5 42.7 19.9
Marechal  4.2  7.7  5.3 0.9  0.8 16.7
```

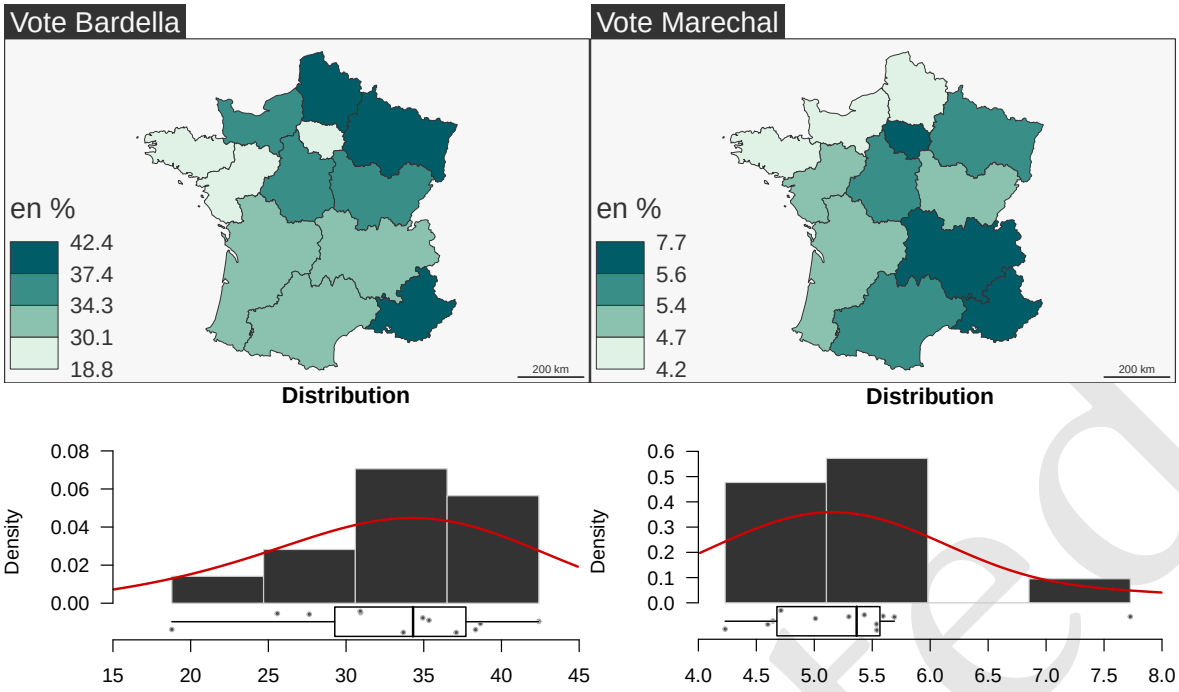
Commentaire : La liste Bardella obtient une moyenne (non pondérée) de 32.9% dans les 12 régions avec des scores allant de 18.8% en Ile-de-France à 42.7% dans les Hauts-de-France. La liste Maréchal affiche quant à elle des scores de 4.2% en Bretagne à 7.7% en PACA avec une moyenne de 5.3%.

La variation absolue des résultats, mesurée par l'écart-type est logiquement beaucoup plus forte pour Bardella ($\sigma = 6.5$) que pour Maréchal ($\sigma = 0.9$). Mais les variations relatives mesurées par le coefficient de variation (rapport entre l'écart-type et la moyenne) sont assez voisines avec 19.9% pour Bardella et 16.7% pour Maréchal.

1.1.2 Distribution spatiale

On cartographie la distribution des deux variables en quatre classes à l'aide de la méthode des quantiles (trois régions par classe) et on examine la forme des histogrammes correspondant.

```
mf_theme(cex = 1.5, line = 1.5, mar =  c(1, 1, 1.5, 1))

# Chargement du fond de carte
map <- readRDS("data/net/map_regi.RDS")

# Jointure fiond de carte et donées
mapdon <- left_join(map, don)


# Carte Bardella
mf_map(mapdon,
       type = "choro",
       var = "Bardella",
       nbreaks = 4,
       method = "quantile",
       leg_title = "en %",
       leg_val_rnd = 1,
       leg_title_cex = 1.5,
       leg_val_cex = 1.2,
       leg_size = 1.5)

mf_layout("Vote Bardella", frame = TRUE, credits = "", arrow = FALSE)
# ---

# Carte Maréchal
mf_map(mapdon,
       type = "choro",
       var = "Marechal",
       nbreaks = 4,
       method = "quantile",
       leg_title = "en %",
       leg_val_rnd = 1,
       leg_title_cex = 1.5,
       leg_val_cex = 1.2,
       leg_size = 1.5)

mf_layout("Vote Marechal", frame = TRUE, credits = "", arrow = FALSE)
# ---


# Distribution Bardella
mf_distr(don$Bardella, nbins = 4, bw = sd(don$Bardella))
# ---

# Distribution Maréchal
mf_distr(don$Marechal, nbins = 4, bw = sd(don$Marechal))
```



Commentaire : la distribution des votes Bardella est légèrement dissymétrique à droite avec une valeur exceptionellement faible correspondant à l'Ile-de-France. La distribution de Maréchal est au contraire dissymétrique à gauche avec une valeur exceptionnellement forte correspondant à la région PACA. La comparaison des deux distributions spatiales ne semble pas révéler à première vue de corrélation positive ou négative ce qui est confirmé par les coefficients de Pearson ($r = 0.20$, $p = 0.53$) ou de Spearman ($\rho = +0.03$, $p = 0.94$).

1.2 Matrices de dissimilarité

En amont d'une classification ou d'une régionalisation, la création d'une matrice de dissimlilarité entre les unités spatiales est une étape essentielle qui conditionne la suite des analyses. Deux choix essentiels interviennent alors :

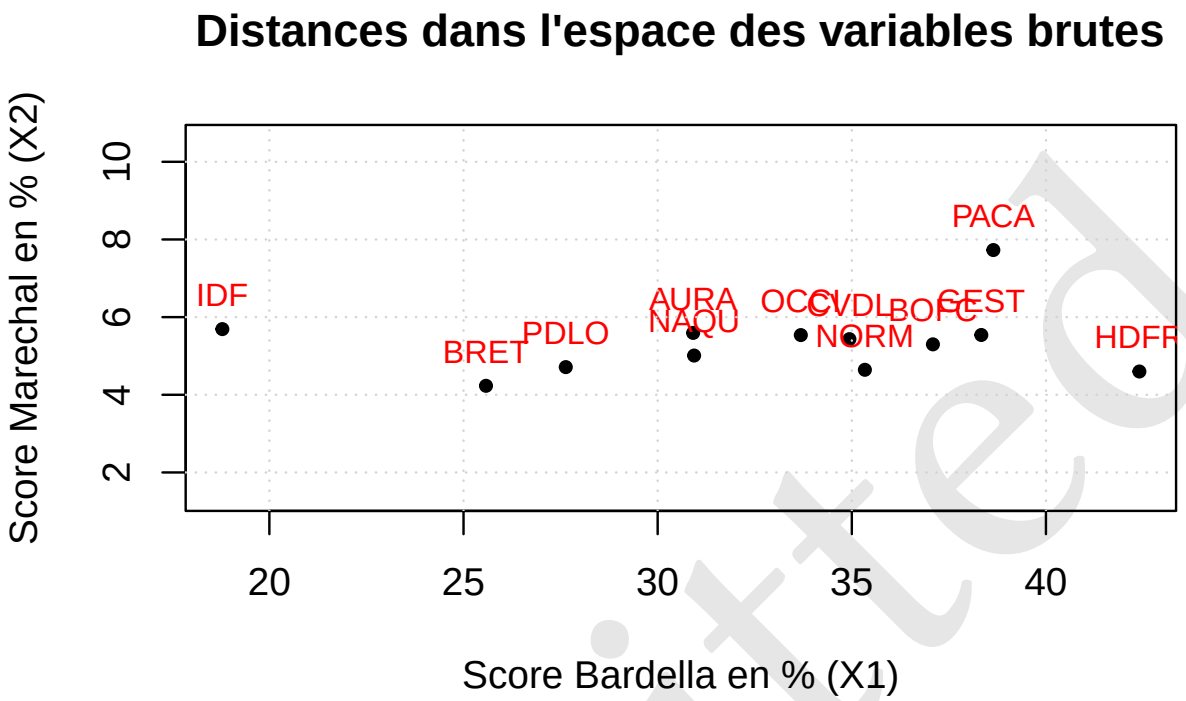
- le choix d'une transformation ou non des indicateurs
- le choix d'une métrique

1.2.1 Espace des variables brutes

La variance des scores de la variable X1 (*Bardella*) est beaucoup plus forte que celle de la variable X2 (*Maréchal*), ce qui signifie que si l'on s'en tient aux variables brutes, les différences entre régions seront liées essentiellement aux variations de la liste X1. Les différentes unités spatiales se positionneront alors dans un espace de la forme suivante :

```
plot(x = don$Bardella,
     y = don$Marechal,
     asp = 1,
     xlab = "Score Bardella en % (X1)",
     ylab = "Score Marechal en % (X2)",
     main = "Distances dans l'espace des variables brutes",
     pch = 20)
```

```
text(x = don$Bardella, y = don$Marechal, labels = don$regi_code,
     pos = 3, cex = 0.8, col = "red")

grid()
```



Commentaire : sur la figure ci-dessus, on a pris soin de construire deux axes orthonormées où une différence d'un point de pourcentage correspond à la même distance horizontalement et verticalement. Il est donc logique que la figure soit beaucoup plus étendue dans le sens horizontal que dans le sens vertical puisque le vote Bardella crée plus de différences entre les régions en valeur absolue que le vote Maréchal

On voit visuellement sur la figure précédente que les points représentant les unités spatiales sont plus ou moins éloignés, la distance qui les sépare étant une mesure de leur dissimilarité en matière de vote pour les deux listes considérées. Deux mesures de distances peuvent alors classiquement être utilisées pour convertir les positions en matrice de distance :

- la distance euclidienne : $D^{Euc}(i,j) = \sqrt{\sum_{k=1}^{K}(X_{ik} - X_{jk})^2}$
- la distance de Manhattan : $D^{Man}(i,j) = \sum_{k=1}^{K}|X_{ik} - X_{jk}|$

Les deux solutions donnant des résultats assez voisins, on se limitera ici à l'analyse de la matrice des distances euclidiennes.

```
DS_eucl <- as.matrix(dist(don[, 4:5],
                          method = "euclidean",
                          upper = TRUE,
                          diag = FALSE))

colnames(DS_eucl) <- don$regi_code
rownames(DS_eucl) <- don$regi_code
```

Matrice des distances euclidiennes :

Commentaire : La plus forte dissimilarité est observée entre la région Ile-de-France (IDF) et la

Dissimilarité en distance euclidienne brute

	IDF	CVDL	BOFC	NORM	HDFR	GEST	PDLO	BRET	NAQU	OCCI	AURA	PAC
IDF	0.0	16.2	18.3	16.6	23.6	19.5	8.9	6.9	12.2	14.9	12.1	20
CVDL	16.2	0.0	2.2	0.9	7.5	3.4	7.3	9.4	4.0	1.3	4.0	4
BOFC	18.3	2.2	0.0	1.9	5.4	1.3	9.5	11.6	6.2	3.4	6.2	2
NORM	16.6	0.9	1.9	0.0	7.1	3.1	7.7	9.8	4.4	1.9	4.5	4
HDFR	23.6	7.5	5.4	7.1	0.0	4.2	14.8	16.8	11.5	8.8	11.5	4
GEST	19.5	3.4	1.3	3.1	4.2	0.0	10.7	12.8	7.4	4.6	7.4	2
PDLO	8.9	7.3	9.5	7.7	14.8	10.7	0.0	2.1	3.3	6.1	3.4	11
BRET	6.9	9.4	11.6	9.8	16.8	12.8	2.1	0.0	5.4	8.2	5.5	13
NAQU	12.2	4.0	6.2	4.4	11.5	7.4	3.3	5.4	0.0	2.8	0.6	8
OCCI	14.9	1.3	3.4	1.9	8.8	4.6	6.1	8.2	2.8	0.0	2.8	5
AURA	12.1	4.0	6.2	4.5	11.5	7.4	3.4	5.5	0.6	2.8	0.0	8
PACA	20.0	4.4	2.9	4.5	4.9	2.2	11.4	13.5	8.2	5.4	8.0	0

région Hauts-de-France (HDFR) et la plus faible entre les régions Centre Val de Loire (CVDL) et Normandie (NORM).

En comparant la matrice de dissimilarité au graphique orthonormé précédent, on comprend que les différences entre unités spatiales sont essentiellement produites par les variations du vote Bardella qui possède une plus forte variance que le vote Maréchal. Ce dernier n'introduit que des différenciations secondaires.

1.2.2 Espace des variables standardisées

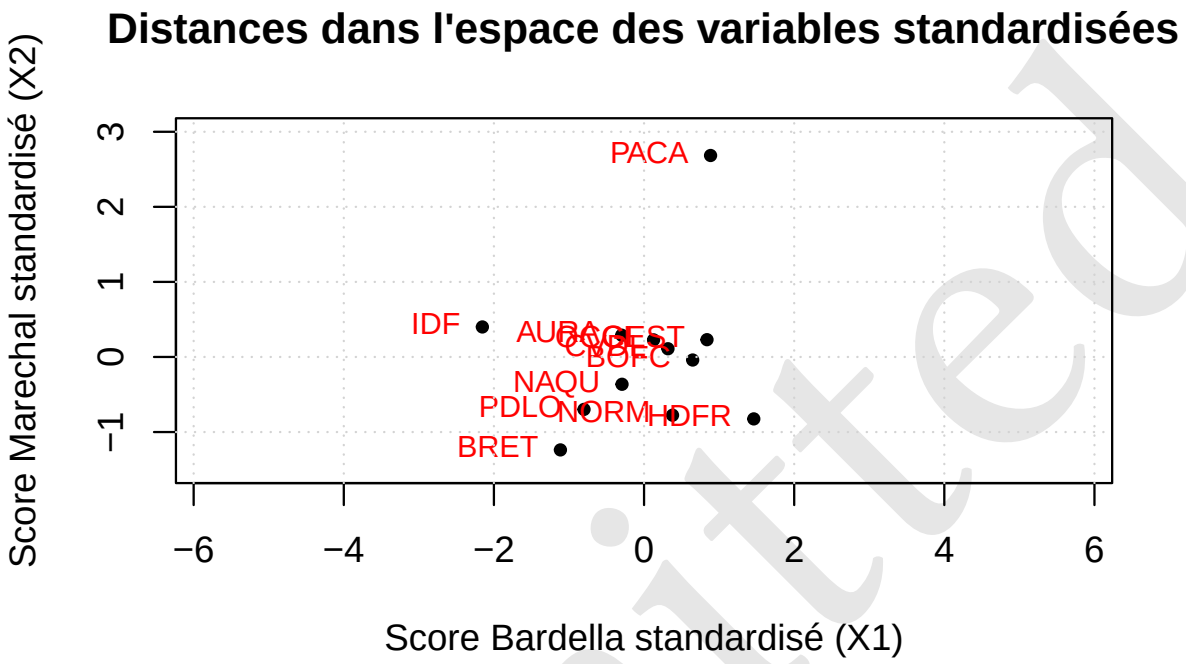
Si le choix de la métrique euclidienne ou de la métrique de Manhattan introduit peu de différences dans les matrices de dissimilarité, il en va tout autrement de la standardisation des variables qui consiste à ramener chaque indicateur à une même moyenne ($\mu = 0$) et surtout un même écart-type ($\sigma = 1$).

$$X_i^* = \frac{X_i - \mu_X}{\sigma_X}$$

Pour bien apprécier la différence, on peut commencer par visualiser les distances (donc les dissimilarités) dans l'espace des variables standardisées en adoptant comme précédemment un repère orthonormé mais dont l'unité de mesure est l'écart-type et non plus les points de pourcentage :

```
# Standardisation des deux variables
don$Bardella_std <- as.double(scale(don$Bardella))
don$Marechal_std <- as.double(scale(don$Marechal))

# Représentation graphique
plot(x = don$Bardella_std,
     y = don$Marechal_std,
     asp = 1,
     xlim = c(-2.5, 2.5),
     ylim = c(-1.5, 3),
     xlab = "Score Bardella standardisé (X1)",
     ylab = "Score Marechal standardisé (X2)",
```

```
     main = "Distances dans l'espace des variables standardisées",
     pch = 20)

text(x = don$Bardella_std, y = don$Marechal_std,
     labels = don$regi_code, pos = 2, cex = 0.8, col = "red")

grid()
```



Les distances euclidiennes dans ce nouvel espace des variables standardisées sont évidemment différentes de celles que l'on avait obtenu dans l'espace des variables brutes.

```
DS_eucl_std <- as.matrix(dist(don[, 6:7],
                         method = "euclidean",
                         upper = TRUE,
                         diag = FALSE))

colnames(DS_eucl_std) <- don$regi_code
rownames(DS_eucl_std) <- don$regi_code
```

Matrice des distances euclidiennes standardisées :

Commentaire : Par rapport à la représentation dans l'espace non-standardisé il y a désormais un étirement comparable du nuage de point dans les deux directions de l'espace standardisé. Ce résultat est logique puisque les écart-types sont désormais égaux pour les deux candidats ce qui signifie que les différences liées au vote Maréchal vont jouer le même rôle que celles liées au vote Bardella. Les deux unités spatiales les plus différentes ne sont plus l'Ile-de-France (IDF) et les Hauts-de-France (HDFR) mais la Bretagne (BRET) et la région Provence-Alpes-Côte d'Azur (PACA). Ce que l'on peut facilement vérifier en calculant la distance euclidienne sur variables standardisées.

1.2.3 Espace des variables ordinales

On pourrait transformer nos deux variables X_1 et X_2 en rang pour en faire des distributions uniformes insensibles au jeu des valeurs exceptionnelles. Si l'on effectue une transformation en rang, la géométrie de l'espace devient celle d'une grille de 12 x 12 positions en fonction des rangs obtenus

Dissimilarité en distance euclidienne standardisée

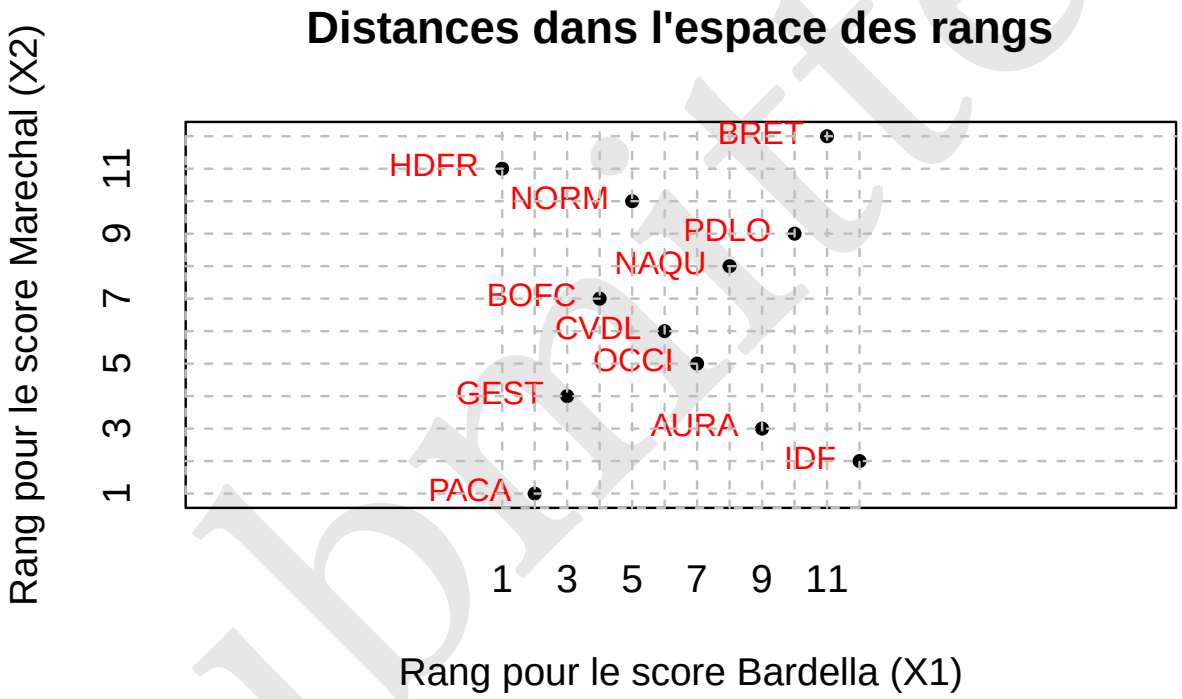
	IDF	CVDL	BOFC	NORM	HDFR	GEST	PDLO	BRET	NAQU	OCCI	AURA	PAC
IDF	0.0	2.5	2.8	2.8	3.8	3.0	1.7	1.9	2.0	2.3	1.9	3.
CVDL	2.5	0.0	0.4	0.9	1.5	0.5	1.4	2.0	0.8	0.2	0.6	2.
BOFC	2.8	0.4	0.0	0.8	1.1	0.3	1.6	2.1	1.0	0.6	1.0	2.
NORM	2.8	0.9	0.8	0.0	1.1	1.1	1.2	1.6	0.8	1.0	1.3	3.
HDFR	3.8	1.5	1.1	1.1	0.0	1.2	2.3	2.6	1.8	1.7	2.1	3.
GEST	3.0	0.5	0.3	1.1	1.2	0.0	1.9	2.4	1.3	0.7	1.1	2.
PDLO	1.7	1.4	1.6	1.2	2.3	1.9	0.0	0.6	0.6	1.3	1.1	3.
BRET	1.9	2.0	2.1	1.6	2.6	2.4	0.6	0.0	1.2	1.9	1.7	4.
NAQU	2.0	0.8	1.0	0.8	1.8	1.3	0.6	1.2	0.0	0.7	0.7	3.
OCCI	2.3	0.2	0.6	1.0	1.7	0.7	1.3	1.9	0.7	0.0	0.4	2.
AURA	1.9	0.6	1.0	1.3	2.1	1.1	1.1	1.7	0.7	0.4	0.0	2.
PACA	3.8	2.6	2.7	3.5	3.6	2.5	3.8	4.4	3.3	2.6	2.7	0.

par les unités spatiales pour le vote Bardella ou le vote Maréchal. Dans cet espace discret (sauf en cas de valeurs ex aequo) il semble logique d'utiliser la somme des différences de rang en valeur absolue, c'est-à-dire la distance de Manhattan sur les variables transformées. Cette distance correspond au plus court chemin en suivant la grille qui croise les rangs de X1 et X2 :

```
# Calcul des rangs
don$Bardella_rnk <- rank(-don$Bardella)
don$Marechal_rnk <- rank(-don$Marechal)

# Représentation graphique
plot(x = don$Bardella_rnk,
     y = don$Marechal_rnk,
     asp = 1,
     xlab = "Rang pour le score Bardella (X1)",
     ylab = "Rang pour le score Marechal (X2)",
     frame = TRUE,
     axes = FALSE,
     xlim = c(1, 12),
     ylim = c(1, 12),
     main = "Distances dans l'espace des rangs",
     pch = 20)

text(x = don$Bardella_rnk, y = don$Marechal_rnk,
   labels = don$regi_code, pos = 2, cex = 0.8, col = "red")

axis(1, at = seq(1, 12, 1), tck = 1, lty = 2, col = "gray")
axis(2, at = seq(1, 12, 1), tck = 1, lty = 2, col = "gray")
```

Dissimilarité de Manhattan sur les rangs

	IDF	CVDL	BOFC	NORM	HDFR	GEST	PDLO	BRET	NAQU	OCCI	AURA	PAC
IDF	0	10	13	15	20	11	9	11	10	8	4	1
CVDL	10	0	3	5	10	5	7	11	4	2	6	
BOFC	13	3	0	4	7	4	8	12	5	5	9	
NORM	15	5	4	0	5	8	6	8	5	7	11	1
HDFR	20	10	7	5	0	9	11	11	10	12	16	1
GEST	11	5	4	8	9	0	12	16	9	5	7	
PDLO	9	7	8	6	11	12	0	4	3	7	7	1
BRET	11	11	12	8	11	16	4	0	7	11	11	2
NAQU	10	4	5	5	10	9	3	7	0	4	6	1
OCCI	8	2	5	7	12	5	7	11	4	0	4	
AURA	4	6	9	11	16	7	7	11	6	4	0	
PACA	11	9	8	12	11	4	16	20	13	9	9	



Calcul de la matrice de distance sur les rangs :

```
DS_Man_rnk <- as.matrix(dist(don[, 8:9],
                             method = "manhattan",
                             upper = TRUE,
                             diag = FALSE))

colnames(DS_Man_rnk) <- don$regi_code
rownames(DS_Man_rnk) <- don$regi_code
```

Commentaire : On trouve désormais une distance maximale de 20 qui place à égalité la paire IDF-HDFR (plus forte distance euclidienne brute) que la paire BRET-PACA (plus forte distance euclidienne standardisée) Cette troisième solution offre donc ici une sorte de compromis entre les deux

précédentes, même si elle est en réalité plus proche de la méthode standardisée que de la méthode brute.

Il existe de nombreuses autres solutions permettant de transformer le petit tableau de données en d'autres matrices de dissimilarité tout aussi légitimes que les trois présentées ci-dessus. On pourrait par exemple utiliser une autre métrique telle que distance de Tchebychev qui est la magnitude absolue maximale des différences entre les coordonnées des points.

Le point important à retenir avant de passer à la suite des analyses est que **le choix de la matrice de dissimilarité exerce une influence cruciale sur les résultats des méthodes de classification ou de régionalisation qui vont être mise en oeuvre**. Or, ce choix est trop souvent implicite dans les logiciels de statistiques qui proposent par défaut des méthodes fondées sur la **variance** c'est-à-dire sur le **carré des distances euclidiennes standardisées**. Ce choix est le plus souvent justifié car il évite aux débutants en statistique des erreurs fatales telles que le fait de ne pas standardiser un jeu de variables hétérogènes ayant des unités de mesure et des ordres de grandeur différents. Mais il peut aussi aboutir à des résultats discutables ou du moins pas forcément les plus adaptés à la problématique.

1.3 Classification

1.3.1 Choix du critère à optimiser

Les méthodes de classification et de régionalisation ascendante hiérarchiques ont pour point commun d'opérer un regroupement des unités spatiales en allant des plus ressemblantes au moins ressemblantes. Elles fournissent un arbre de regroupement qui permet de visualiser chaque étape du regroupement et des critères permettant d'opérer un compromis entre l'homogénéité interne des classes ou régions et leur nombre.

Une bonne classification (ou une bonne régionalisation) devra comporter le moins de classes ou régions pour offrir un bon résumé. Mais également un nombre suffisant pour éviter de constituer des ensembles trop hétérogène. On utilise souvent la part de variance expliquée par la partition pour mesurer cette qualité. Mais ce choix conduit à imposer une métrique (distance euclidienne) et un algorithme (critère de Ward). Il est plus intéressant de prendre un critère plus général fondé sur le rapport entre les dissimilarités internes et externes des entités constituées. Si on s'en tient à la définition de classes ou régions homogènes comme des **groupes d'unités spatiales qui se ressemblent plus entre elles qu'elles ne ressemblent aux unités spatiales des autres groupes**, alors notre critère à optimiser H prendra une des formes suivantes :

$$H = \frac{Dissimilaritintergroupe}{Dissimilaritintragroupe}$$

ou

$$H = \frac{Dissimilaritintergroupe}{Dissimilarittotale}$$

ou

$$H = 1 - \frac{Dissimilaritintragroupe}{Dissimilarittotale}$$

1.3.2 Choix de l'algorithme de regroupement

Une classification ascendante hiérarchique peut s'opérer selon différents algorithmes qui correspondent à différents critères d'optimisation. Le critère qui semble intuitivement le plus simple est la minimisation des **distances moyennes** intra-classes et la maximisation des **distances moyennes** inter-classes. Cette méthode du *average linkage* est la plus simple à comprendre. Mais il existe beaucoup d'autres algorithmes cherchant par exemple à minimiser les distances miÉcnimales (*single linkage*), les distance maximales (*complete linkage*), les distances médianes, etc... La méthode par défaut de la plupart des logiciels de statistiques est appelée méthode de *Ward* qui consiste à minimiser la somme des distances entre les centres de gravité des classes ce qui la place l'analyse dans le cadre de l'analyse de la variance (Ward, 1963). Cette méthode comporte toutefois des variantes qui produisent des résultats différents comme cela a été démontré par Murtagh et Legendre (2014) et on distingue en pratique deux méthodes *Ward.D* et *Ward.D2* qui s'appliquent à des distances simples ou des distances élevées au carré.

Pour assurer une bonne comparabilité des résultats de classification et de régionalisation, nous utiliserons ici la fonction R-base `hclust()` (*hierarchical clustering*) plutôt que la fonction `HCPC()` du package `FactoMineR` qui est plus puissante mais introduit souvent des modifications de l'algorithme de base à l'insu de l'utilisateur non averti (notamment le fait d'optimiser a posteriori les classes par une méthode de type k-means). La régionalisation sera faite à l'aide de la fonction `constr.clust()` du package `adespatial` qui reproduit fidèlement la méthode de la fonction `hclust()` en y ajoutant simplement une contrainte de contiguïté des unités regroupées. Pour plus de détail on se reportera à la description de la classification avec contrainte de contiguïté dans Guénard et Legendre (2022).

1.3.3 Comparaison des classifications

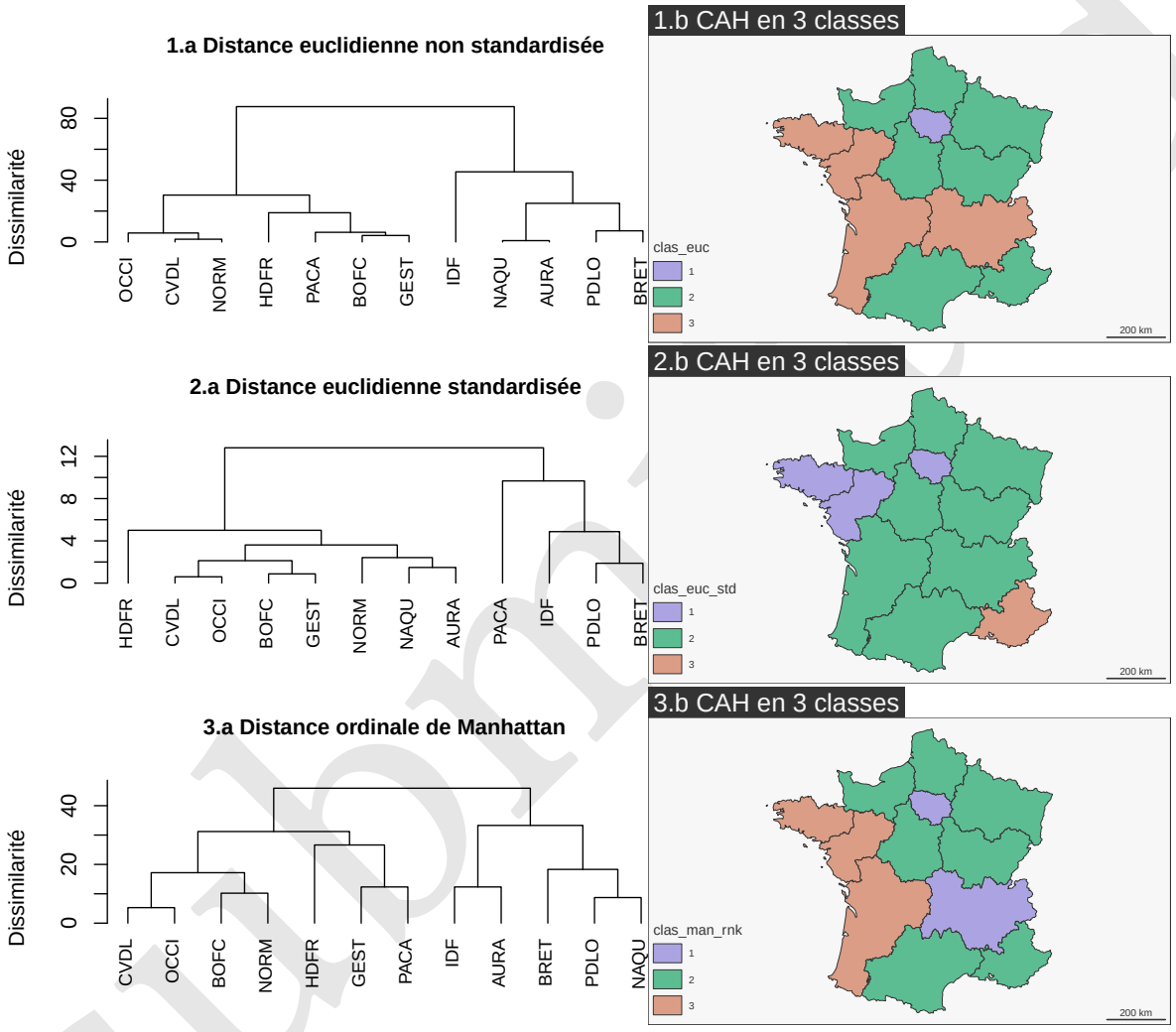
Nous allons examiner les résultats des classifications opérées sur les matrices de dissimilarité en distance euclidienne sur variables standardisées ou non standardisées et en distance de Manhtattan sur variables ordinales avec la même méthode *Ward.D*. Nous examinerons également dans chaque cas la distribution géographique des résultats pour une partition en deux classes afin de voir si les classes obtenues correspondent ou non à une régionalisation de la France

```
# CAH - Euclidienne non standardisée
cah_euc <- hclust(dist(DS_eucl), method = "ward.D")

# CAH - Euclidienne standardisée
cah_euc_std <- hclust(dist(DS_eucl_std), method = "ward.D")

# CAH - Manhattan ordinale
cah_man_rnk <- hclust(dist(DS_Man_rnk), method = "ward.D")

# Découpage en 3 classes
clas_euc <- as.factor(cutree(cah_euc, k = 3))
clas_euc_std <- as.factor(cutree(cah_euc_std, k = 3))
clas_man_rnk <- as.factor(cutree(cah_man_rnk, k = 3))

# Ajout des classifications (colonnes) dans la couche géographique
map$clas_euc  <- clas_euc
map$clas_euc_std <- clas_euc_std
map$clas_man_rnk <- clas_man_rnk
```

Représentation graphique des classifications :

```r
## 1.a Arbre CAH - Euclidienne non standardisée
plot(cah_euc,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "1.a Distance euclidienne non standardisée",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 1.b Carte - CAH - Euclidienne non standardisée
mf_map(map, type = "typo", var = "clas_euc")
mf_layout("1.b CAH en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
# ---


## 2.a Arbre CAH - Euclidienne standardisée
plot(cah_euc_std,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "2.a Distance euclidienne standardisée",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 2.b Carte - CAH - Euclidienne non standardisée
mf_map(map, type = "typo", var = "clas_euc_std")
mf_layout("2.b CAH en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
# ---


## 3.a Arbre CAH - Ordinale de Manhattan
plot(cah_man_rnk,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "3.a Distance ordinale de Manhattan",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 3.b Carte - CAH - Ordinale de Manhattan
mf_map(map, type = "typo", var = "clas_man_rnk")
mf_layout("3.b CAH en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
```

Les trois classifications aboutissent logiquement à des regroupements différents puisqu'elles sont


1.a Distance euclidienne non standardisée
Dissimilarité
0
40
80
OCCI
CVDL
NORM
HDFR
PACA
BOFC
GEST
IDF
NAQU
AURA
PDLO
BRET
1.b CAH en 3 classes
clas_euc
1
2
3
200 km
2.a Distance euclidienne standardisée
Dissimilarité
0
4
8
12
HDFR
CVDL
OCCI
BOFC
GEST
NORM
NAQU
AURA
PACA
IDF
PDLO
BRET
2.b CAH en 3 classes
clas_euc_std
1
2
3
200 km
3.a Distance ordinale de Manhattan
Dissimilarité
0
20
40
CVDL
OCCI
BOFC
NORM
HDFR
GEST
PACA
IDF
AURA
BRET
PDLO
NAQU
3.b CAH en 3 classes
clas_man_rnk
1
2
3
200 km

fondées sur des matrices de dissimilarité différentes. La région Ile-de-France ne se regroupe jamais avec les régions voisines car son score pour la liste Bardella est beaucoup plus faible et son score pour la liste Maréchal un peu plus élevé. Elle se regroupe fréquemment avec les régions de l'Ouest (Bretagne, Pays-de Loire, Aquitaine) qui se caractérisent par la faiblesse relative du vote d'extrême-droite. La région PACA se regroupe quant-à elle surtout avec sa voisine d'Occitanie avec laquelle elle partage une fort vote Bardella et Maréchal. Mais elle diffère trop de la région Auvergne-Rhône-Alpes pour former un regroupement avec les régions du Nord et de l'Est. Au total, **aucune des classifications n'aboutit à une régionalisation** c'est-dire à une division de la France en trois sous-ensembles connexes de régions voisines.

1.4 Régionalisation

La fonction `constr.hclust()` du package `adespatial` permet de réaliser une **classification ascendante hiérarchique sous contrainte de contiguïté** en suivant un algorithme strictement comparable à celui d'une classification. La seule différence réside dans le fait d'éliminer des solutions en interdisant le regroupement d'unités spatiales si elles ne sont pas voisines ou, plus précisément connexes.

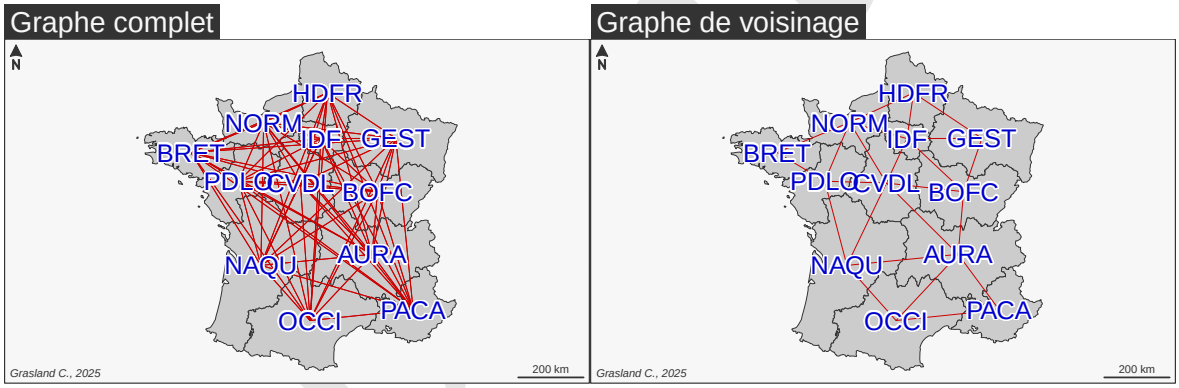
1.4.1 Graphe de proximité

Pour bien comprendre la différence entre classification et régionalisation, il est intéressant de visualiser cartographiquement les matrices de contiguïté associés à chacune des deux méthodes.

- la **classification** fait appel implicitement à un *graphe complet* qui est non planaire et dans lequel toutes les fusions d'unités spatiales en classes sont autorisées, qu'elles soient voisines ou non, connexes ou non.
- la **régionalisation** fait de son côté appel à un *graphe de contiguïté* qui est de type planaire et que l'on obtient - dans l'exemple présenté ici - en détectant les régions qui ont une frontière commune. Il est facile d'obtenir ce graphe en utilisant par exemple la fonction `st_intersects()` du package `sf`.

```
#### GRAPHE COMPLET
# Construction d'un tableau de lien (i, j) complet
reg_link_full <- expand.grid(i = mapdon$regi_code,
                             j = mapdon$regi_code,
                            stringsAsFactors = FALSE)

# Création de la couche géographique des liens
reg_links <- mf_get_links(x = mapdon,
                          df = reg_link_full,
                          x_id = "regi_code",
                          df_id = c("i","j"))

# Cartographie
mf_map(mapdon)
mf_map(reg_links, col = "red3", add = TRUE)
mf_label(mapdon, var = "regi_code", cex = 1.3, col = "blue3", halo = TRUE, bg = "white")
mf_layout("Graphe complet", frame = TRUE, credits = "Grasland C., 2025")
# ---


#### GRAPHE DE VOISINAGE (contiguité)
```

```r
# Calcul matrice de contiguïté
mat_conti <- st_intersects(mapdon, mapdon, sparse = FALSE)
colnames(mat_conti) <- mapdon$regi_code
rownames(mat_conti) <- mapdon$regi_code

# Suppression de la moitié de la matrice (et de la diagonale)
mat_conti[lower.tri(mat_conti, diag = TRUE)] <- FALSE

# Construction d'un tableau de lien (i, j) de contiguïté
reg_link_contig <- as.data.frame.table(mat_conti, responseName = "contig") |>
                              filter(contig == TRUE)

# Création de la couche géographique de liens
reg_links_contig  <- mf_get_links(x = mapdon,
                                  df = reg_link_contig,
                                  x_id = "regi_code",
                                  df_id = c("Var1","Var2"))

# Cartographie
mf_map(mapdon)
mf_map(reg_links_contig , col = "red3", add = TRUE)
mf_label(mapdon, var = "regi_code", cex = 1.3, col = "blue3", halo = TRUE, bg = "white")
mf_layout("Graphe de voisinage", frame = TRUE, credits = "Grasland C., 2025")
```



Dans les analyses de classification précédents, aucune contrainte de contiguïté spatiale n'était introduite et l'on pouvait par exemple fusionner dans une même classe la Bretagne et l'Ile-de-France qui ont des profils similaires en matière de faible vote pour les listes d'extrême-droite. Dans une analyse de régionalisation, il n'est plus possible de réunir ces deux unités spatiales sauf si on y ajoute d'autres régions les reliant telles que la Normandie ou les Pays de Loire et le Centre Val de Loire. On peut donc dire qu'**une régionalisation est une classification avec contraintes de proximité spatiale** ou, inversement, qu'**une classification est une régionalisation sans contraintes de proximité spatiale**.

Il découle de ce qui précède une conséquence fondamentale qui est le fait qu'**une régionalisation suppose un double choix en ce qui concerne la matrice de dissimilarité, d'une part, et la matrice de proximité d'autre part**. Or, si le choix de la contiguïté administrative paraît évident dans le cas étudié ici, d'autres solutions seraient possibles pour établir un graphe de proximité aboutissant à d'autres formes de régionalisation. On peut en donner rapidement deux exemples.

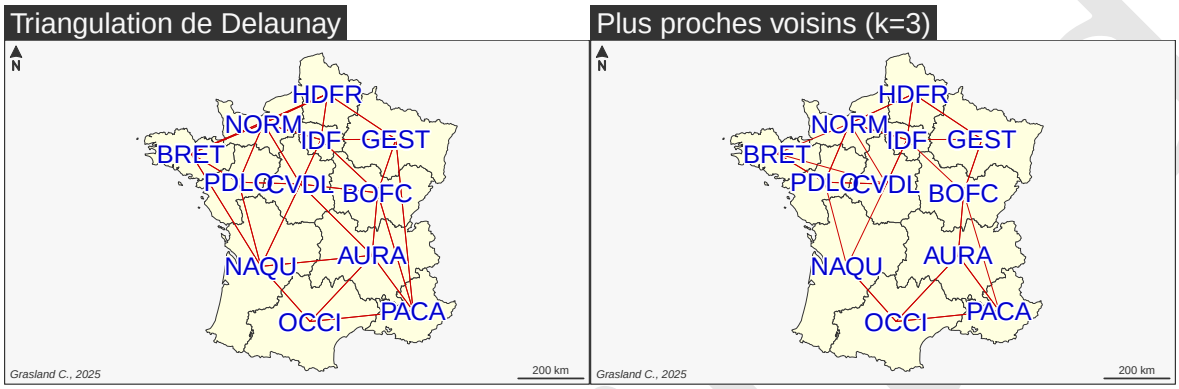
— Une **triangulation de Delaunay** pourrait par exemple être établie entre les centres des uni-

tés spatiales, qui aboutirait également à un graphe planaire mais ne respecterait pas forcément le critère de présence d'une frontière commune. On peut la réaliser facilement avec la fonction `tri2nb()` du package spdep.

— La **méthode des k plus proches voisins** pourrait également servir à déterminer pour chaque unité spatiale les k plus proches en prenant comme critère la distance à vol d'oiseau entre leurs centres. On réalise facilement le graphe à l'aide des fonctions `knearneigh()` et `knn()` du package spdep. On obtient alors un graphe non planaire mais où chaque unité spatial aurait des nombres de voisins plus proches que dans le cas du graphe de contiguïté (mais pas forcément égal).

```
#### GRAPHE DE VOISINAGE (triangulation de Delaunay)
## Matrice de contiguïté
x <- tri2nb(coords = st_coordinates(st_centroid(mapdon)))
mat_contig_delaunay <- nb2mat(x)
colnames(mat_contig_delaunay) <- mapdon$regi_code
rownames(mat_contig_delaunay) <- mapdon$regi_code

# Construction d'un tableau de lien (i, j) de contiguïté
reg_contig_delaunay <- as.data.frame.table(mat_contig_delaunay,
                                           responseName = "contig_voronoi") |>
                                      filter(contig_voronoi > 0)

# Création de la couche géographique de liens
reg_links_contig_delaunay  <- mf_get_links(x = mapdon,
                                           df = reg_contig_delaunay,
                                           x_id = "regi_code",
                                           df_id = c("Var1","Var2"))


# Cartographie
mf_map(mapdon, col="lightyellow")
mf_map(reg_links_contig_delaunay , col = "red3", add = TRUE)
mf_label(mapdon, var = "regi_code", cex = 1.3, col = "blue3", halo = TRUE, bg = "white")
mf_layout("Triangulation de Delaunay", frame = TRUE, credits = "Grasland C., 2025")
# ---


#### GRAPHE DE VOISINAGE (méthode des k plus proches voisins)
## Matrice de contiguïté
x <- knearneigh(x = st_coordinates(st_centroid(mapdon)), k = 3)
x <- knn2nb(x)
mat_contig_kvoisins <- nb2mat(x)
colnames(mat_contig_kvoisins) <- mapdon$regi_code
rownames(mat_contig_kvoisins) <- mapdon$regi_code

# Construction d'un tableau de lien (i, j) de contiguïté
mat_contig_kvoisins <- as.data.frame.table(mat_contig_kvoisins,
                                           responseName = "k_voisins") |>
                                    filter(k_voisins > 0)

# Création de la couche géographique de liens
```

```
reg_links_contig_kvoisins  <- mf_get_links(x = mapdon,
                                           df = mat_contig_kvoisins,
                                           x_id = "regi_code",
                                           df_id = c("Var1","Var2"))

# Cartographie
mf_map(mapdon, col="lightyellow")
mf_map(reg_links_contig_kvoisins , col = "red3", add = TRUE)
mf_label(mapdon, var = "regi_code", cex = 1.3, col = "blue3", halo = TRUE, bg = "white")
mf_layout("Plus proches voisins (k=3)", frame=T, credits = "Grasland C., 2025")
```



Comme on peut le voir sur les cartes ci-dessus, il est possible de produire des régionalisations avec contrainte de proximité spatiale qui ne s'appuient pas obligatoirement sur le critère de contiguïté et de présence d'une frontière commune. Dans le cas de la triangulation de Delaunay, il devient possible de regrouper par exemple la région PACA avec la région BOFC sans être obligé d'y inclure la région AURA. Inversement, dans le cas de la méthode des trois plus proches voisins il n'est plus possible de fusionner directement les régions AURA et NAQU bien qu'elles possèdent une frontière commune. Les résultats seront toujours des régionalisations dans la mesure où il existera bien une contrainte de proximité spatiale. Mais le résultat fera apparaître des groupes d'unités spatiales qui semblent disjointes sur une carte mais ne le sont pas dans le graphe de proximité choisi.

1.4.2 Régionalisation

Comme dans le cas de la classification, il existe de nombreux algorithmes possible pour regrouper les unités spatiales en cherchant à minimiser les dissimilarités intra-régionales. Nous nous limiterons ici à l'algorithme de régionalisation réalisé par la fonction `constr.hclust()` du package `adespatial` qui présente l'intérêt d'utiliser exactement les mêmes formules de calcul que la fonction `hclust()` de R-base et offre une parfaite possibilité de comparaison des résultats entre les deux approches. Pour éviter de multiplier les exemples, nous nous limiterons ici à l'analyse des régionalisations fondées sur une matrice de contiguïté, en reprenant les trois matrices de dssimilarité précédentes.

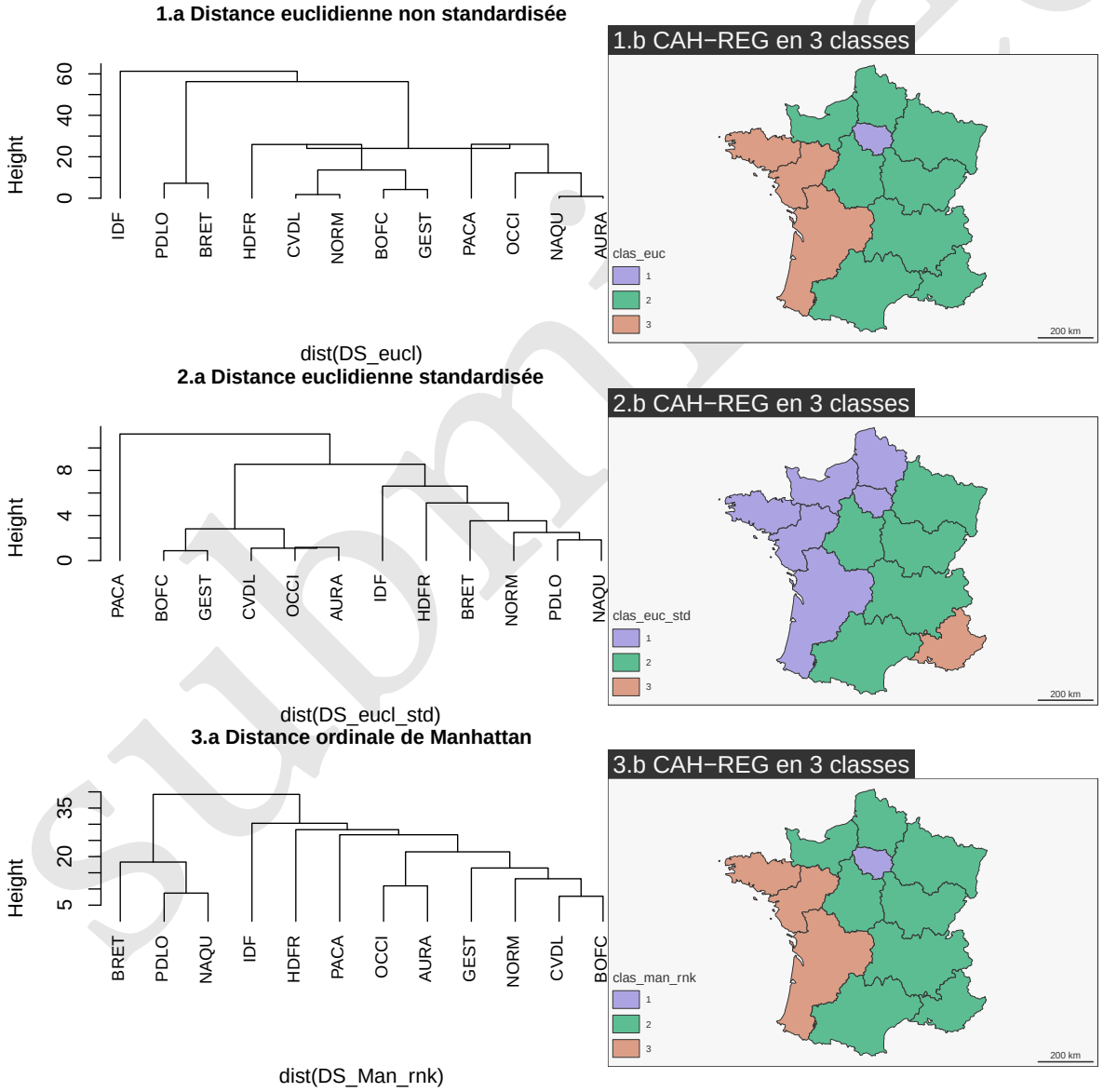
```
# CAH contrainte - Euclidienne non standardisée
reg_euc <- constr.hclust(d = dist(DS_eucl),
                         method = "ward.D",
                         links = reg_link_contig)

# CAH - Euclidienne standardisée
reg_euc_std <- constr.hclust(d = dist(DS_eucl_std),
                             method = "ward.D",
                             links =reg_link_contig)
```

```
# CAH - Manhattan ordinale
reg_man_rnk <- constr.hclust(d = dist(DS_Man_rnk),
                             method = "ward.D",
                             links =reg_link_contig)

# Découpage en 3 classes
clas_euc <- as.factor(cutree(reg_man_rnk, k=3))
clas_euc_std <- as.factor(cutree(reg_euc_std, k=3))
clas_man_rnk <- as.factor(cutree(reg_man_rnk, k=3))

# Ajout des classifications (colonnes) dans la couche géographique
map$clas_euc  <- clas_euc
map$clas_euc_std <- clas_euc_std
map$clas_man_rnk <- clas_man_rnk
```

Représentation graphique des classifications :

```
## 1.a Arbre CAH - Euclidienne non standardisée
plot(reg_euc,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "1.a Distance euclidienne non standardisée",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 1.b Carte - CAH - Euclidienne non standardisée
mf_map(map, type = "typo", var = "clas_euc")
mf_layout("1.b CAH-REG en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
# ---


## 2.a Arbre CAH - Euclidienne standardisée
plot(reg_euc_std,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "2.a Distance euclidienne standardisée",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 2.b Carte - CAH - Euclidienne non standardisée
mf_map(map, type = "typo", var = "clas_euc_std")
mf_layout("2.b CAH-REG en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
# ---
```

```r
## 3.a Arbre CAH - Ordinale de Manhattan
plot(reg_man_rnk,
     hang = -1,
     cex = 0.8,
     cex.main = 1,
     main = "3.a Distance ordinale de Manhattan",
     ylab = "Dissimilarité",
     sub = NA,
     xlab = "")
# ---

# 3.b Carte - CAH - Ordinale de Manhattan
mf_map(map, type = "typo", var = "clas_man_rnk")
mf_layout("3.b CAH-REG en 3 classes", frame = TRUE, arrow = FALSE, credits = "")
```



Comme dans le cas de la classification (cf. partie 2.4.2) on observe tout d'abord une forte variation des

résultats selon le choix de la matrice de dissimilarité. On retrouve également une tendance à l'isolement des régions IDF et PACA qui forment à nouveau du singleton puisqu'elles sont fortement différentes des autres unités spatiales et de leurs voisines en particulier. L'apport spécifique de la régionalisation consiste surtout ici à mettre en valeur la proximité des trois régions atlantiques (BRET, PDLO et NAQU) qui se regroupent du fait de leur proximité à la fois politique et spatiale. Une comparaison avec les arbres de classification précédents montre logiquement des regroupements plus tardifs du fait de l'impossibilité de rassembler certaines régions non voisines. **Une régionalisation aboutit nécessairement à des regroupements moins homogènes qu'une classification du fait des contraintes spatiales qui lui sont imposées.**

1.5 Conclusion

Au final, cet exercice souligne la complexité des options possibles du fait du nombre de choix qu'il faut opérer pour réaliser une classification et, a fortiori une régionalisation. Encore n'avons nous pas fait état de l'ensemble des solutions alternatives, notamment celles qui se fondent sur des méthodes de classification descendantes (ref. ???) ou sur des méthodes de type noyau mobile.

Mais la question la plus fondamentale est probablement la suivante : **quel est l'apport d'une régionalisation par rapport à une classification pour l'analyse d'un phénomène social** ? Puisque nous avons vu qu'une régionalisation est par définition moins efficace qu'une classification pour constituer des groupes homogènes, il faut que la prise en compte des contraintes spatiales apporte un avantage décisif à la régionalisation pour choisir de la mettre en oeuvre. Ce qui suppose que la matrice de proximité spatiale ait un sens pour la personne qui va interpréter les résultats.

C'est ce point que nous allons maintenant explorer en étudiant l'ensemble des résultats des élections européennes à trois niveaux d'agrégation.

2 Échelle départementale : classification et régionnalisation hiérarchiques

La réalisation d'une classification et d'une régionalisation des résultats des élections européennes va être menée à différentes échelles, depuis le niveau des régions jusqu'à celui des circonscriptions en passant par le liveau départemental. L'objectif sera de construire des classes ou des régions présentant des profils électoraux homogènes en matière de vote.

Préalablement à ces analyses, il est important d'analyser la distribution des votes afin de distinguer l'implantation spatiale des listes candidates au scutin afin de repérer celles qui vont le plus contribuer aux différenciations au niveau national ou au niveau local.

2.1 Analyse des listes

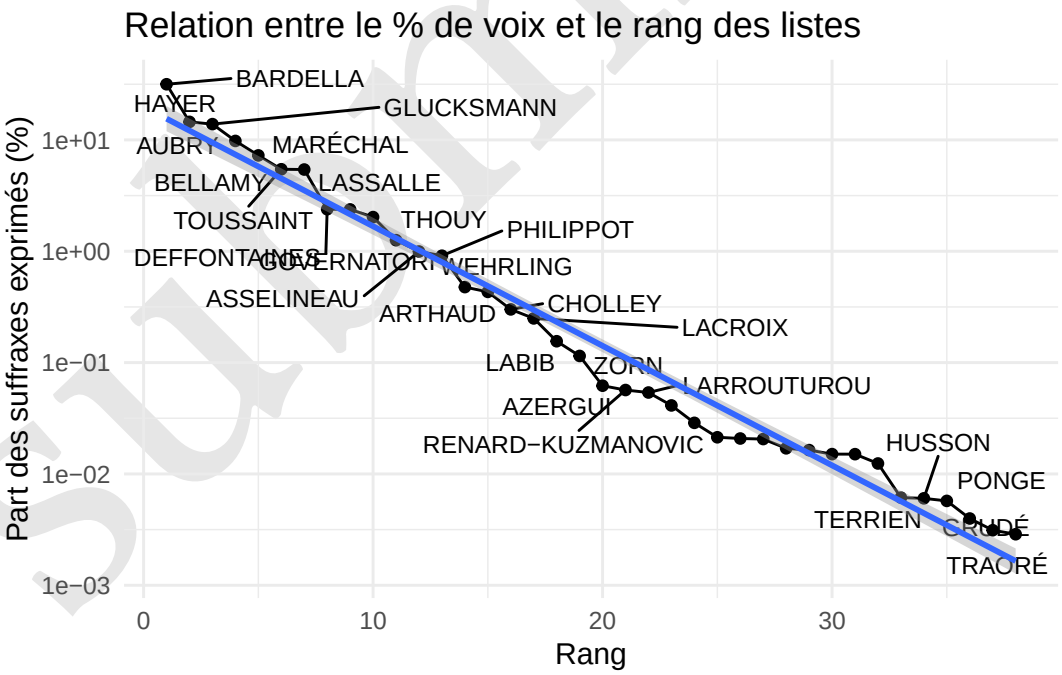
Les électeurs français ont eu le choix entre 38 listes lors des élections européennes de juin 2024 (cf données). Mais seule une partie d'entre elles a connu une audience nationale et beaucoup de petites listes n'ont même pas été capable de fournir des bulletins dans tous les bureaux de votes.

2.1.1 Loi rang-taille ?

La distribution du pourcentage de votes en fonction du rang des listes suit une loi exponentielle presque parfaite ($r^2 = 0.98$, $p < 0.001$)

```
# Chargement des métadonnées sur les listes électorales
listes <- readRDS("data/net/don_listes.RDS")
```

```r
# Chargement des résultats de vote par circonscriptions
don_cir <- readRDS("data/net/don_circ.RDS")
result_vote <- don_cir[, 12:49]

# Calcul % national de vote
listes$pct <- 100 * apply(result_vote, 2, sum) / sum(result_vote)

# Calcul indice de Gini
listes$gini <- apply(result_vote, 2, ineq::Gini)

# Calcul du rang en fonction du pourcentage de vote
listes$rang <- rank(-listes$pct)

# Régression linéaire
mod <- lm(log(listes$pct) ~ listes$rang)

# Graphique rang-taille
ggplot(listes, aes(x = rang, y = pct, label = tete_nom)) +
  geom_point() +
  geom_line() +
  geom_text_repel(cex = 3) +
  scale_x_continuous("Rang") +
  scale_y_log10("Part des suffraxes exprimés (%)") +
  geom_smooth(method = "lm") +
  theme_minimal() +
  ggtitle("Relation entre le % de voix et le rang des listes")
```



Résumé statistique de la régression linéaire :

Variable dépendante

% de votes reçus par une liste (log)

Rang de la liste

-0.247***

(0.006)

Constant

2.989***

(0.133)

Observations

38

R2

0.980

Adjusted R2

0.979

Residual Std. Error

0.402 (df = 36)

F Statistic

1,725.409*** (df = 1 ; 36)

Note :

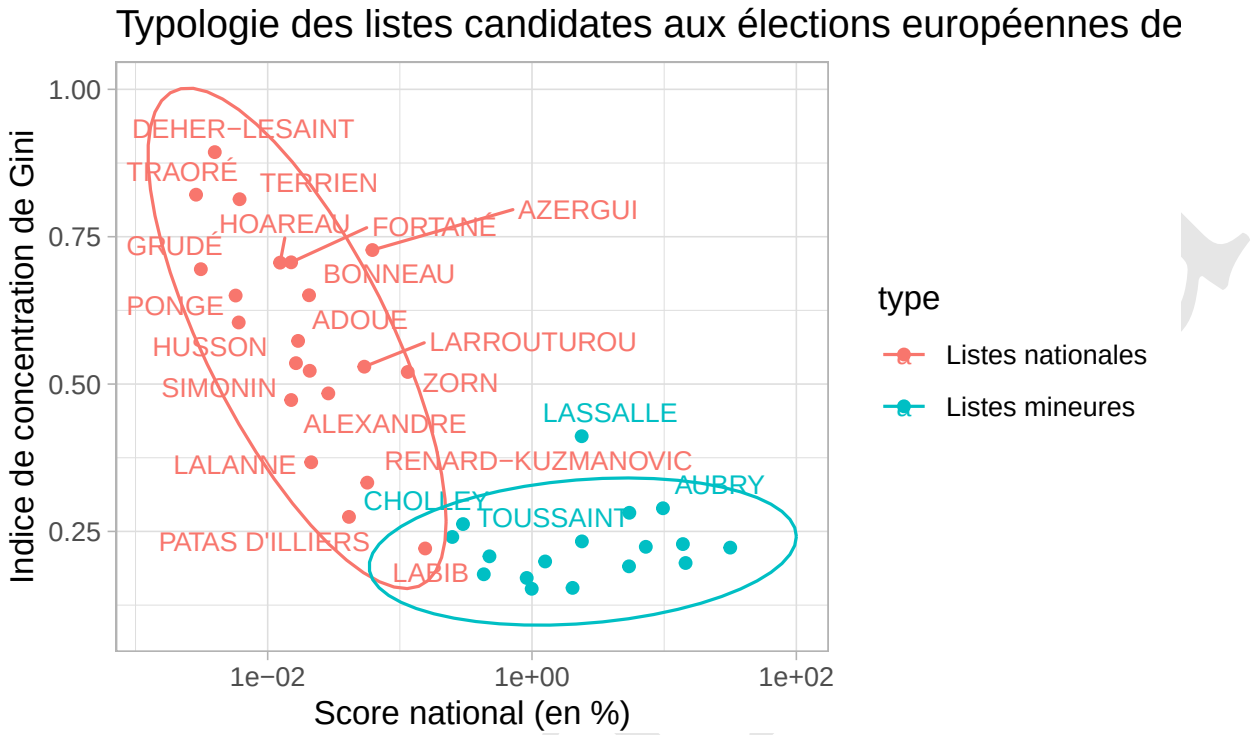
p<0.1 ; ***p<0.05 ;*** p<0.01

2.1.2 Typologie

La régularité de la loi précédente ne permet pas d'établir une rupture nette permettant de séparer grandes et petites listes. Mais une typologie combinant le logarithme du score national en % et l'indice de concentration de Gini par circonscription permet de mieux distinguer des listes mineures ayant obtenu des votes dans un petit nombre de circonscription et des listes d'audience nationale ayant obtenu des voix dans un nombre plus important de circonscriptions même lorsque leur score est faible.

```
# Passage en logarithme du % national et l'indice de Gini
tab_log <- cbind(log(listes$pct), listes$gini)

# Classification par la méthode des k-means
w <- kmeans(tab_log, centers = 2, iter.max = 1000)

# Récupération de la classification (colonne "type")
listes$type <- as.factor(w$cluster)
levels(listes$type) <- c("Listes nationales", "Listes mineures")

# Représentation graphique de la classification
ggplot(listes, aes(x = pct, y = gini, label = tete_nom, colour = type)) +
  geom_point() +
```

```
  geom_text_repel(cex = 3) +
  stat_ellipse() +
  scale_x_log10("Score national (en %)") +
  scale_y_continuous("Indice de concentration de Gini") +
  theme_light() +
  ggtitle("Typologie des listes candidates aux élections européennes de juin 2024")
```



Il existe une corrélation négative entre le score national d'une liste et sa concentration mesurée par l'indice de Gini. Les listes les plus importantes sont en général celles qui sont le mieux réparties tandis que les petites listes ont en général concentrés les suffrages dans quelques circonscriptions. Cette règle connaît toutefois des exceptions. Ainsi la liste "*Alliance rurale*" conduite par Jean Lassalle, bien implanté dans le Sud-Ouest, a obtenu un score national assez élevé (2.4%) tout en affichant un indice de concentration assez fort (0.41). Inversement, la liste du parti NPA "*Pour un Monde sans frontières ni patrons …*" conduite par Selma Labib n'a recueilli que très peu de voix (0.16%) mais beaucoup mieux réparties dans un nombre important de circonscriptions avec un indice de concentration faible (0.16) comparable à celui des listes les plus importantes.

2.2 Classification

2.2.1 Choix de la matrice de dissimilarité

On choisit comme matrice de dissimilarité le coefficient de divergence c'est-à-dire la part des électeurs qui devrait changer de votes pour que les deux unités spatiales affichent le même profil électoral. Cet indice correspond à la moitié de la distance de Manhattan entre les profils en pourcentage :

$$\frac{1}{2}\sum_{p=1}^{38}\left|\frac{X_{ip}}{X_{i.}}-\frac{X_{jp}}{X_{j.}}\right|$$

On peut illustrer le calcul en prenant l'exemple de la plus forte dissimilarité qui est observée entre le département de l'Aisne (02) et le département de Paris (75) :

```r
# Chargement des résultats par département et d'un fond de carte
map_dept <- readRDS("data/net/map_dept.RDS")
don_dept <- readRDS("data/net/don_dept.RDS")

# Calcul répartition % de vote par département pour chaque liste
result_vote_dep <- don_dept[, 11:48]
mat_vote_dep <- 100 * result_vote_dep / apply(result_vote_dep, 1, sum)
rownames(mat_vote_dep) <- don_dept$dept
colnames(mat_vote_dep) <- listes$tete_nom

# Calcul des différences de entre le l'Aisne et Paris
tab_diff <- data.frame(t(mat_vote_dep[c("02", "75"), ]))
names(tab_diff) <- c("Aisne (02)", "Paris (75)")
tab_diff$dif <- tab_diff[, 1] - tab_diff[, 2]
# Différence absolue
tab_diff$difabs <- abs(tab_diff$dif)

# Calcul des totaux
tab_diff <- rbind(tab_diff, apply(tab_diff, 2, sum))
row.names(tab_diff)[nrow(tab_diff)] <- "Total"

# Affichage de la table
kable(tab_diff, digits = 1 )
```

	Aisne (02)	Paris (75)	dif	difabs
DEHER-LESAINT	0.0	0.0	0.0	0.0
PONGE	0.0	0.0	0.0	0.0
MARÉCHAL	5.0	5.9	-0.9	0.9
AUBRY	5.3	16.8	-11.5	11.5
BARDELLA	50.6	8.5	42.1	42.1
TOUSSAINT	2.4	10.7	-8.3	8.3
AZERGUI	0.0	0.0	0.0	0.0
THOUY	2.4	1.2	1.2	1.2
TERRIEN	0.0	0.0	0.0	0.0
ZORN	0.1	0.4	-0.3	0.3
HAYER	11.3	17.7	-6.4	6.4
ALEXANDRE	0.0	0.0	0.0	0.0
CHOLLEY	0.2	0.4	-0.3	0.3
WEHRLING	0.3	0.3	0.0	0.0
ASSELINEAU	0.9	0.8	0.1	0.1
SIMONIN	0.0	0.0	0.0	0.0
FORTANÉ	0.0	0.0	0.0	0.0
BELLAMY	6.2	10.5	-4.2	4.2
ARTHAUD	0.7	0.3	0.5	0.5
LARROUTUROU	0.0	0.1	-0.1	0.1
RENARD-KUZMANOVIC	0.1	0.1	0.0	0.0
LABIB	0.1	0.1	0.0	0.0
ADOUE	0.0	0.0	0.0	0.0
PHILIPPOT	0.9	0.6	0.3	0.3

	Aisne (02)	Paris (75)	dif	difabs
HUSSON	0.0	0.0	0.0	0.0
BONNEAU	0.0	0.0	0.0	0.0
GLUCKSMANN	7.8	22.9	-15.1	15.1
HOAREAU	0.0	0.0	0.0	0.0
LASSALLE	2.2	0.4	1.8	1.8
LALANNE	0.0	0.0	0.0	0.0
LACROIX	0.2	0.2	0.0	0.0
ELMAYAN	0.0	0.0	0.0	0.0
DEFFONTAINES	2.3	1.4	0.9	0.9
COSTE-MEUNIER	0.0	0.0	0.0	0.0
GOVERNATORI	0.8	0.6	0.2	0.2
TRAORÉ	0.0	0.0	0.0	0.0
PATAS D'ILLIERS	0.0	0.0	0.0	0.0
GRUDÉ	0.0	0.0	0.0	0.0
Total	100.0	100.0	0.0	94.5

```
# Calcul matrice de dissimilarité
dissim <- dist.ldc(mat_vote_dep, method = "manhattan") / 2
```

```
Info -- For this coefficient, sqrt(D) would be Euclidean
Info -- This coefficient does not have an upper bound (no fixed D.max)
```

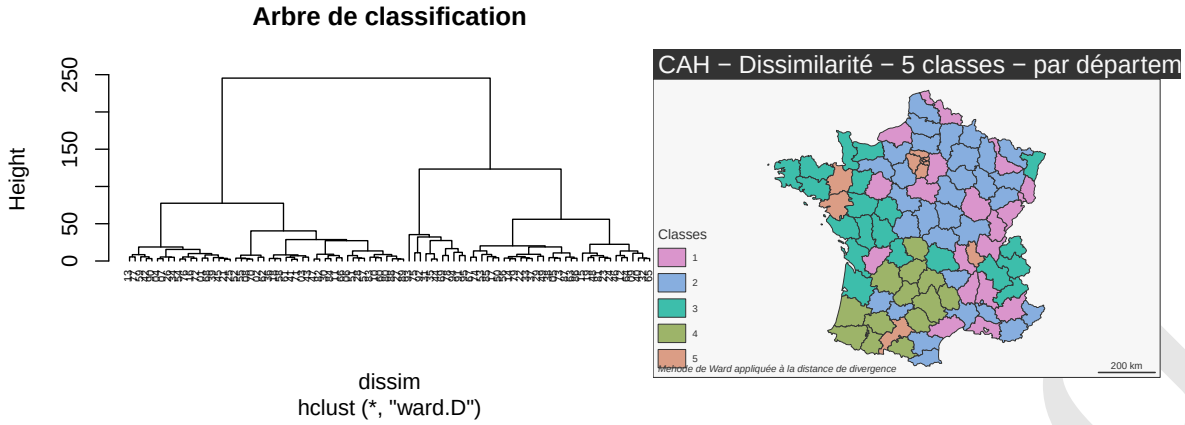
La somme des différences de vote est égale à 94.5 points de pourcentage. En divisant par deux on obtient une valeur de 47.2 qui est le pourcentage de vote qu'il faudrait modifier dans l'un ou l'autre département pour aboutir à des profils similaires. Le coefficient de divergence est compris entre 0 (votes identiques) et 100 (aucun vote commun).

2.2.2 Résultats de la classification

L'application d'une méthode de classification ascendante hiérarchique à la matrice de dissimilarité fait apparaître assez nettement cinq classes qui regroupent souvent des départements voisins mais sans pour autant former des régions.

```
# Classification
cah_dissim <- hclust(dissim, method = "ward.D")

# Découpage en 5 classes
clas_dissim <- as.factor(cutree(cah_dissim, 5))

# Classe d'appartenance pour chaque département (fond de carte)
map_dept$cah_dissim <- as.factor(clas_dissim)

## A. Arbre CAH - Dissimilarité
plot(cah_dissim, hang = -1, cex = 0.5, main = "Arbre de classification")
# ---

# B. Carte - CAH - Dissimilarité
mf_map(map_dept, type = "typo", var = "cah_dissim", leg_title = "Classes")
mf_layout("CAH - Dissimilarité - 5 classes - par département",
```

```
  credits = "Mehode de Ward appliquée à la distance de divergence",
  frame = TRUE, scale = TRUE, arrow = FALSE)
```



Une analyse des profils permet ensuite de caractériser ces classes.

```
# Calcul % national de vote
tabres <- data.frame(mat_vote_dep)
tot <- tabres |>  summarise_all(.funs = c("mean"))

# Récupération de la classe d'appartenance
tabres$clas_dissim <- clas_dissim

# Moyenne % vote pour chaque classe
res <- tabres |> group_by(clas_dissim) |> summarise_all(.funs = c("mean"))

# Calcul écarts des classes au profil moyen
mat <- res[, -1]
for (i in 1:5) {mat[i, ] <- mat[i, ] - as.matrix(tot)}
mat <- as.data.frame(t(mat))
colnames(mat) <- c("Classe1", "Classe2", "Classe3", "Classe4", "Classe5")

# Ajout ligne de totaux
mat$Profil <- as.numeric(tot)
```

Écart des classes au profil moyen (listes principales) :

	Classe1	Classe2	Classe3	Classe4	Classe5	Profil
BARDELLA	1.50	7.52	-3.97	-2.43	-13.69	33.88
HAYER	-0.55	-1.31	2.56	-1.05	1.23	14.11
GLUCKSMANN	-0.91	-2.78	2.01	2.46	2.85	13.43
AUBRY	0.74	-1.89	-1.24	-1.94	8.59	8.28
BELLAMY	-0.32	-0.17	0.19	-0.34	1.06	7.12
MARÉCHAL	0.10	0.38	-0.42	-0.36	-0.04	5.34
TOUSSAINT	0.03	-1.47	1.15	-0.38	2.44	4.90
LASSALLE	-0.79	-0.12	-0.36	3.66	-1.94	3.08
DEFFONTAINES	-0.10	-0.03	-0.08	0.75	-0.47	2.53
THOUY	0.09	0.20	-0.07	-0.24	-0.30	2.07
GOVERNATORI	0.11	-0.13	0.18	-0.16	0.04	1.23

	Classe1	Classe2	Classe3	Classe4	Classe5	Profil
ASSELINEAU	0.04	-0.02	-0.05	0.11	-0.03	1.02
PHILIPPOT	0.04	0.05	-0.05	0.04	-0.17	0.95

- la **classe 1** est assez proche du **profil moyen** avec une légère sur-représentation des votes Bardella (+1.5) et Aubry (+0.79), associée à une sous-représentation des votes Glucksman (-0.91), Lassalle (_0.79), Hayer (-0.55) et Bellamy (-0.32).
- la **classes 2** est caractérisée par la **très forte surreprésentation du vote d'extrême droite** pour Bardella (+7.5), Maréchal (+0.38) ou Philippot (+0.05) ainsi que le parti animaliste (+0.2) associé à une sous-représentation des autres partis, en particulier de Glucksmann (-2.78) et Toussaint.
- la **classe 3** surreprésente les votes des **partis centristes**, qu'il s'agisse du centre-gauche (Hayer : +2.56), du centre-droit (Glucksman : +2.01) ou des écologistes (Toussaint : +1.15) et elle sous-représente les partis d'extrême-droite mais aussi d'extrême gauche.
- la **classe 4** s'inscrit plutôt dans **une spécificité régionale du Sud-Ouest** caractérise par l'importance du vote Lassalle (+3.66) et du vote Deffontaines (+0.75), associé au vote de centre-gauche de la liste Glucksmann (+2.46). Comme dans le cas précédent, on observe une faiblesse du vote pour les partis d'extrême droite ou d'extrême gauche.
- la **classe 5** correspond enfin à un **vote des grandes métropoles** caractérisé par un score exceptionnel de la liste Aubry (+8.59), associé à une surreprésentation des votes pour les autres partis de gouvernement de droite (Bellamy : +1.06, Hayer : +1.23) ou de gauche (Glucksmann : +2.85, Toussaint : +2.44)

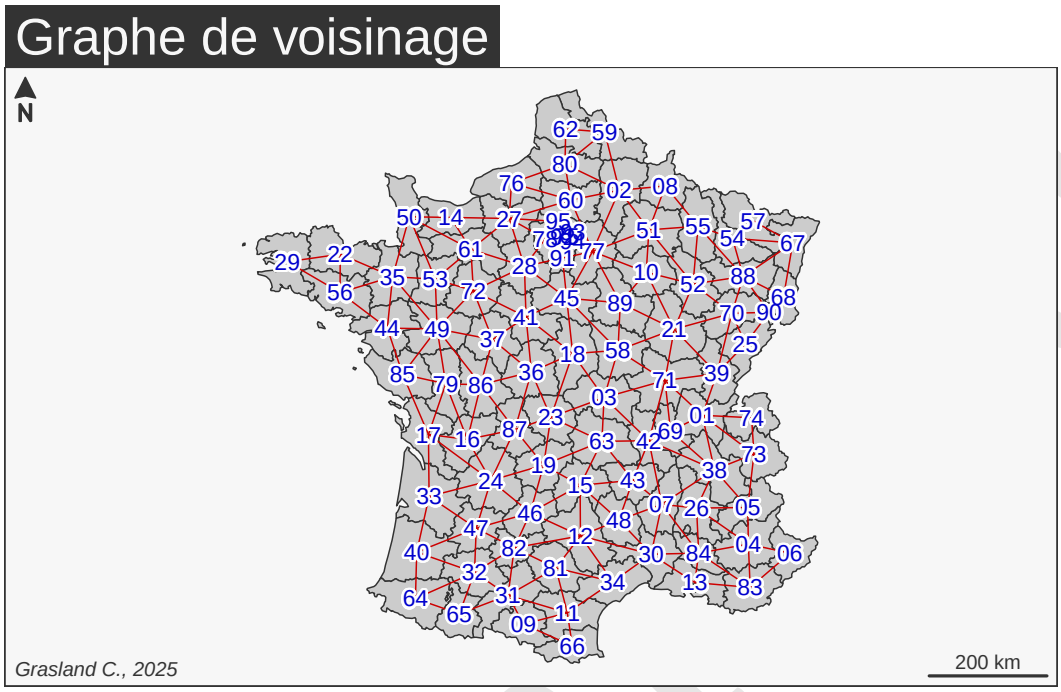
2.3 Régionalisation

2.3.1 Matrice de contiguïté

On calcule la matrice de contiguïté au niveau départemental à l'aide du package `sf`. Puis, on les visualise cartographiquement.

```
mapdon_dept <- left_join(map_dept, don_dept)


#### GRAPHE DE VOISINAGE (contiguité)
# Calcul matrice de contiguïté
mat_conti <- st_intersects(mapdon_dept, mapdon_dept, sparse = FALSE)
colnames(mat_conti) <- mapdon_dept$dept
rownames(mat_conti) <- mapdon_dept$dept

# Suppression de la moitié de la matrice (et de la diagonale)
mat_conti[lower.tri(mat_conti, diag = TRUE)] <- FALSE

# Construction d'un tableau de lien (i, j) de contiguïté
reg_link_contig <- as.data.frame.table(mat_conti, responseName = "contig") |>
                           filter(contig == TRUE)

# Création de la couche géographique de liens
reg_links_contig  <- mf_get_links(x = mapdon_dept,
                                  df = reg_link_contig,
                                  x_id = "dept",
                                  df_id = c("Var1","Var2"))
```

```
# Cartographie
mf_map(mapdon_dept)
mf_map(reg_links_contig , col = "red3", add = TRUE)
mf_label(mapdon_dept, var = "dept", col = "blue3", halo = TRUE, bg = "white")
mf_layout("Graphe de voisinage", frame = TRUE, credits = "Grasland C., 2025")
```

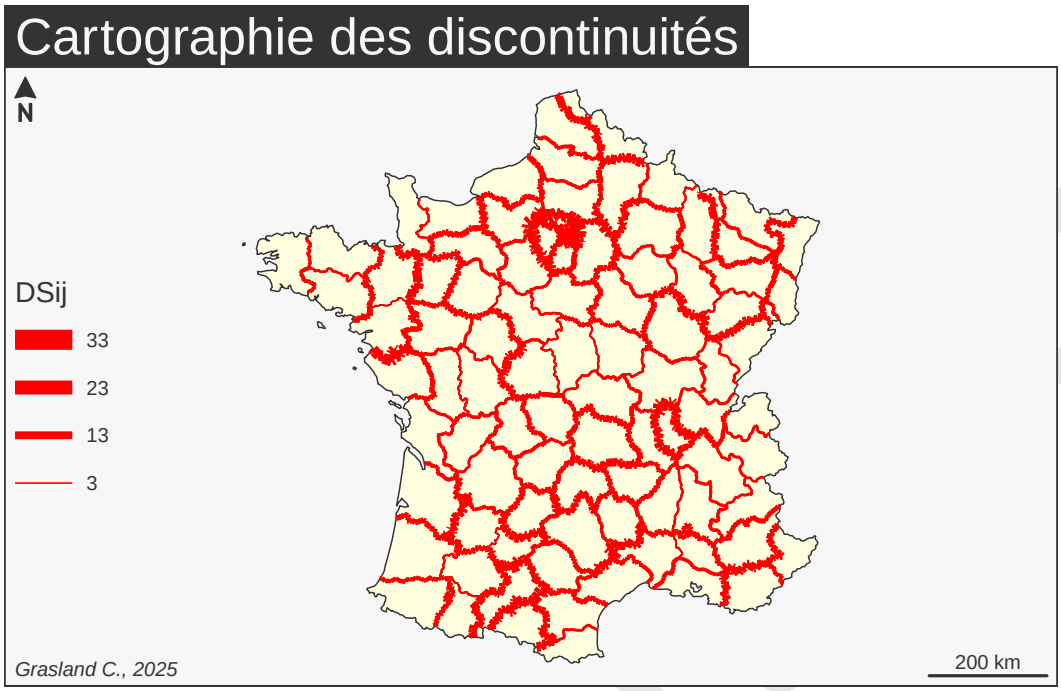


2.3.2 Dissimilarités locales

Avant de procéder à la régionalisation, on peut visualisser les discontinuités en extrayant les frontières des unités spatiales à l'aide de la fonction `mf_get_borders()`du package `mapsf` et en effectuant une jointure avec les valeurs de dissimilarité (Grasland, 1997). On pourra ainsi repérer les limites qui séparent des départements très ressemblants (donc susceptibles de se regrouper en régions) ou au contraire très différents (qui seront probablement localisés dans des régions différentes).

```
# Conversion de la matrice de dissimilarité en tableau long
m <- as.matrix(dissim)
tab_dis <- cbind(expand.grid(dimnames(m)), value = as.vector(m))
names(tab_dis ) <- c("i", "j", "DSij")

# Extraction des frontières d'unités spatiale
map_border_dep <- mf_get_borders(mapdon_dept)[, c("dept", "dept.1")]
names(map_border_dep) <- c("i", "j", "geometry")

# Jointure
map_border_dep <- merge(map_border_dep, tab_dis, by = c("i", "j"))

# Cartographie des dissimilarités les plus fortes
mf_map(mapdon_dept, type = "base", col = "lightyellow")
mf_map(map_border_dep,
       type = "prop",
```

```
        col = "red",
        var = "DSij",
        val_max = 70,
        leg_pos = "left")
mf_layout("Cartographie des discontinuités", frame = TRUE, credits = "Grasland C., 2025")
```



Les discontinuités les plus remarquables sont celles qui séparent les départements d'Ile-de-France du reste du Bassin parisien (ex. dissimilarité de 26 points entre Yvelines et Eure) mais aussi les départements franciliens entre eux (ex. dissimilarité de 33 points entre Seine-Saint-Denis et Paris). On retrouve également de très fortes différences entre les départements qui abritent les grandes métropoles de province (Lyon, Toulouse, Nantes, Lille, ...) et leurs voisins. Mais il apparaît également des discontinuités entre certains départements plus ruraux. A l'inverse, on repère des groupes de départements peu différents les uns des autres dans les Alpes, le sud du Bassin Parisien ou le Centre-Ouest. La carte des discontinuités permet donc d'anticiper les regroupements les plus probables qui vont intervenir au cours de l'étape de régionalisation.

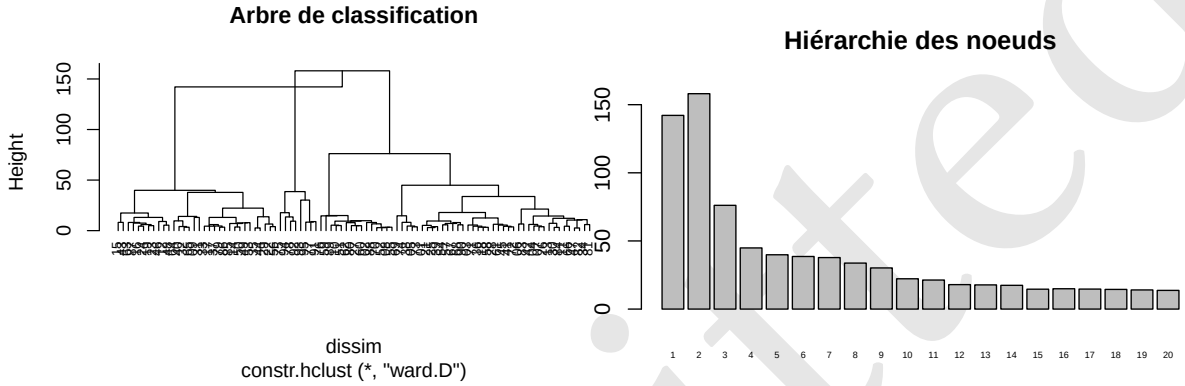
Une approche différentes, proposée par les écologues, consiste à mesurer la contribution des unités spatiales et des variables les décrivant à la production des dissimilarités au niveau global et local. Cette approche est classiquement menée à l'aide de mesures basées sur la **variance,** mais les auteurs proposent de la généraliser à **une mesure quelconque de dissimilarité** ce qui permet une meilleure adéquation à la problématique (Legendre et De Cáceres, 2013). Et qui permet d'appliquer la méthode non pas à l'ensemble des dissimilarités (comme dans une ACP ou une CAH) mais uniquement aux dissimilarités locales.

2.3.3 Résultats de la regionalisation

La réalisation d'une régionalisation ascendante hiérarchique est très simple avec la fonction `constr.hclust()`du package `adespatial`.

```
# Régionalisation ascendante hiérarchique
regio <- constr.hclust(d = dissim, method = "ward.D", links = reg_link_contig)
```

```r
# Arbre de classification
plot(regio,
     main = "Arbre de classification",
     hang = -1,
     cex = 0.7)
# ---

# Hiérarchie des noeuds
barplot(rev(regio$height)[1:20],
        main = "Hiérarchie des noeuds",
        names.arg = 1:20,
        cex.names = 0.4)
```

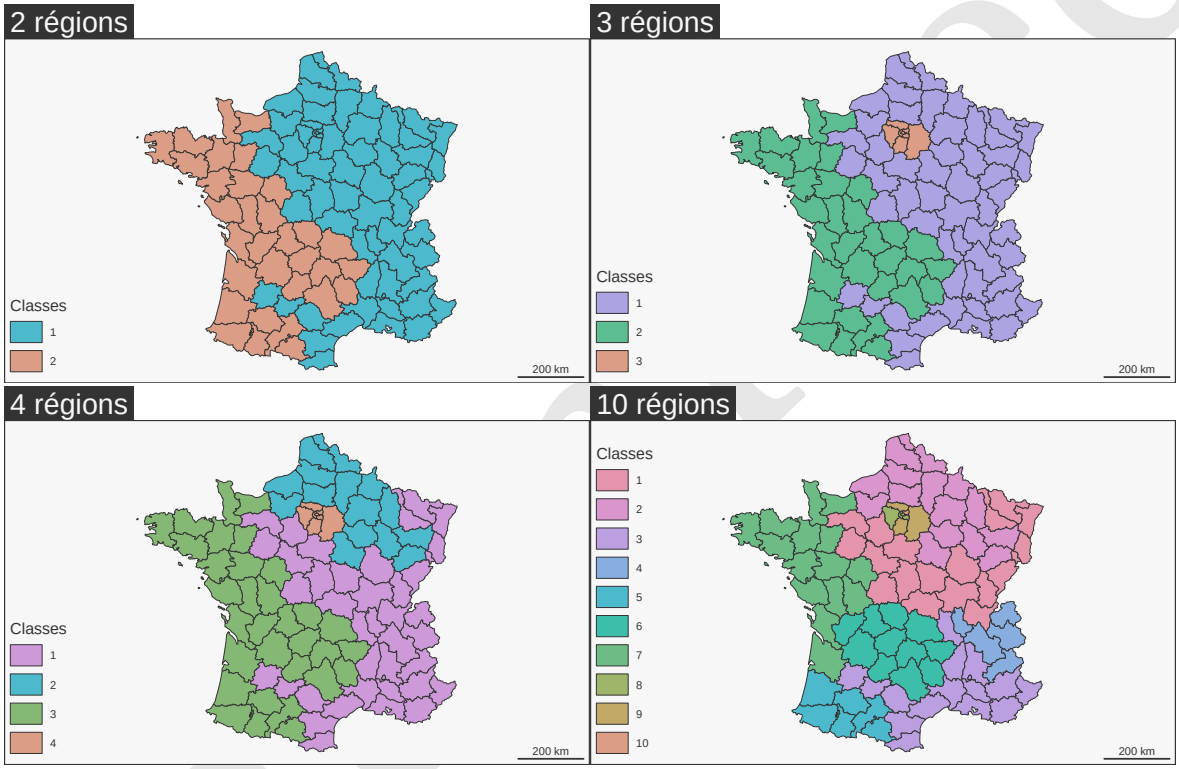


L'arbre de classification et l'indice de hiérarchie des noeuds mettent tout d'abord en valeur les partitions en 2, 3 ou 4 classes qui se détachent très clairement des regroupements ultérieurs. On observe toutefois que la régionalisation en 2 classes est moins efficaces que la partition en deux classes ce qui peut suprendre un utilisateur habitué à utiliser des méthodes fondées sur la distance euclidienne au carré et la variance. Ce résulat est en fait logique dans la mesure où nouss avons utilisé une métrique non euclidienne (Guénard et Legendre, 2022). Dans notre exemple, il signale que le premier niveau de découpage de la France en régions électoral n'est pas une oppostion nord-est/sud-ouest mais un découpage en trois entités qui isole la région Ile-de-France. Quant au découpage en quatre régions, il met en valeur à l'intérieur de la France du nord-est le cas de la partie nord et est du bassin parisien qui est singulièrement différente du reste de la France du Nord-Est. Au delà de cette partition en quatre classes, on observe une suite de partition de niveau voisins jusqu'au 9e noeuds de l'arbre où apparaît une discontinuité nette, ce qui incite à retenir une partition en 10 régions. Le niveau de dissimilarité de ce découpage en 10 régions sera approximativment le même que celui que nous avios utilisé précédemment pour réaliser une classification comportant cinq classes. Ce qui confirme qu'une régionalisation est par définition moins efficace qu'une classification puisqu'elle doit comporter deux fois plus de groupes pour aboutir au même niveau d'homogénéité.

On peut représenter les quatre niveaux de régionalisation en effectuant un découpage de l'arbre à l'aide de la fonction `cutree()` et d'un logiciel quelconque de cartographie thématique dans R comme `mapsf`.

```r
mapdon_dept$reg2 <- as.factor(cutree(regio, 2))
mapdon_dept$reg3 <- as.factor(cutree(regio, 3))
mapdon_dept$reg4 <- as.factor(cutree(regio, 4))
mapdon_dept$reg10 <- as.factor(cutree(regio, 10))
```

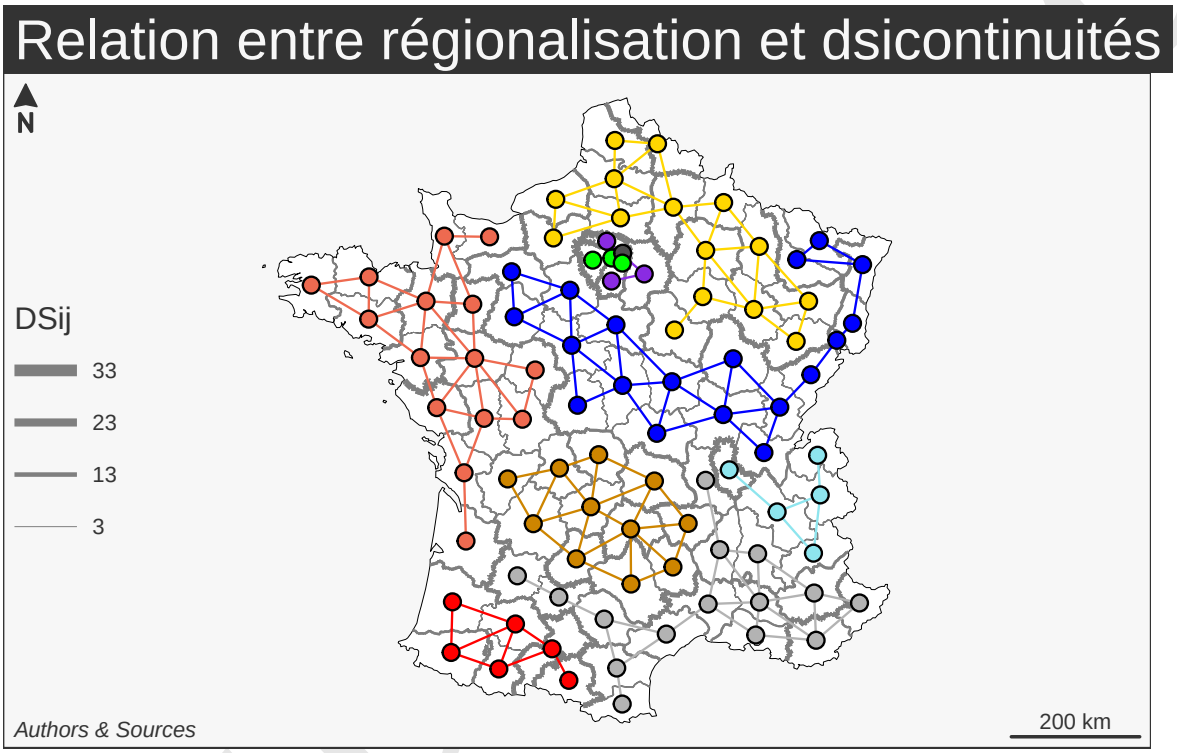
```
mf_map(mapdon_dept, var="reg2",type="typo", leg_title = "Classes")
mf_layout("2 régions", credits = "", scale=TRUE, frame=TRUE, arrow=FALSE)
#---

mf_map(mapdon_dept, var="reg3",type="typo", leg_title = "Classes")
mf_layout("3 régions", credits = "", scale=TRUE, frame=TRUE, arrow=FALSE)
#---

mf_map(mapdon_dept, var="reg4",type="typo", leg_title = "Classes")
mf_layout("4 régions", credits = "", scale=TRUE, frame=TRUE, arrow=FALSE)
#---

mf_map(mapdon_dept, var="reg10",type="typo", leg_title = "Classes")
mf_layout("10 régions", credits = "", scale=TRUE, frame=TRUE, arrow=FALSE)
```



Mais on peut également utiliser la fonction `plot : constr.hclust()` du package `adespatial` à condition de lui fournir les centroïdes des unités spatiales. On peut alors visualiser la façon dont le graphe de contiguïté a été segmenté pour aboutir à une régionalisation. Il est alors intéressant d'y superposer la carte des discontinuités pour mieux voir comment les régions réspectent dans la mesure du possible les frontières correspondant aux plus fortes différences entre unités voisines.

```
# Extraction des coordonnées des centroïdes des départements
dep_centroide <- st_coordinates(st_centroid(mapdon_dept))

# Régionalise
regio <- constr.hclust(d = dissim,
                       method = "ward.D",
                       links = reg_link_contig,
                       coords = dep_centroide)
```

```
# Trace le fonds de carte et les discontinuités
mf_map(mapdon_dept, type="base", col="white",border="black",lwd=0.2)
mf_map(map_border_dep,
       type = "prop",
       col = "gray50",
       var = "DSij",
       lwd_max = 5,
       leg_pos = "left",
       add = TRUE)
mf_layout("Relation entre régionalisation et dsicontinuités", frame = TRUE)
plot(regio, k = 10, links = TRUE, axes = FALSE, plot = FALSE, hybrids = "no")
```



On procède maintenant à l'analyse des écarts au profil moyen en reprenant la même procédure que pour la classification. Pour faciliter l'analyse, on recode les noms de régions pour combiner les partitions en trois régions (Nord-Est = NE, Sud-Ouest = SO, Ile-de-France = IF) et la partition en 10 (les quatres sous-régions du Nord-Est sont codées NE1, NE2, NE3, NE4, les trois régions du Sud-Ouest SO1, SO2, SO3 et les trois régions d'Ile-de-France IF1, IF2, IF3)

```
# Suppression colonne classification précédente
tabres <- tabres[, -39]

# Ajout classification en 10 classe
tabres$clas_10 <- cutree(regio, 10)

# calcul moyennes des classes
res <- tabres |> group_by(clas_10) |> summarise_all(.funs = c("mean"))
mat <- res[, -1]
```

```r
# Calcul écarts des classes au profil moyen
for (i in 1:10) {mat[i, ] <- mat[i, ] - as.matrix(tot)}

# Transposition de la matrice to dataframe
mat <- as.data.frame(t(mat))

# Renommage des colonnes
colnames(mat) <- c("NE1","NE2","NE3","NE4",
                   "SO1","SO2","SO3",
                   "IDF1","IDF2","IDF3")

# Ajout ligne de totaux
mat$Profil <- as.numeric(tot)
```

Écart des régions au profil moyen (listes principales) :

	NE1	NE2	NE3	NE4	SO1	SO2	SO3	IDF1	IDF2	IDF3	Profil
BARDELLA	2.57	9.65	4.04	-4.51	-4.33	-1.37	-5.42	-18.93	-5.51	-17.00	33.88
HAYER	0.29	-1.26	-2.24	0.67	-0.73	-0.98	3.59	3.38	-0.90	-5.80	14.11
GLUCKSMANN	-1.49	-3.85	-0.93	1.15	4.02	1.53	2.81	3.81	-1.09	-0.67	13.43
AUBRY	-1.12	-1.84	0.20	1.11	-1.16	-1.84	-1.54	8.29	9.55	28.84	8.28
BELLAMY	0.48	-0.18	-1.28	0.33	-1.79	0.77	0.30	3.40	-0.52	-3.08	7.12
MARÉCHAL	0.11	-0.28	1.12	0.37	-0.27	-0.54	-0.70	0.91	-0.13	-1.63	5.34
TOUSSAINT	-0.62	-1.63	-0.46	2.32	0.31	-0.46	1.45	2.79	0.05	1.75	4.90
LASSALLE	-0.51	-0.41	-0.14	-1.16	3.69	2.27	-0.36	-2.43	-1.99	-2.44	3.08
DEFFONTAINES	-0.11	-0.14	-0.03	-0.63	0.42	0.99	-0.15	-0.69	-0.35	0.29	2.53
THOUY	0.21	0.38	-0.16	-0.24	-0.31	-0.09	-0.14	-0.42	0.29	-0.45	2.07
GOVERNATORI	0.06	-0.11	-0.10	0.25	-0.07	-0.11	0.18	-0.12	0.12	-0.34	1.23
ASSELINEAU	0.00	-0.13	0.13	0.09	0.12	0.04	-0.13	-0.06	0.19	0.09	1.02
PHILIPPOT	0.07	-0.03	0.12	0.00	0.04	-0.03	-0.10	-0.21	-0.03	-0.18	0.95

Trois des quatre sous-régions qui composent la région Nord-Est se caractérisent par une surreprésentation générale des votes pour les listes de droite (Bellamy) ou d'extrême-droite (Bardella, Maréchal).

- La **région NE1 de type droite et extrême droite** occupe les franges sud du bassin parisien ainsi que l'Alsace et le nord de la Lorraine. Elle se caractérise par une légère sur-représentation du vote Bardella (+2.57) combinée à une sur-représentation des autres votes de droite (Bellamy +0.48, Hayer +0.29, Maréchal +0.11, Philippot +0.07) et une sous-représentaiton des listes de gauche (Glucksman -1.49, Aubry -1.12, Toussaint -0.62).
- la **region NE2 de type bastion RN rural et ouvrier** occupe le nord et l'est du bassin parisien de la Normandie à la Lorraine en passant par le Nord et la Champagne. Sa caractéristique principale est un score exceptionnellement élevé pour la liste Bardella (+9.65) et une faiblesse relative de toutes les autres listes à l'exception de la liste Thouy du parti animaliste.
- la région **NE3 de type bastion d'extrême-droite diversifié** correspond à un vote d'extrême droite mélangeant davantage le vote RN de la liste Bardella (+4.04) avec d'autres avatars de l'extrême-droite se traduisant par une surreprésentation des listes Maréchal (+1.12), Asselineau (+0.13) ou Philippot (+0.12). Comme dans le cas précédent, les autres listes de droite classique ou de gauche sont sous-représentés à l'exception de la liste Aubry (+0.2).

— la région **NE4 de type métropolitain écologiste** constitue une enclave à l'intérieur de la région NE regroupant la métropole Lyonnaise et le nord des Alpes. Elle affiche des caractéristiques très différentes voire opposées aux types prédédents. Elle se caractéristique par un score très élevé des écologistes (Toussaint +2.32, Gobernatori +0.23), ainsi que des partis de gauche (Aubry : +1.11) de centre-gauche (Glucksman +1.15) et de centre-droit (Hayer +0.67). Le vote Bardella y est nettement sous-représenté (-4.51) mais pas le vote de droite (Bellamy +0.33) ou d'extrême droite dans d'autres versions (Maréchal +0.37, Asselineau +0.09).

La région Sud-Ouest affiche un profil général très différent caractérisé par la faiblesse conjointe des votes d'extrême-droite (Bardella, Maréchal) et d'extrême-gauche et une surreprésentation des listes portées par les partis centristes de gouvernement (Hayer, Glucksman). Mais elle affiche trois variantes bien typées en raison du rôle de deux listes à forte composante régionale. - la région **SO1 de type identité régionale sud-ouest** regroupe les départements situés au Nord des Pyrénées, du pays Basque à Toulouse. Son originalité fondamentale réside dans le poids exceptionnel du vote pour la liste Alliance Rurale portée par Jean Lassalle (+3.69) combinée par un vote très élevé pour les listes socialistes (Glucksman +3.62) et communiste (Deffontaines +0.42). - la région **SO2 de type radical-socialiste** prolonge la région précédente vers le massif central, exception faite de la vallée de la Garonne acquise à l'extrême-droite. Elle conserve des caractéristiques voisines de S01 mais en moins accentué. Elle aurait probablement fusionné avec la précédente sans l'obstacle constitué par les départements conquis par l'extrême-droite qui font obstacle à l'unification en une seule région. - la région **SO3 de type ouest chrétien-démocrate** associe les départements de Bretagne, Pays de Loire, Basse Normandie et nord de l'Aquitaine. Elle affiche une forte résistance au vote d'extrême-droite (Bardella -5.42, Maréchal -0.70) comme d'extrême-gauche (Aubry -1.54, Deffontaines -0.36) et concentre ses suffrages sur les listes des partis de centre-gauche (Glucksmann +2.81), de centre-droit (Hayer +3.59) ainsi que les écologistes (Toussaint +1.45, Governatori +0.18)

La région Ile-de-France forme la troisième région, caractérisée par une résistance générale au vote d'extrême droite et une performance exceptionnellement élevée de la liste LFI portée par Aubry. Elle n'en comporte pas moins de très forts contrastes internes.

-la région **IF1 de type métropolitain central** regroupe Paris, les Hauts-de Seine, les Yvelines et le Val-de-Marne dans une catégorie caractérisée par le partage des votes entre listes des partis de gouvernement de centre-gauche (Glucksman +3.81) et de centre-droit (Hayer +3.38) ainsi que par des scores très élevés pour la liste LFI (Aubry +8.3), les écologistes (Toussaint +2.8) et la droite classique (Bellamy +3.4) ou les formes d'extrême-droite élitiste (Maréchal +0.70). -la région **IF2 de type métropolitain périphérique** regroupe les départements de grande couronne du Val d'Oise, de l'Essone et de Seine-et-Marne avec un rejet du rassemblement national beaucoup moins marqué (-5.4) et un vote toujours plus important pour la liste LFI de M. Aubry (+9.55). Les partis centristes ont désormais des scores légèrement plus faibles que leur moyenne nationale. - la région **IF3 de type bastion LFI** se limite à l'unique département de Seine-Saint-Denis dont la caractéristique unique est le score exceptionnel de la liste Aubry (+28.8) et à un degré bien moindre des écologistes (Toussaint +1.75) et communistes (Deffontaines +0.29)

2.4 Discussion

Quels sont les apports respectifs des deux approches de régionalisation et de classification ?

2.4.1 Intérêt et limites de la classification

L'analyse de classification offre obligatoirement un meilleur résumé de l'information contenue dans la matrice de dissimilarité dans la mesure où elle ne subit pas la contrainte de contiguïté qui est imposée à la régionalisation. Même si la méthode de classification ascendante hiérarchique n'aboutit

pas nécessairement à une solution optimale en matière de maximisation de l'homogénéité intra-classe et de l'hétérogénéité inter-classe (la méthode des k-means est a priori plus efficace mais plus coûteuse en temps de calcul), elle présente l'avantage de fournir des résumés à différents niveaux d'agrégation et de distinguer des types et des sous-types à l'intérieur de ceux-ci.

La limite de la méthode concerne sa visualisation cartographique qui laisse apparaître des blocs régionaux mais qui correspondent rarement à une classe unique. Les résultats n'ont pas vocation à produire une géographie du vote même si le commentaire des résultats fait appel à des notions de proximité et de localisation.

2.4.2 Intérêt et limites de la régionalisation

L'analyse de la régionalisation possède les mêmes propriétés de regroupement hiérarchique en régions qui se subdivisent ensuite en sous-région ce qui permet une analyse nuancée des oppositions principales et secondaires. L'analyse géographique des résultats permet donc bien de construire un commentaire multiscalaire partant des divisions principales ("Nord-Est/ Nord-Ouest/ Ile-de-France) pour extraire ensuite des subdivisions secondaires ce qui est la procédure habituelle de la description d'un espace géographique.

La limite de l'analyse tient ici au poids de la contrainte de contiguïté qui oblige à regrouper les entités à l'intérieur d'un ensemble d'unités voisines même lorsqu'elles sont séparées par des discontinuités extrêmement élevées. Ce qui aboutit à une hétérogénéité parfois très élevé des entités regroupées.

2.4.3 Autocorrélation et diffusion spatiale des comportements électoraux

Finalement le choix de l'une ou l'autre méthode dépend des hypothèses que l'on formule sur l'origine et les conséquences de l'autocorrélation spatiale des comportements électoraux.

- Si l'on suppose que les causes du votes sont principalement d'ordre social et liées à des causes individuelles qui ne dépendent pas de la localisation géographique, alors la classification semble la solution la plus logique. Une fois identifiées les classes correspondant à tel ou tel type de comportement électoral, on pourra les mettre en rapport avec d'autres attributs des lieux tels que la richesse des habitants, les modes d'habitat, l'accessibilité au services, etc.
- Si l'on suppose au contraire que les comportements électoraux de propagent dans l'espace à la faveur de processus d'imitation ou d'identification, alors il semble pertinent de regrouper des lieux proches en région qui sont susceptibles de voir leurs attitudes électorales converger au cours du temps. La régionalisation est alors un outil pertinent de prospective ou de stratégie.

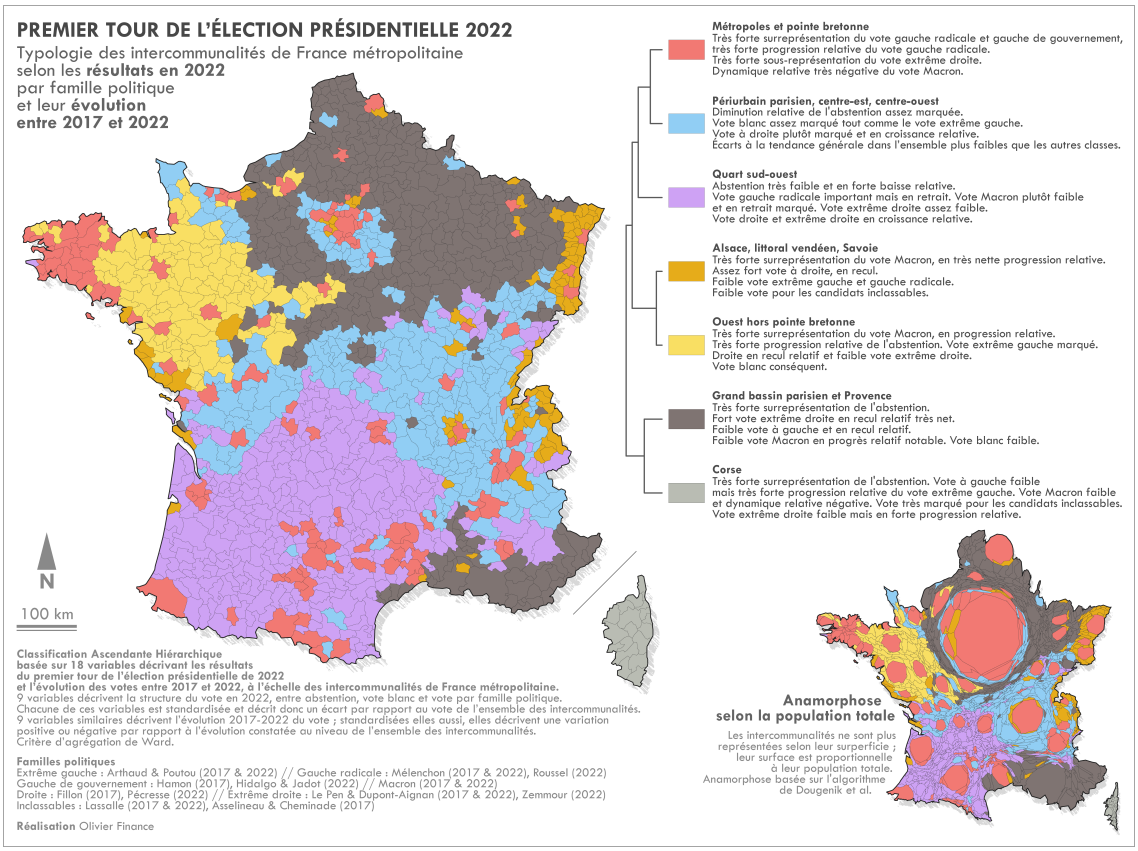
3 Échelle des circonscriptions : gradients urbains ou discontinuités ?

La reproduction des analyses précédentes au niveaux des 535 circonscriptions législatives constitue de prime abord un avantage puisque ces unités spatiales ont des populations beaucoup plus proches entre elles que les départements. La loi impose en effet des seuils miniumum et maximum de population à ces unités afin d'assurer une représentation équitable des citoyens à l'Assemblee Nationale. Malgré les exceptions (départements peu peuplés ayant au moins un député) et les manipulations de limites pour favoriser tel ou tel parti (*gerrymandering*),les circonscriptions sont un cadre idéal d'observation des résultats des élections européennes … surtout lorsqu'elles sont suivies d'une dissolution de l'Assembléen Nationale comme ce fut le cas en 2024.

Ce changement d'échelle entraîne toutefois un saut de complexité dans l'analyse puisque les oppositions entre les espaces ruraux, périurbain et métropolitain qui étaient encore peu visibles à l'échelle

d'observation des départements sont désormais fondamentaux et créent pour beaucoup de partis politiques des distribution en "peau de léopard" composés de taches isoles (e.g.liste LFI présente surtout en ville) ou de nappes percées de trous (e.g. vote RN majoritaire dans les zones rurales et fortement réduit dans les métropoles). La question est alors de savoir si la transition entre espaces métropolitains et ruraux s'opère de façon graduelle (hypothèse du gradient d'urbanité) ce qui autoriserait la création de régions de proche en proche. Ou si on passe brutalement d'un comportement à un autre ce qui ferait des métropoles des enclaves bien délimitées cernées par des discontinuités.

Une carte publiée par O. Finance dans Cybergeo à propos du premier tour des élections présidentielle de 2022 à l'échelle des intercommunalités montre clairement l'existence d'une double structure à la fois régionale et métropolitaine :



L'auteur précise que la carte combine en fait des variables de niveau (structures des votes en 2022) et des variables d'évolution (entre les élections de 2017 et 2022) :

> Cette carte a été construite à l'aide d'une Classification Ascendante Hiérarchique. Elle synthétise 9 variables décrivant la structure du vote en 2022 (abstention, vote blanc, vote pour chaque famille politique) et 9 variables similaires décrivant l'évolution du vote entre 2017 et 2022. Ces variables sont toutes standardisées et décrivent donc pour les 9 premières des écarts par rapport au vote de l'ensemble des intercommunalités, pour les 9 suivantes des variations positives ou négatives par rapport à l'évolution constatée au niveau de l'ensemble des intercommunalités. Source : Finance O., 2022, Cybergeo Conversation

La structure obtenue combine à la fois un archipel métropolitain (classe représentée en rouge) et des blocs régionaux bien identifiables indiquant une forte auto-corrélation spatiale des votes dans

les espaces non métropolitains.

3.1 Matrice de dissimilarité

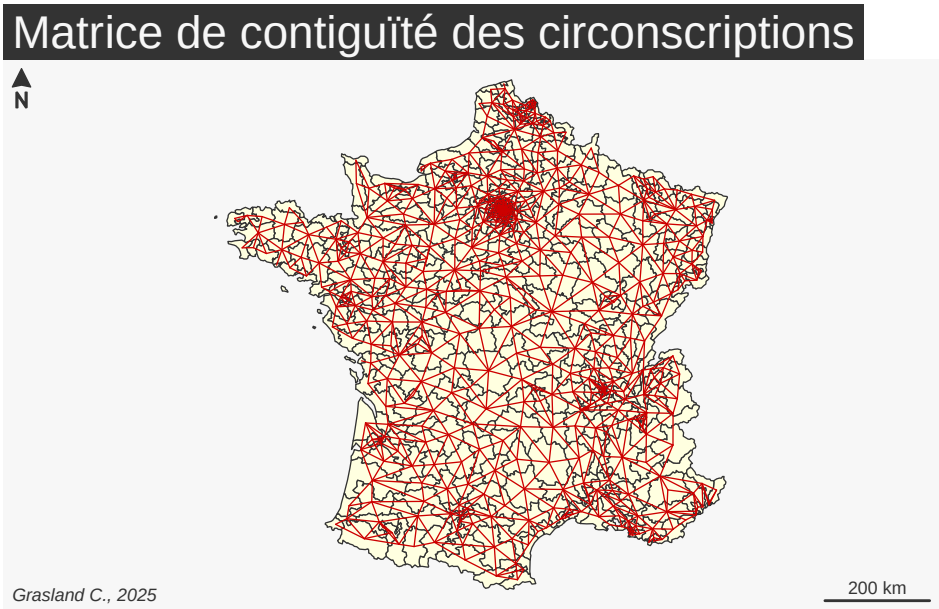
On charge les fichiers de circonscriptions et on construit la matrice de dissimilarité en utilisant la même procédure que pour les départements (cf. partie 2.2).

```
# Chargement des résultats par circonscription et d'un fond de carte
map_circ <- readRDS("data/net/map_circ.RDS")
don_circ <- readRDS("data/net/don_circ.RDS")

# Calcul répartition % de vote par circonscription pour chaque liste
result_vote_circ <- don_circ[, 12:49]
mat_vote_circ <- 100 * result_vote_circ / apply(result_vote_circ, 1, sum)
rownames(mat_vote_circ) <- don_circ$circ
colnames(mat_vote_circ) <- listes$tete_nom

# Dissimilarité
dissim <- dist.ldc(mat_vote_circ, method = "manhattan") / 2
```

```
Info -- For this coefficient, sqrt(D) would be Euclidean
Info -- This coefficient does not have an upper bound (no fixed D.max)
```

On prépare ensuite la la matrice de contiguïté des circonscriptions en suivant là encore la procédure utilisée pour les départements (cf. partie 2.3).

```
# Jointure fond de carte et résultat de vote
mapdon_circ <- left_join(map_circ, don_circ)

# Matrice de contiguïté
mat_conti <- st_intersects(mapdon_circ, mapdon_circ, sparse = FALSE)
colnames(mat_conti) <- mapdon_circ$circ
rownames(mat_conti) <- mapdon_circ$circ

# Suppression de la moitié de la matrice (et de la diagonale)
mat_conti[lower.tri(mat_conti, diag = TRUE)] <- FALSE

# Construction d'un tableau de lien (i, j) de contiguïté
circ_link_contig <- as.data.frame.table(mat_conti, responseName = "contig") |>
                          filter(contig == TRUE)

# Création de la couche géographique de liens
circ_links_contig  <- mf_get_links(x = mapdon_circ,
                                   df = circ_link_contig,
                                   x_id = "circ",
                                   df_id = c("Var1","Var2"))

# Cartographie
mf_map(mapdon_circ,col="lightyellow")
mf_map(circ_links_contig , col = "red3", add = TRUE)
mf_layout("Matrice de contiguïté des circonscriptions", credits = "Grasland C., 2025")
```



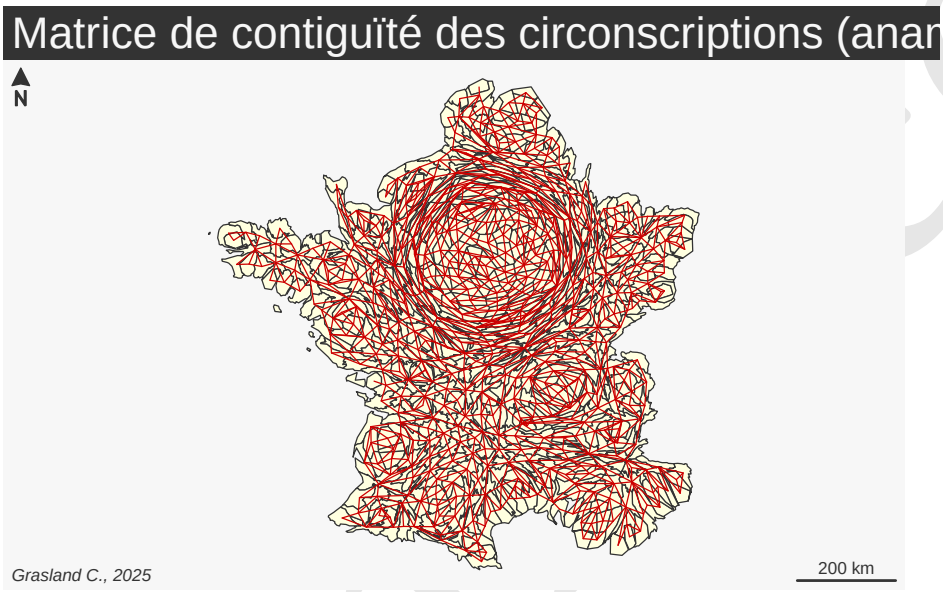
Pour mieux visualiser les zones urbaines, on peut créer une carte par anamorphose à l'aide de la fonction `cartogramR()` du package du même nom. On prend comme variable de poids le nombre de votants ce qui donne des surfaces approximativement égales aux unités spatiales.

```
# Construction du cartogramme en fonction du nombre de votants
cartogram_R <- cartogramR(mapdon_circ,
                          count = "vot",
                          method = "dcn",
                          options = list(L = 4096, maxit = 200))$cartogram |>
               st_as_sf()
```

```
 The maximum number of increases (3) in the criterion between 2 stages is exceeded (see option
 Main loop exit too early:
  Objective err. is not met: actual error=0.7092 > objective=0.01
 If the result does not satisfy your needs, please
  - increase verbosity level (to understand the problem),
  - increase maxit,
  - increase maxinc (risky),
  - increase maxrelError and maxrelTol
 in cartogramR() options.
```

```
# Jointure cartogramme et résultats de vote
cartogram_circ <- left_join(cartogram_R, don_circ)

# Matrice de contiguïté du cartogramme
mat_conti <- st_intersects(cartogram_circ, cartogram_circ, sparse = FALSE)
colnames(mat_conti) <- cartogram_circ$circ
rownames(mat_conti) <- cartogram_circ$circ

# Suppression de la moitié de la matrice (et de la diagonale)
mat_conti[lower.tri(mat_conti, diag = TRUE)] <- FALSE

# Construction d'un tableau de lien (i, j) de contiguïté
circ_link_contig <- as.data.frame.table(mat_conti, responseName = "contig") |>
```

```
                                filter(contig == TRUE)

# Création de la couche géographique de liens
circ_links_contig  <- mf_get_links(x = cartogram_circ,
                                   df = circ_link_contig,
                                   x_id = "circ",
                                   df_id = c("Var1","Var2"))

mf_map(cartogram_R, col = "lightyellow")
mf_map(circ_links_contig , col = "red3", add = TRUE)
mf_layout("Matrice de contiguïté des circonscriptions (anamorphosée)",
          credits = "Grasland C., 2025")
```



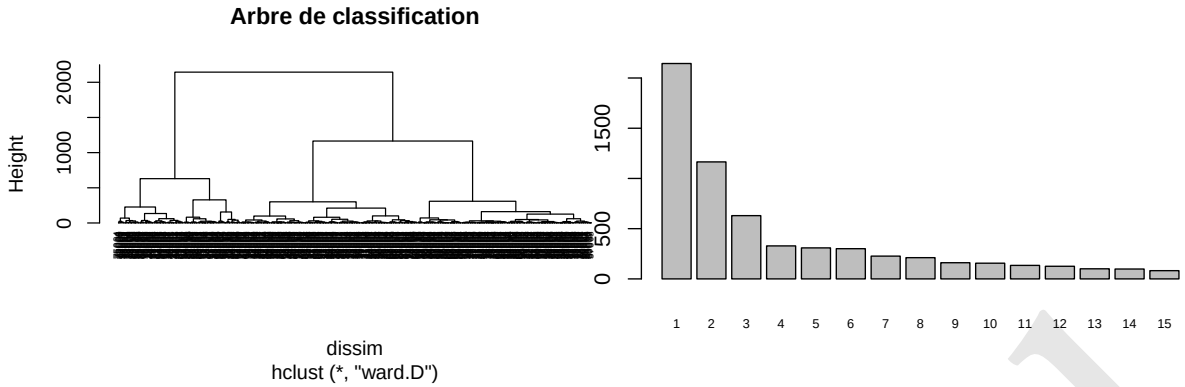
3.2 Classification

La classification fait nettement ressortir une division en 4 classes, sans rupture manifeste au delà de ce seuil.

```
# Classification
cah <- hclust(dissim, method = "ward.D")


## A. Arbre CAH - Dissimilarité
plot(cah, hang = -1, cex = 0.5, main = "Arbre de classification")
# ---

# B. Carte - CAH - Dissimilarité
barplot(rev(cah$height)[1:15],
        names.arg = 1:15,
        cex.names = 0.6)
```

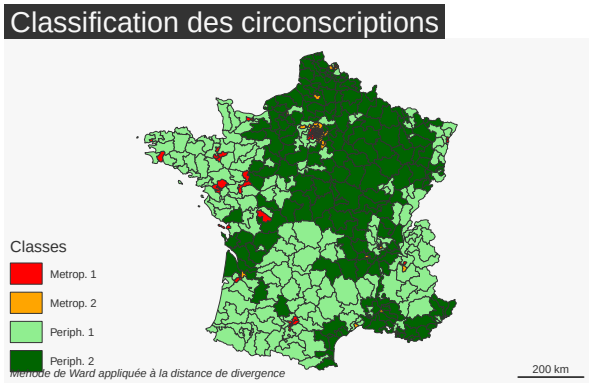
La cartographie de ces classes met en évidence une coupure évidente entre les espaces métropolitains et les espaces périphériques, cahcun d'entre eux se subdivisant ensuite en deux sous-types.

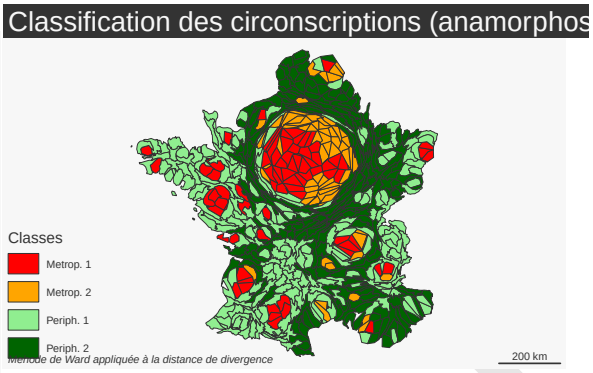
```
# Découpage en 4 classes
```




```
clas <- as.factor(cutree(cah, 4))
levels(clas) <- c("Metrop. 1", "Metrop. 2", "Periph. 1", "Periph. 2")
mapdon_circ$cah <- clas
cartogram_R$cah <- clas

# Création d'unee palette de couleur
mypal <- c("red", "orange", "lightgreen", "darkgreen")

# Carte - découpage réel des circonscriptions
mf_map(mapdon_circ,
       type = "typo",
       var = "cah",
       leg_title = "Classes",
       lwd = 0.5,
       leg_pos = "bottomleft",
      pal = mypal)

mf_layout(title = "Classification des circonscriptions",
          credits = "Mehode de Ward appliquée à la distance de divergence",
          arrow = FALSE)
# ---


# Carte - cartogramme en fonction du nombre de votants
mf_map(cartogram_R,
       type = "typo",
       var = "cah",
       leg_title = "Classes",
       lwd = 0.5,
       leg_pos = "bottomleft",
       pal = mypal)

mf_layout(title = "Classification des circonscriptions (anamorphose)",
          credits = "Mehode de Ward appliquée à la distance de divergence",
          arrow = FALSE)
```

Le profil des quatre classes est assez simple à interpréter puisqu'il s'ordonne presque parfaitement en fonction du score de la liste du RN de Bardella.






```
# Calcul % national de vote
tabres <- data.frame(mat_vote_circ)
tot <- tabres |>  summarise_all(.funs = c("mean"))

# Récupération de la classe d'appartenance
tabres$clas <- clas

# Moyenne % vote pour chaque classeclasse
res <- tabres |> group_by(clas) |> summarise_all(.funs = c("mean"))

# Calcul écarts des classes au profil moyen
mat <- res[, -1]
for (i in 1:5) {mat[i, ] <- mat[i, ] - as.matrix(tot)}
mat <- as.data.frame(t(mat))
colnames(mat) <- c("Metrop.1", "Metrop.2", "Periph.1", "Periph.2")

# Ajout ligne de totaux
mat$Profil <- as.numeric(tot)
```

Écart des régions au profil moyen (listes principales) :

	Metrop.1	Metrop.2	Periph.1	Periph.2	NA	Profil
BARDELLA	-16.65	-10.18	-0.97	10.55	NA	31.57
HAYER	3.61	-3.10	1.38	-1.74	NA	14.28
GLUCKSMANN	5.97	-0.20	1.29	-3.48	NA	13.63
AUBRY	3.01	16.82	-3.08	-3.47	NA	10.57
BELLAMY	2.78	-1.91	0.21	-0.63	NA	7.16
TOUSSAINT	3.69	0.84	0.27	-1.94	NA	5.37
MARÉCHAL	0.32	-0.95	-0.04	0.23	NA	5.37
DEFFONTAINES	-0.70	0.16	-0.03	0.25	NA	2.41
LASSALLE	-1.52	-1.52	0.75	0.30	NA	2.32
THOUY	-0.45	-0.16	-0.03	0.25	NA	2.03
GOVERNATORI	-0.03	-0.15	0.16	-0.10	NA	1.24
ASSELINEAU	-0.13	0.08	0.05	-0.03	NA	1.00
PHILIPPOT	-0.21	-0.14	0.05	0.08	NA	0.91

— **Les espaces métropolitains centraux (Metrop. 1)** votent beaucoup moins pour le rassemblement National (Bardella) et les partis à implantation régionale (Lassalle, Deffontaines),

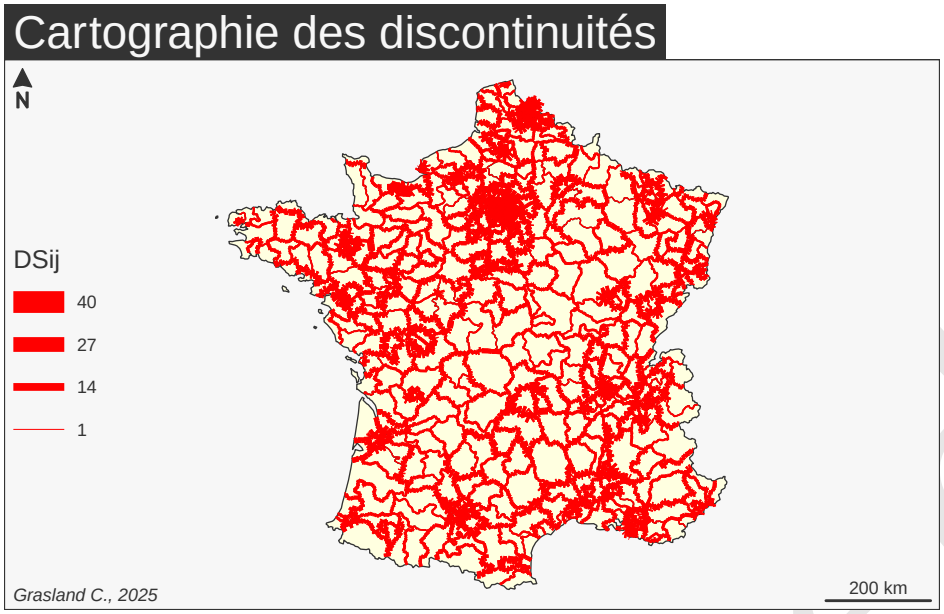
privilégiant les partis traditionnels de gouvernement (Hayer, Glucksman, Bellamy) ainsi que les écologistes (Toussaint), LFI (Aubry) ou la liste d'extrême-droite de Marechal.

- **Les espaces métropolitains périphériques (Metrop. 2)** correspondent aux zones d'implantation privilégiée de la France insoumise (Aubry), associée à une surreprésentation légère des votes communistes ou écologistes.
- **Les espaces périphériques intégrés (Periph. 1)** ont un profil moyen avec une légère surreprésentation des votes pour les partis de centre-droit ou de centre-gauche ainsi que des listes régionalistes (Lassalle).
- **Les espaces périphériques marginalisés (Periph. 2)** se caractérisent par une forte surreprésentation du vote Bardella et une faiblesse du vote pour l'ensemble des autres partis de gouvernement.

3.3 Régionalisation

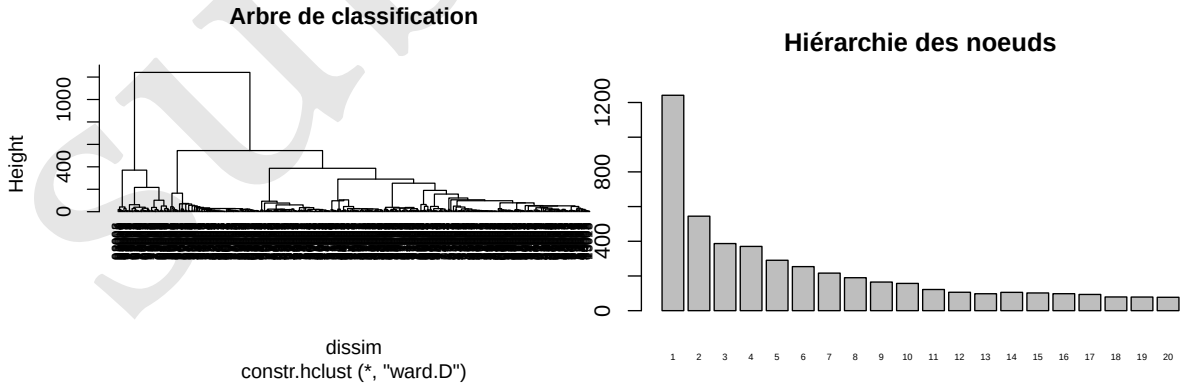
Comme on peut le constater, cette configuration des classes est a priori très défavorable à la constitution de régions sauf à fusionner les différents types mis en évidence par la classification. La carte des discontinuités entre les circonscriptions confirme l'existence de très fortes différences entre les zones urbaines et les espaces périurbainsou ruraux qui les entourent.

```
# Conversion de la matrice de dissimilarité en tableau long
m <- as.matrix(dissim)
tab_dis <- cbind(expand.grid(dimnames(m)), value = as.vector(m))
names(tab_dis ) <- c("i", "j", "DSij")

# Extraction des frontières d'unités spatiale
map_border_circ <- mf_get_borders(mapdon_circ)[, c("circ", "circ.1")]
names(map_border_circ) <- c("i", "j", "geometry")

# Jointure
map_border_circ <- merge(map_border_circ, tab_dis, by = c("i", "j"))

# Cartographie des dissimilarités les plus fortes
mf_map(mapdon_circ, type = "base", col = "lightyellow")
mf_map(map_border_circ,
       type = "prop",
       col = "red",
       var = "DSij",
       val_max = 70,
       leg_pos = "left")
mf_layout("Cartographie des discontinuités", frame = TRUE, credits = "Grasland C., 2025")
```



L'application de l'algorithme de régionalisation conduit pourtant à identifier des niveaux de découpage pertinents en 2, 3, 5 ou 12 régions.

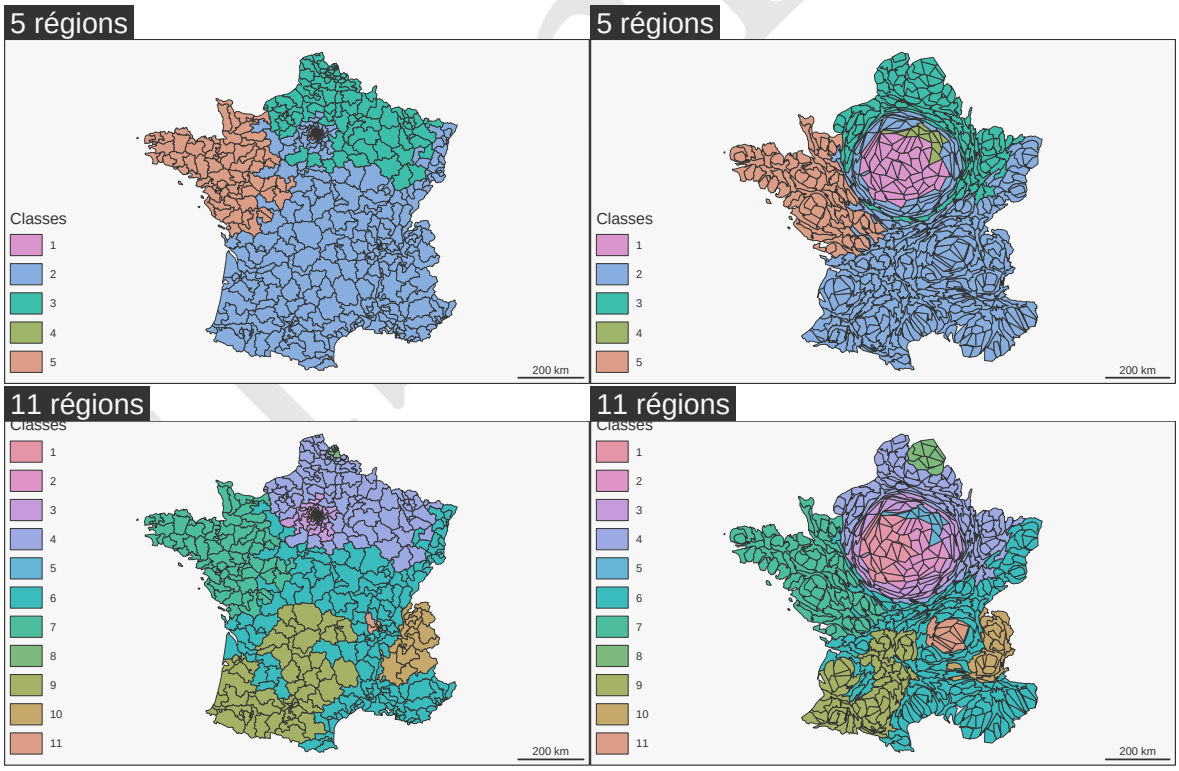
```
# Régionalisation ascendante hiérarchique
regio <- constr.hclust(d = dissim, method = "ward.D", links = circ_link_contig)

# A. Arbre de classification
plot(regio,
     main = "Arbre de classification",
     hang = -1,
     cex = 0.7)
# ---

# B. Hiérarchie des noeuds
barplot(rev(regio$height)[1:20],
        main = "Hiérarchie des noeuds",
        names.arg = 1:20,
        cex.names = 0.4)
```



La cartographie des découpages en 5 et 11 régions produit des résultats intéressants même si leur pouvoir explicatif est plus faible que celui de la classification.

```
# Découpage en 5 classes
```

```r
mapdon_circ$reg5 <- as.factor(cutree(regio, 5))
cartogram_R$reg5 <- as.factor(cutree(regio, 5))

# Découpage en 11 classes
mapdon_circ$reg11 <- as.factor(cutree(regio, 11))
cartogram_R$reg11 <- as.factor(cutree(regio, 11))

# 1.a CAH - 5 classes
mf_map(mapdon_circ, var="reg5",type="typo", leg_title = "Classes",lwd=0.4)
mf_layout("5 régions",credits = "", scale=T, frame=T, arrow=F)
# ---

# 1.b CAH - 5 classes
mf_map(cartogram_R, var="reg5",type="typo", leg_title = "Classes", lwd=0.4)
mf_layout("5 régions",credits = "", scale=T, frame=T, arrow=F)
# ---

# 2.a CAH - 11 classes
mf_map(mapdon_circ, var="reg11",type="typo", leg_title = "Classes", lwd=0.4)
mf_layout("11 régions",credits = "", scale=T, frame=T, arrow=F)
# ---

# 2.b CAH - 11 classes
mf_map(cartogram_R, var="reg11",type="typo", leg_title = "Classes", lwd=0.4)
mf_layout("11 régions",credits = "", scale=T, frame=T, arrow=F)
```



Sans reprendre en détail l'analyse des profils de classe, on voit que la régionalisation en cinq classes est assez proche des résultats obtenus à l'échelle des départements. On retrouve en effet la singu-

larité de l'Ile de France, de la Seine-Saint-Denis, de l'Ouest et dunord du bassin parisien. Quant à la régionalisation en 11 classes, elle met en valeur la singularité des trois plus grandes métropoles provinciales (Lille, Lyon, Marseille) ainsi qu les spécificiés du Sud-Ouest et des Alpes.

Le changement d'échelle ne modifie donc pas radicalement les conclusions obtenues au niveau départemental puisque les métropoles de taille moyenne (Rennes, Nantes, Bordeaux, Toulouse, Strasbourg, ...) sont absorbées par les circonscriptions voisines. Seules les métropoles de taille suffisante pour se subdiviser en plusieurs circonscriptions législatives arrivent à émerger comme régions à cette échelle.

Bibliographie

Benzecri, J. P. (1973). L'Analyse des données : la Taxinomie, vol. 1. *Dunod, Paris*, *31*.

Grasland, C. (1997). L'analyse des discontinuités territoriales : l'exemple de la structure par âge des régions européennes vers 1980. *L'espace géographique*, 309326. https://www.jstor.org/stable/44381820

Guénard, G. et Legendre, P. (2022). Hierarchical clustering with contiguity constraint in R. *Journal of statistical software*, *103*, 126. https://www.jstatsoft.org/article/view/v103i07

Husson, F., Josse, J. et Pagès, J. (2010). *Analyse de données avec R-Complémentarité des méthodes d'analyse factorielle et de classification* (p. nc). https://inria.hal.science/inria-00494779/

Husson, F., Lê, S. et Pagès, J. (2016). *Analyse de données avec R*.

Lê, S., Josse, J. et Husson, F. (2008). FactoMineR : an R package for multivariate analysis. *Journal of statistical software*, *25*, 118. https://www.jstatsoft.org/article/view/v025i01

Legendre, P. et De Cáceres, M. (2013). Beta diversity as the variance of community data : dissimilarity coefficients and partitioning. *Ecology Letters*, *16*(8), 951-963. https://doi.org/10.1111/ele.12141

Legendre, P. et Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, *80*, 107138. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/BF00048036&casa_token=HeSYprqF-4cAAAAA:x09D1Jj79TyMaBGTTT7jzTHkkY372rT2gMRokzrHsDdsg9eZTK

Murtagh, F. et Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward's Criterion ? *Journal of Classification*, *31*(3), 274-295. https://doi.org/10.1007/s00357-014-9161-z

Sanders, L. (1989). L'analyse statistique des données en géographie. *GIP Reclus*. https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=6516569

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Annexes

Préparation des données

A. Données tabulaires

- le fichier **résultats-définitifs-par-circonscriptions.csv** est accesible sur le site data.gouv.fr en suivant ce lien. Il présente les résultats définitifs des élections européennes et a pour origine le Ministère de l'Intérieur. Comme il est très complexe (beaucoup de colonnes redondantes) nous l'avons modifié pour créer des fichiers ne contenant que les colonnes indispensables (effectifs)
- le fichier **candidats-eur-2024.xlsx** est accessible sur le site data.gouv.fr en suivant ce lien. Produit par le ministère de l'intérieur il fournit une information detaillée sur les candidats de chacune des listes. Nous allons en extraire uniquement les caractéristiques des têtes de liste afin de produire un tableau de métadonnées sur les 38 têtes de listes.

— le fichier **indic-stat-circonscriptions-legislatives-2022.xls** a été produit par l'INSEE et est accessible en suivant ce lien. Il fournit un ensemble de données de cadrage sociales et économiques sur les circonscriptions législatives de France à partir des données du recensement de 2022 et de quelques autres sources. Il ne sera pas utilisé directement mais peut servir pour des exercices complémentaires.
— le fichier **circo_composition.xls** également accessible sur le même lien permet de mettre en rapport les circonscription avec les départements, les régions ou les communes. Sachant qu'une même commune peut participer à deux circonscriptions ou plus. On s'en servira principalement pour établir le lien entre circonscriptions et régions.
— le fichier **france_circonsscriptions_legislatives_2012.json** contient un fonds de carte simplifié des circonscriptions législatives en vigueur depuis 2012. C'est une reprise du travail de Toxicode. L'Atelier de cartographie de Sciences Po à ensuite vérifié, nettoyé et généralisé le fond. Il est accessible sur le site data.gouv.fr en suivant ce lien. Il est beaucoup plus léger et mieux généralisé que le fonds de carte fournit par l'INSEE avec les deux ressources précédentes

B. Données géographiques

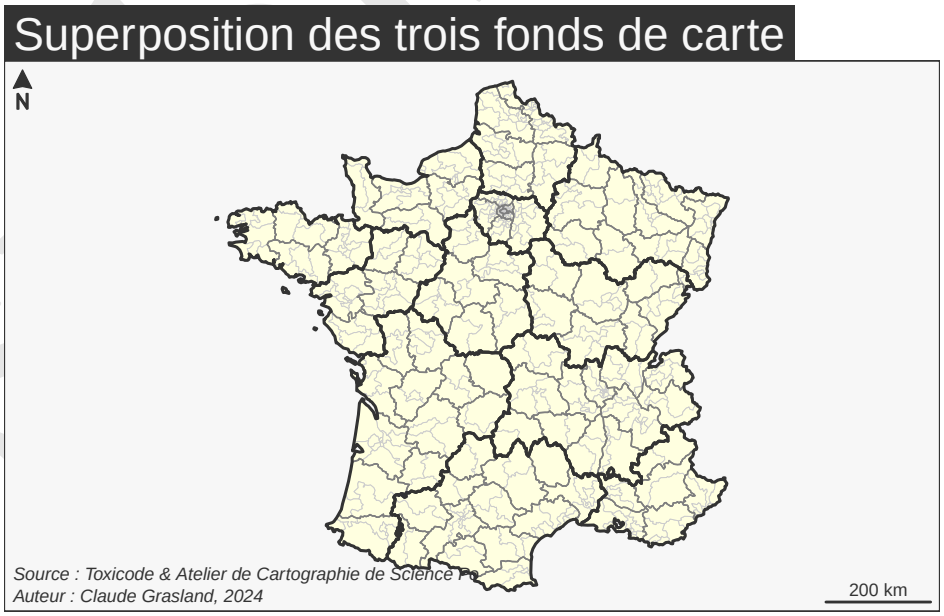
On charge le fichier des circonscriptions en ne conservant que les données de France métropolitaine hors Corse, soit 533 circonscriptions. On le projette en EPSG 2154 puis on l'agrège par département et régions pour disposer de trois fonds de cartes.

```
# Charge le fonds de carte
map <- st_read("data/raw/france-circonscriptions-legislatives-2012.json")


# Charge la table de correspondance entre circonscriptions, départements et régions
meta <- readxl::read_xlsx("data/raw/circo_composition.xlsx", sheet= "table")

# Harmonise les noms et codes de départements et régions
meta <- meta |>
          select(circ=circo,
                 dept= DEP,
                 dept_nom=libdep,
                 regi=REG,
                 regi_nom = libreg) |>
          filter(substr(circ, 1, 2) == dept) |>   # CORRIGE DES ERREURS DE CODAGE DE L'INSEE
          unique()


# Crée la carte des circonscriptions
map_circ <- map |>
              mutate(circ = ID) |>
              select(circ) |>
              left_join (meta) |>
              filter(nchar(dept)<3,                      # Élimine les DROM
                     !dept %in% c("2A", "2B")) |>       # Élimine la Corse
              arrange(regi, dept,circ) |>
              st_transform(2154)                         # Change la projection

saveRDS(map_circ, "data/net/map_circ.RDS")
```

```r
# Agrégation par département
map_dept <- map_circ |>
                group_by(dept) |>
                summarize(dept_nom = min(dept_nom),
                          regi = min(regi),
                          regi_nom = min(regi_nom),
                          .groups = "drop")

saveRDS(map_dept, "data/net/map_dept.RDS")

# Agrégation par région
map_regi <- map_dept |>
                group_by(regi) |>
                summarise(regi_nom = min(regi_nom),
                          .groups = "drop")

saveRDS(map_regi, "data/net/map_regi.RDS")
```

On affiche les trois fonds de carte pour vérification :

```r
# Chargement
map_circ <- readRDS("data/net/map_circ.RDS")
map_dept <- readRDS("data/net/map_dept.RDS")
map_regi <- readRDS("data/net/map_regi.RDS")

# Vérification des fonds de carte
mf_map(map_circ, type = "base", col="lightyellow", border="gray80", lwd=0.4)
mf_map(map_dept, type = "base", col=NA, border="gray50", lwd=0.8, add=T)
mf_map(map_regi, type = "base", col=NA, border="gray20", lwd=1.6, add=T)
mf_layout(title = "Superposition des trois fonds de carte",
          credits = "Source : Toxicode & Atelier de Cartographie de Science Po\nAuteur : Claude
```



C. Données électorales

Nous allons extraire du fichier électoral les variables générales de cadrage (inscrits, votants, blancs,nuls, ...) et les effectifs bruts de vote pour les candidats des différentes listes par circonscription. Ces deux tableaux seront ensuite agrégés par départements et régions

```
x <- read.csv2("data/raw/resultats-definitifs-par-circonscription.csv")

# Modification du code des circonscriptions popur adéquation avec les fonds de carte
z <- x$Code.circonscription.législative
x$circ <- paste0(substr(z, 1, 2), "0", substr(z, 3, 4))


# Variables générales
gen <- x  |>
       select(circ ,
              ins = Inscrits,
              vot = Votants,
              abs = Abstentions,
              bla = Blancs,
              nul = Nuls,
              exp = Exprimés)


# Ajout des clés d'agrégation géographiques
map_circ <- readRDS("data/net/map_circ.RDS")
geo <- st_drop_geometry(map_circ)
gen <- left_join(geo, gen)


# Ajout des suffrages par liste
vot <- x[, substr(names(x), 1, 5) == "Voix."]
a <- rep("vot", 38)
b <- 1:38
z <- paste0(a, b)
names(vot) <- z
vot$circ <- x$circ
vot <- vot[, c(39, 1:38)]

# Tableau final des circonscriptions
don_circ <- left_join(gen, vot)
saveRDS(don_circ, "data/net/don_circ.RDS")
kable(head(don_circ), caption = "Résultats électoraux par circonscription")

# Agrégation par départements
don_dept <- don_circ |> group_by(dept, dept_nom, regi, regi_nom) |> summarise_at(2:45, sum) |>

saveRDS(don_dept, "data/net/don_dept.RDS")
kable(head(don_dept), caption = "Résultats électoraux par département")

# Agrégation par régions
```

```
don_regi <- don_circ |> group_by(regi, regi_nom) |> summarise_at(4:47, sum) |> ungroup()
saveRDS(don_regi, "data/net/don_regi.RDS")
kable(head(don_regi), caption = "Résultats électoraux par région")

# Métadonnées sur les têtes de listes
m <- readxl::read_xlsx("data/raw/candidats-eur-2024.xlsx")
m <- m |> filter(Ordre == 1) |>
  select(
    code = `Numéro de panneau`,
    nom = `Libellé de la liste`,
    tete_nom = `Nom sur le bulletin de vote`,
    tete_prenom = `Prénom sur le bulletin de vote`,
    tete_sexe = Sexe,
    tete_nais = `Date de naissance`
  )

# Ajout de la nuance politique selon le ministère de l'intérieur
typol <- as.character(x[1, substr(names(x), 1, 6) == "Nuance"])
m$typol <- typol
saveRDS(m, "data/net/don_listes.RDS")
kable(m, caption = "Description des listes")
```

Contrôle des données

A. Agrégation

On vérifie tout d'abord que la procédure d'agrégation a bien donné bien les mêmes totaux au niveau des circonscriptions, départements et régions. Il apparaît que pour chaque niveau le nombre total d'inscrits est bien le même et il ne semble pas utile de vérifier les autres colonnes.

```
# Chargement des données statistiques
don_circ <- readRDS("data/net/don_circ.RDS")
don_dept <- readRDS("data/net/don_dept.RDS")
don_regi <- readRDS("data/net/don_regi.RDS")
```

Vérification des sommes

```
sum(don_circ$ins)
```

```
[1] 45704587
```

```
sum(don_dept$ins)
```

```
[1] 45704587
```

```
sum(don_regi$ins)
```

```
[1] 45704587
```

B. Jointure

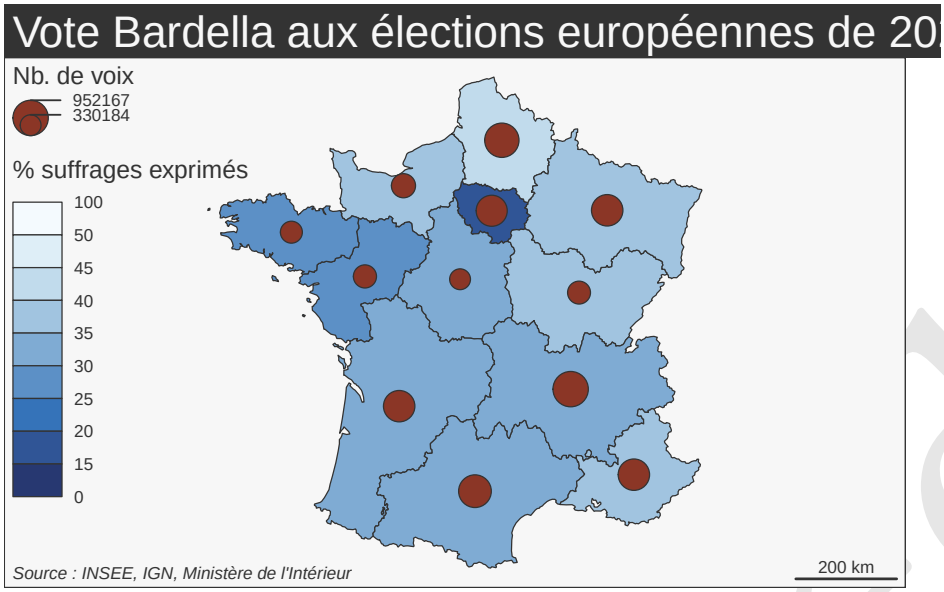
On affiche trois cartes du vote pour la liste n°5 (Bardella) afin de vérifier si les jointures s'opèrent correctement entre données géométriques et statistiques.

```
# Chargement des données géométriques
map_circ <- readRDS("data/net/map_circ.RDS")
```

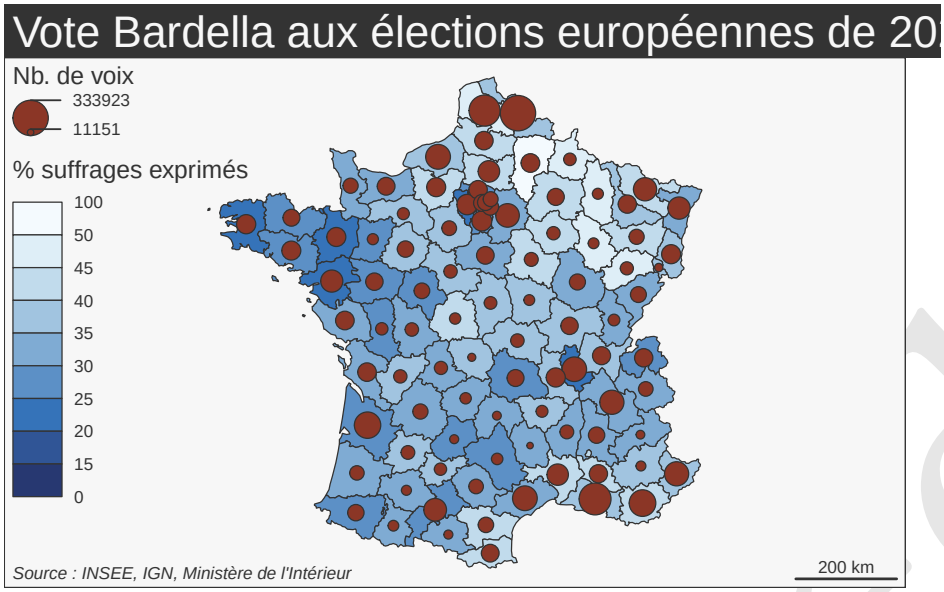
```r
map_dept <- readRDS("data/net/map_dept.RDS")
map_regi <- readRDS("data/net/map_regi.RDS")

# Choix de la palette et des classes
mypal <- hcl.colors(n = 9, palette = "Blues")
mybreaks = c(0, 15, 20, 25, 30, 35, 40, 45, 50, 100)

# Carte par région
mapdon <- left_join(map_regi, don_regi) |>
  mutate(vot = vot5, pct = 100 * vot5 / exp) |>
  select(vot, pct)
```

C. Cartographie

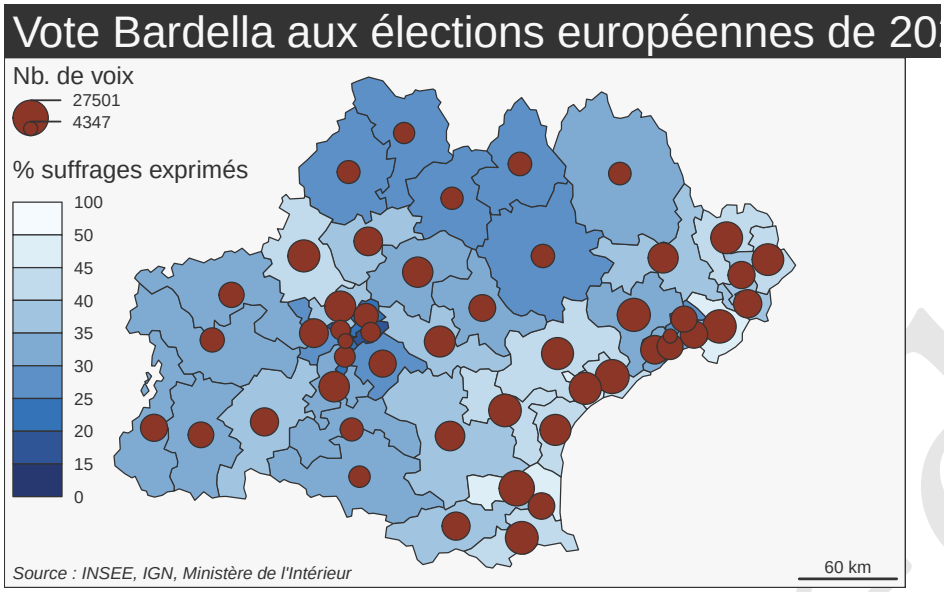
```r
mf_map(
  mapdon,
  type = "choro",
  var = "pct",
  breaks = mybreaks,
  pal = mypal,
  leg_val_rnd = 0,
  leg_pos = "left",
  leg_title = "% suffrages exprimés"
)

mf_map(
  mapdon,
  type = "prop",
  var = "vot",
  inches = 0.1,
  leg_pos = "topleft",
  leg_title = "Nb. de voix"
)

mf_layout(
  title = "Vote Bardella aux élections européennes de 2024 par régions",
  credits = "Source : INSEE, IGN, Ministère de l'Intérieur",
  frame = T,
  arrow = F
)
```




```
# Carte par département
mapdon <- left_join(map_dept, don_dept) |>
              mutate(vot = vot5, pct = 100 * vot5 / exp) |>
              select(vot, pct)

mf_map(
  mapdon,
  type = "choro",
  var = "pct",
  breaks = mybreaks,
  pal = mypal,
  leg_val_rnd = 0,
  leg_pos = "left",
  leg_title = "% suffrages exprimés"
)

mf_map(
  mapdon,
  type = "prop",
  var = "vot",
  inches = 0.1,
  leg_pos = "topleft",
  leg_title = "Nb. de voix"
)

mf_layout(
  title = "Vote Bardella aux élections européennes de 2024 par régions",
  credits = "Source : INSEE, IGN, Ministère de l'Intérieur",
  frame = T,
  arrow = F
)
```




```
# Carte par circonscription (zoom sur l'Occitanie)
mapdon <- left_join(map_circ, don_circ) |>
                filter(regi_nom == "Occitanie") |>
                mutate(vot = vot5, pct = 100 * vot5 / exp) |>
                select(vot, pct)

mf_map(
  mapdon,
  type = "choro",
  var = "pct",
  breaks = mybreaks,
  pal = mypal,
  leg_val_rnd = 0,
  leg_pos = "left",
  leg_title = "% suffrages exprimés"
)

mf_map(
  mapdon,
  type = "prop",
  var = "vot",
  inches = 0.1,
  leg_pos = "topleft",
  leg_title = "Nb. de voix"
)

mf_layout(
  title = "Vote Bardella aux élections européennes de 2024 par circonscription en région Occita
  credits = "Source : INSEE, IGN, Ministère de l'Intérieur",
  frame = T,
  arrow = F
)
```



Bibliographie

Annexes

Informations de session

setting	value
version	R version 4.5.1 (2025-06-13)
os	Fedora Linux 42 (Workstation Edition)
system	x86_64, linux-gnu
ui	X11
language	(EN)
collate	fr_FR.UTF-8
ctype	fr_FR.UTF-8
tz	Europe/Paris
date	2025-08-13
pandoc	3.6.3 @ /usr/libexec/rstudio/bin/pandoc/ (via rmarkdown)
quarto	1.7.33 @ /usr/bin/quarto

package	ondiskversion	source
adespatial	0.3.28	CRAN (R 4.5.0)
cartogramR	1.5.1	CRAN (R 4.5.0)
dplyr	1.1.4	CRAN (R 4.5.1)
ggplot2	3.5.2	CRAN (R 4.5.0)
ggrepel	0.9.6	CRAN (R 4.5.0)
gt	1.0.0	CRAN (R 4.5.0)
ineq	0.2.13	CRAN (R 4.5.0)
knitr	1.50	CRAN (R 4.5.0)
mapsf	0.12.0	CRAN (R 4.5.0)

sf	1.0.20	CRAN (R 4.5.0)
spData	2.3.4	CRAN (R 4.5.0)
spdep	1.3.13	CRAN (R 4.5.0)
stargazer	5.2.3	CRAN (R 4.5.0)