Statistique

Benjamin Bobbia

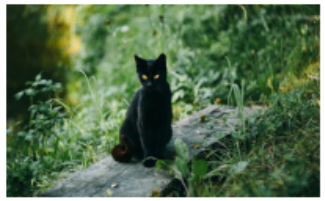
ISAE

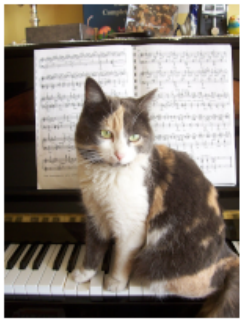
-1.1 Classification

Problème de la classification

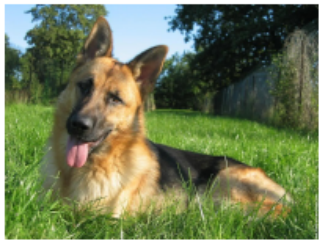
Objectif : regrouper des objets $x_1, \ldots, x_n \in \mathcal{E}$ qui se « ressemblent ».








Problème de la classification

Objectif : regrouper des objets $x_1, \dots, x_n \in \mathcal{E}$ qui se « ressemblent ».

Le problème de la classification est **moins bien posé** que celui de l'apprentissage supervisé car **les classes ne sont pas connues a priori**.

Plusieurs questions se posent :

- que savons-nous de l'**espace** $\mathcal{E}$?
- existe-t-il une **« bonne » classification** ?
- connaissons-nous le **nombre de classes** a priori ?
- comment mesurons-nous la **« ressemblance »** ?
- pouvons-nous définir une notion de **similitude entre les objets** ?
- pouvons-nous définir une notion de **similitude entre des groupes d'objets** ?
- ...

Problème de la classification

Objectif : regrouper des objets $x_1, \ldots, x_n \in \mathcal{E}$ qui se « ressemblent ».

Le problème de la classification est **moins bien posé** que celui de l'apprentissage supervisé car **les classes ne sont pas connues a priori**.

Il existe un (très) grand nombre de méthodes pour aborder ce problème de la classification. Certaines de ces méthodes feront l'objet d'autres cours et nous nous concentrons ici sur deux approches « classiques » :

- **Agrégation autour de centres mobiles** (*a.k.a.* K-means)
  $\longrightarrow$ nombre de classes **connu** a priori
- **Classification ascendante hiérarchique**
  $\longrightarrow$ nombre de classes **inconnu** a priori

Agrégation autour de centres mobiles

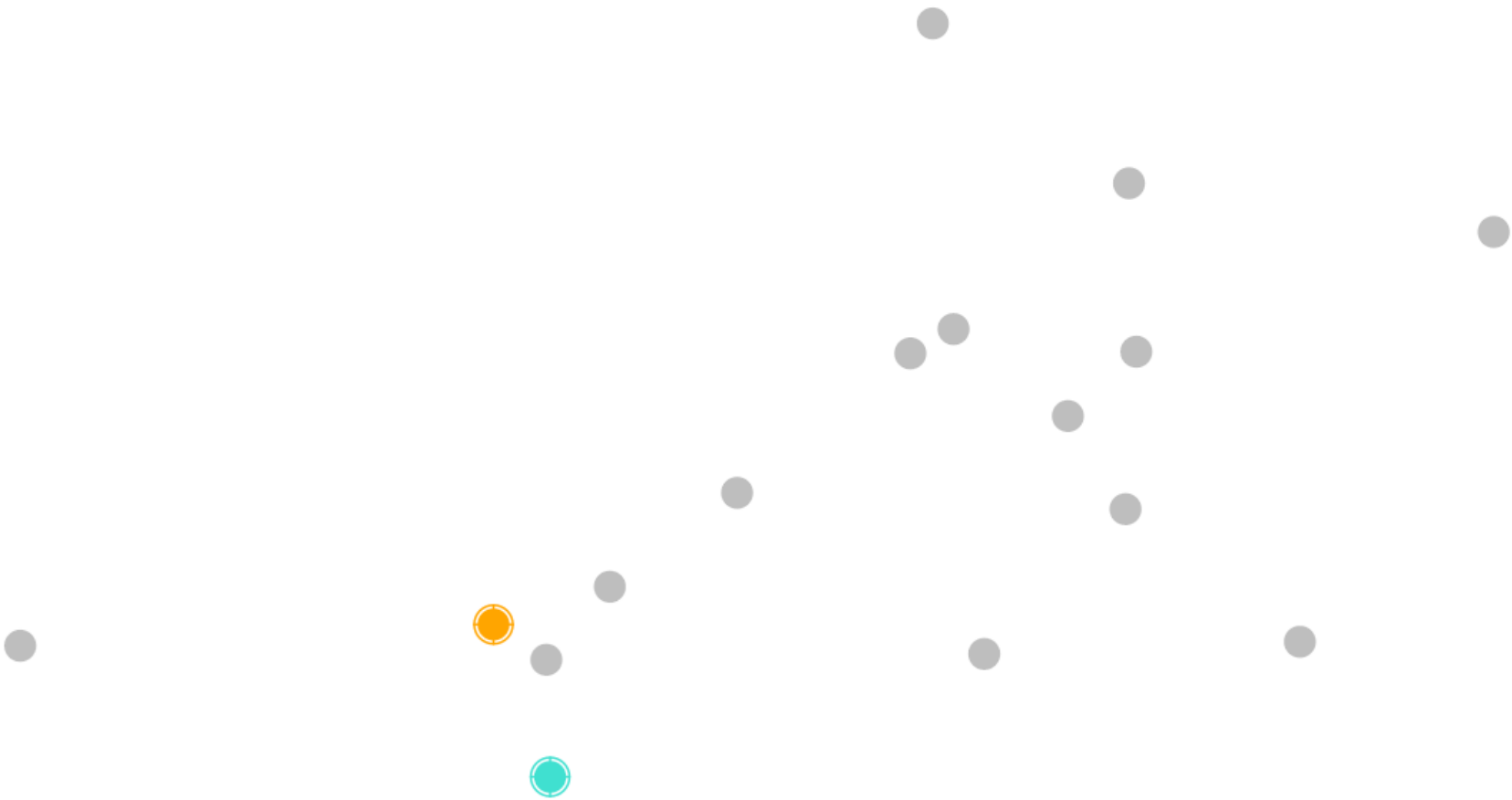
Prérequis : déterminer le **nombre K de classes** soit par une connaissance a priori du phénomène étudié, soit par une autre méthode (nous en reparlerons plus tard).

Algorithme

1. Initialiser K centres distincts (tirages aléatoires ou choix imposés)
2. Répéter les étapes suivantes :
   - Affecter chaque objet au centre le plus proche
   - Recalculer les centres de chaque groupe
3. Terminer lorsque les objets ne changent plus de groupe entre 2 itérations successives

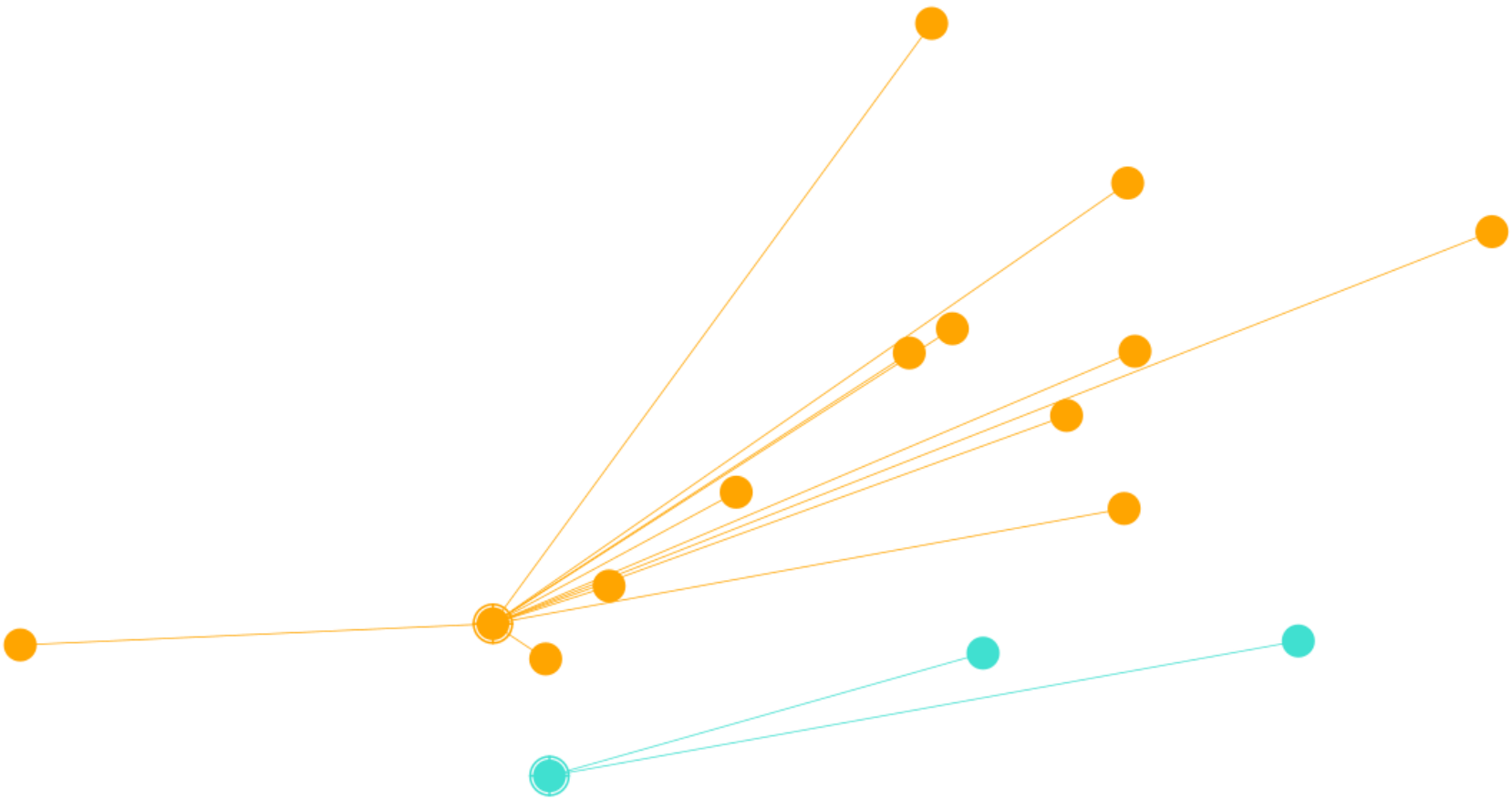
À l'issue de cet algorithme, nous obtenons une classification des données en K groupes.

Illustration de l'algorithme ($K = 2$)



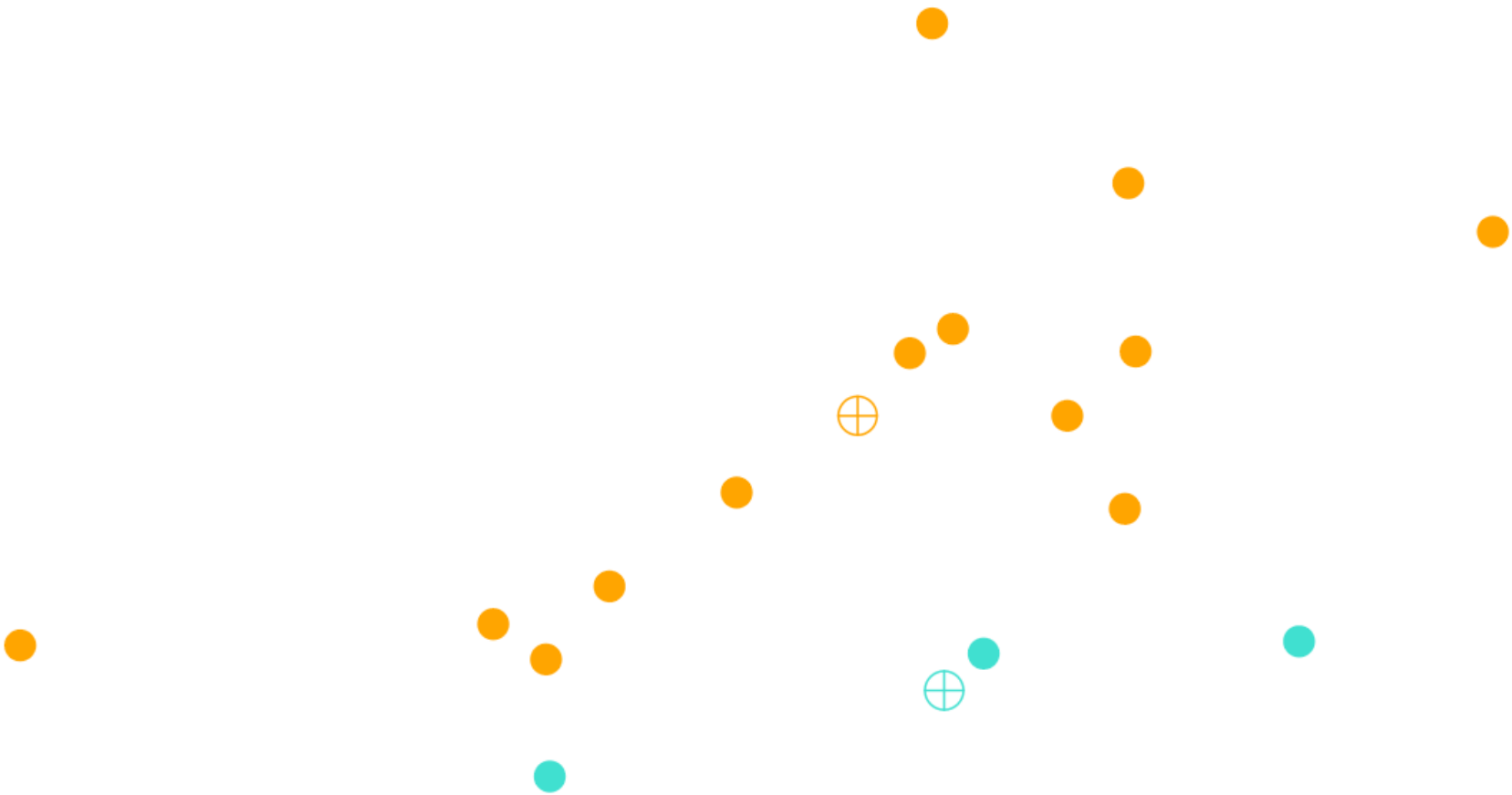
Initialisation de 2 centres aléatoires

Illustration de l'algorithme ($K = 2$)



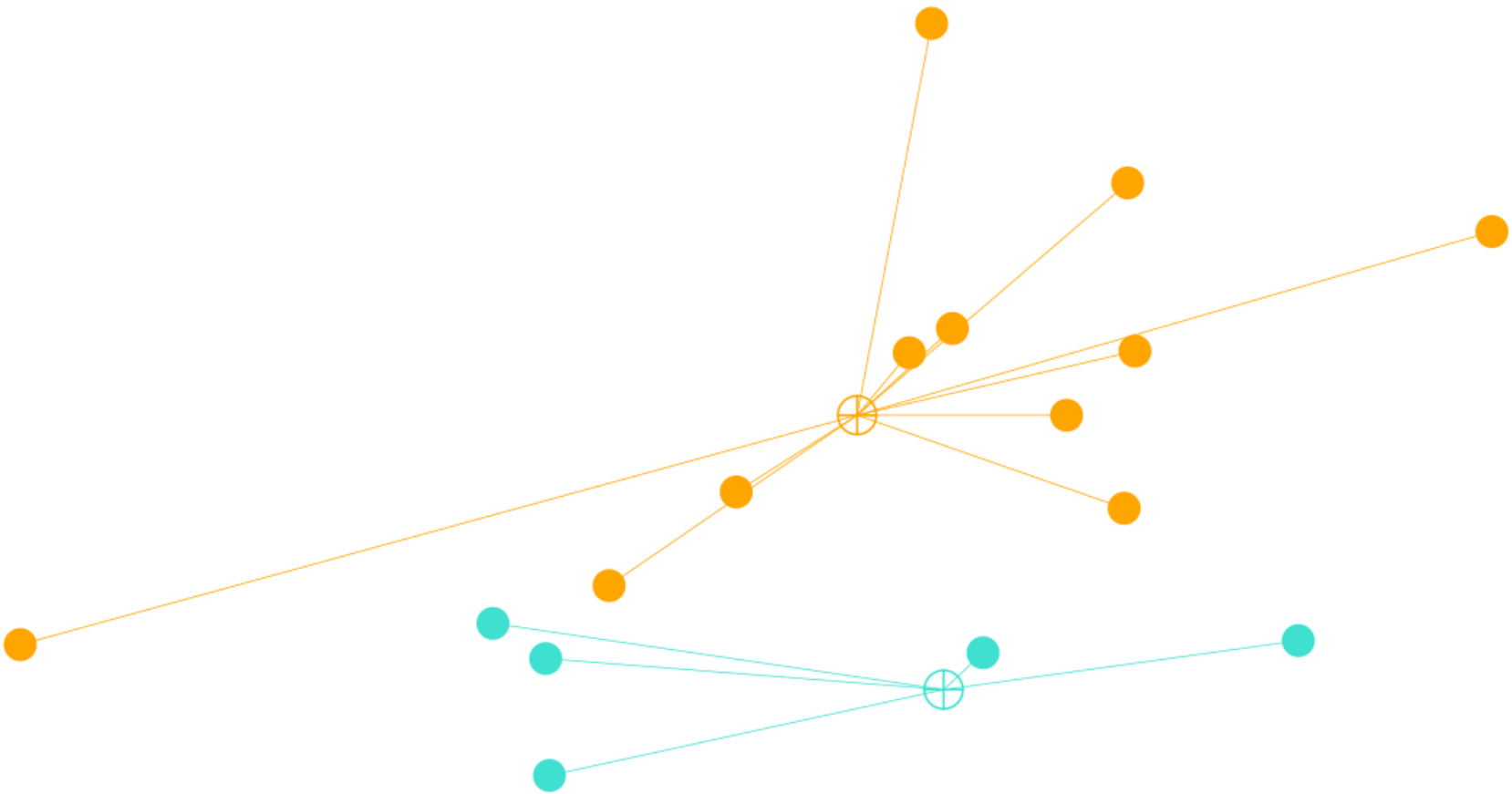
Affectation de chaque point au centre le plus proche

Illustration de l'algorithme ($K = 2$)



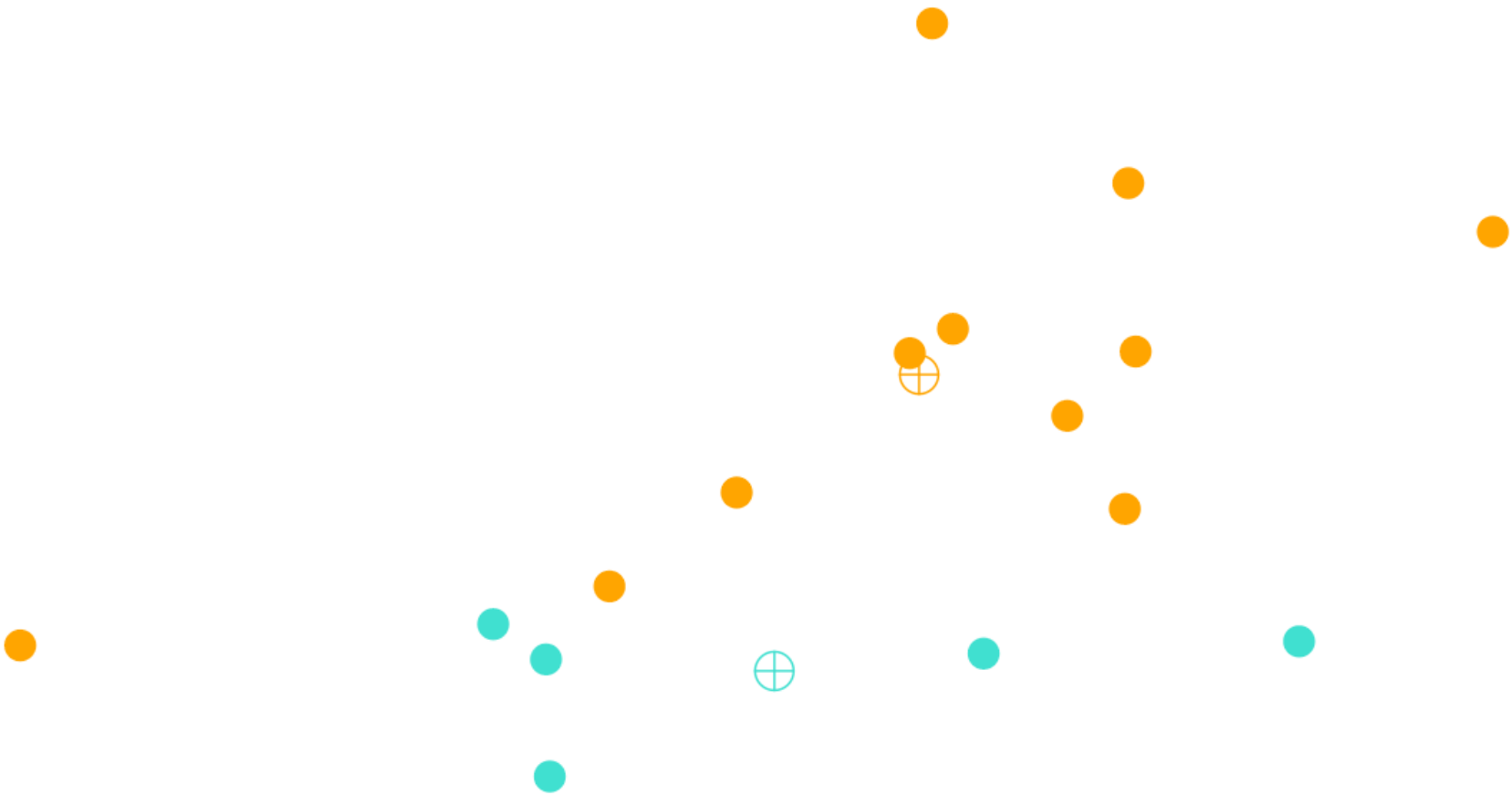
Mise à jour des centres

Illustration de l'algorithme ($K = 2$)



Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)

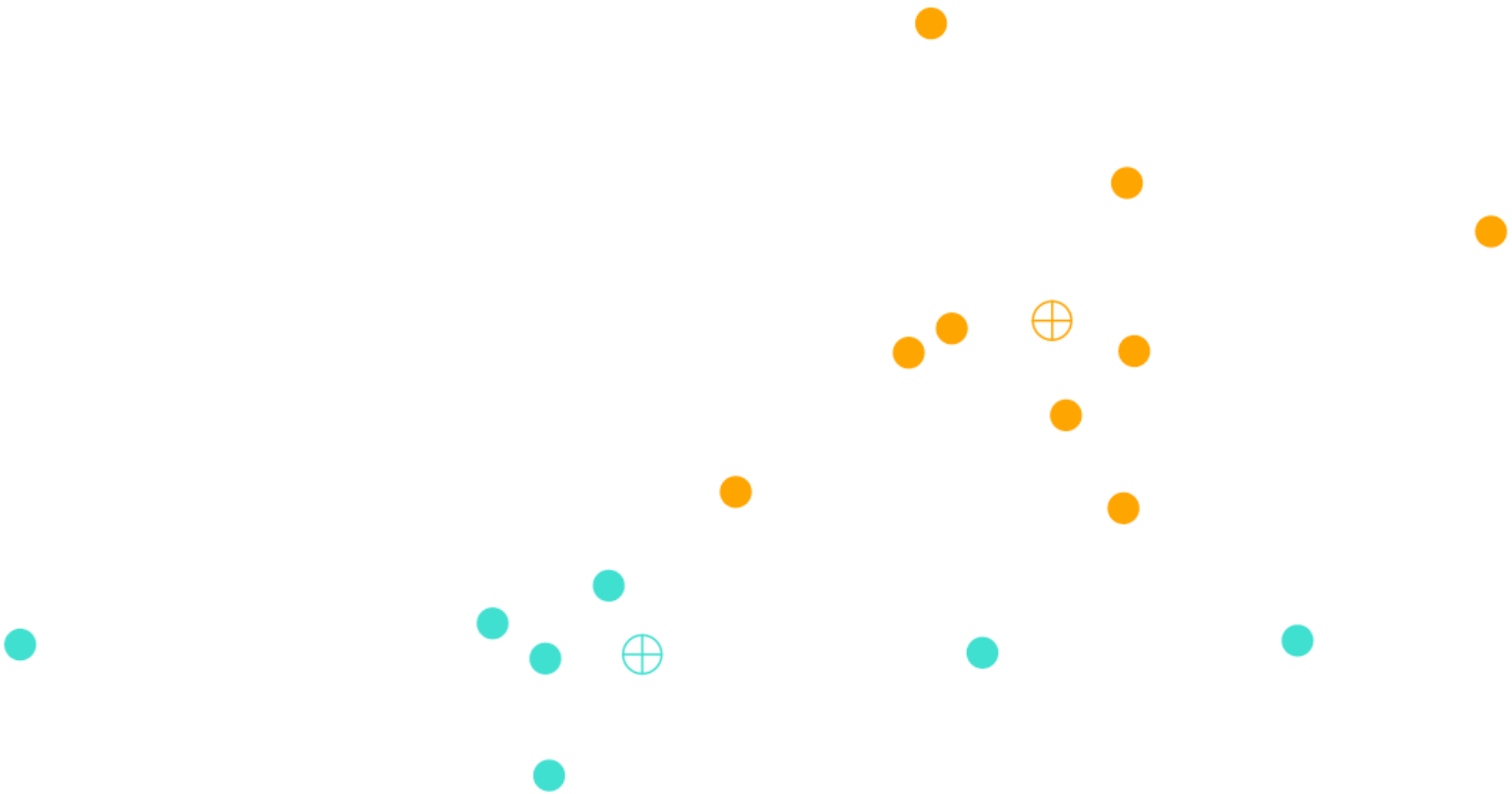


Mise à jour des centres (on itère . . .)

Illustration de l'algorithme ($K = 2$)

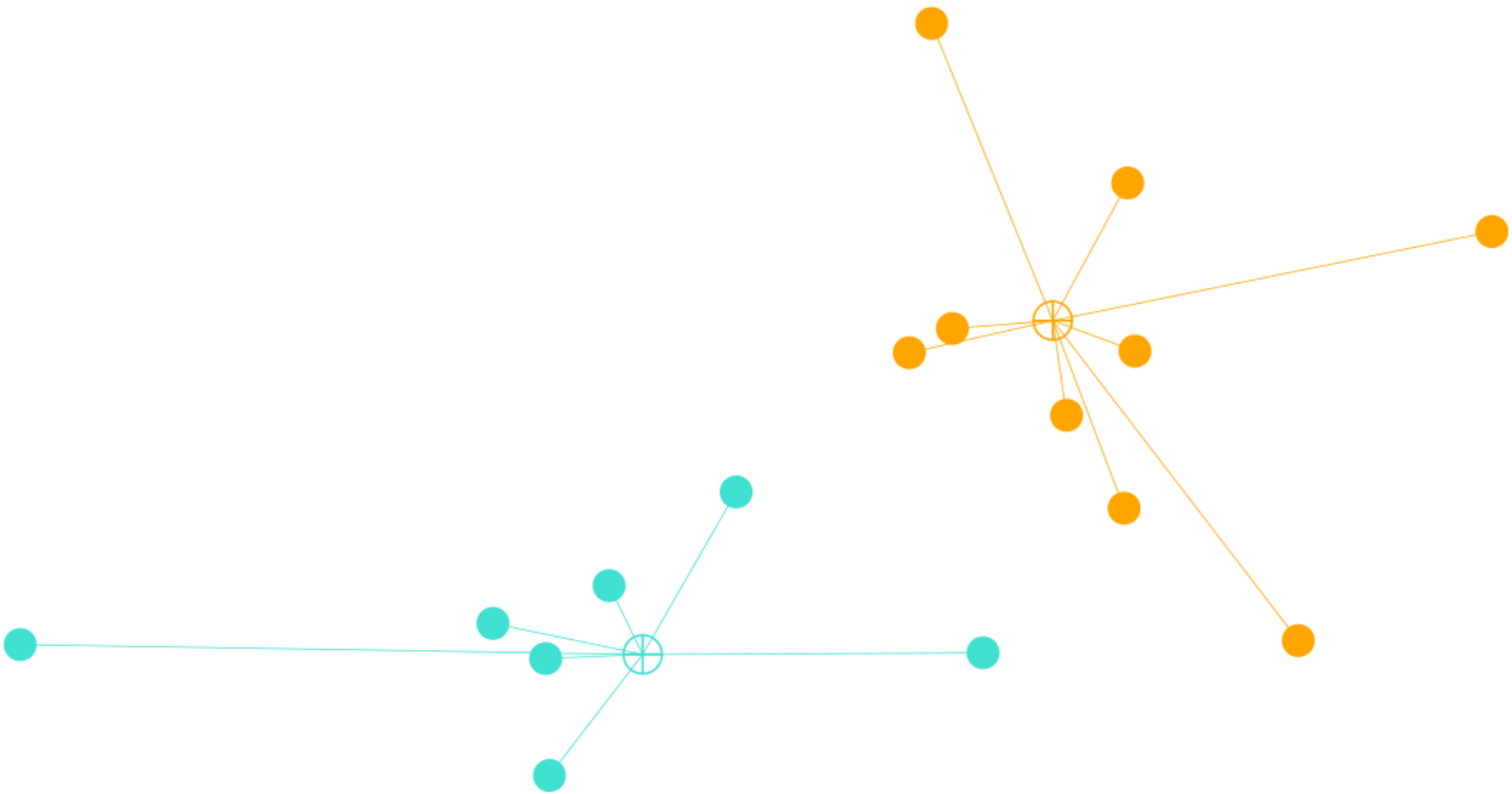
Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)



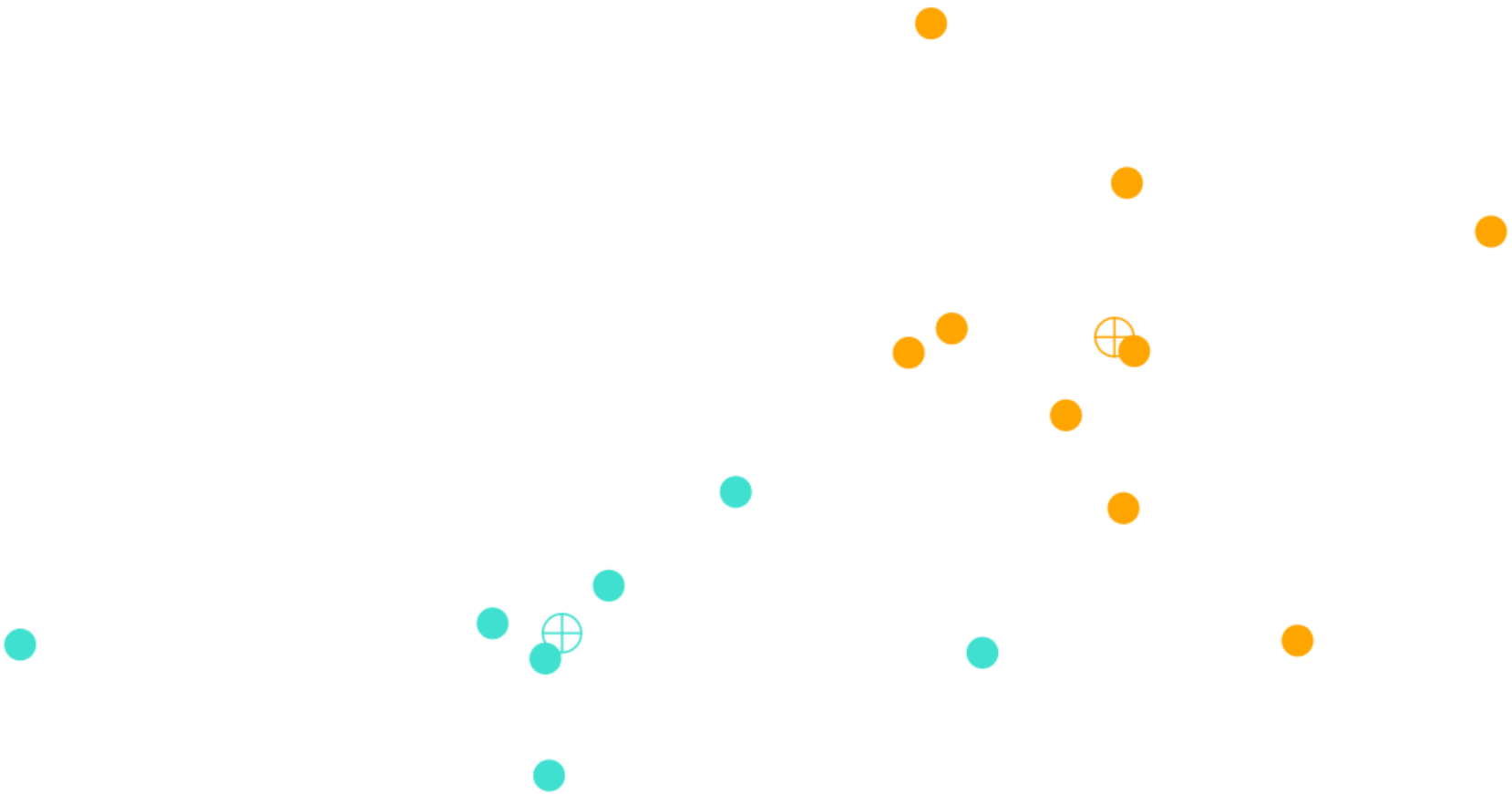
Mise à jour des centres (on itère . . .)

Illustration de l'algorithme ($K = 2$)



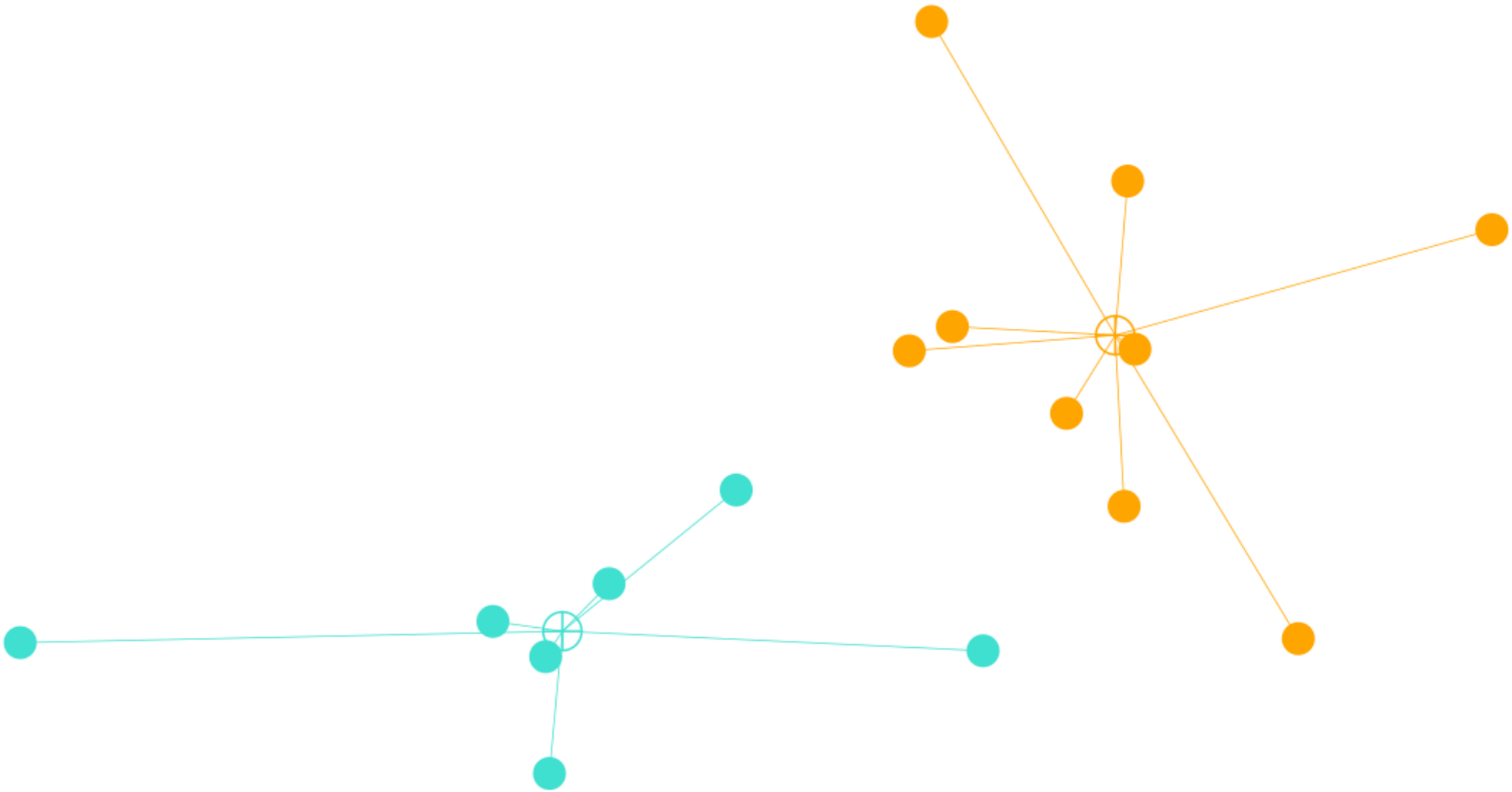
Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)



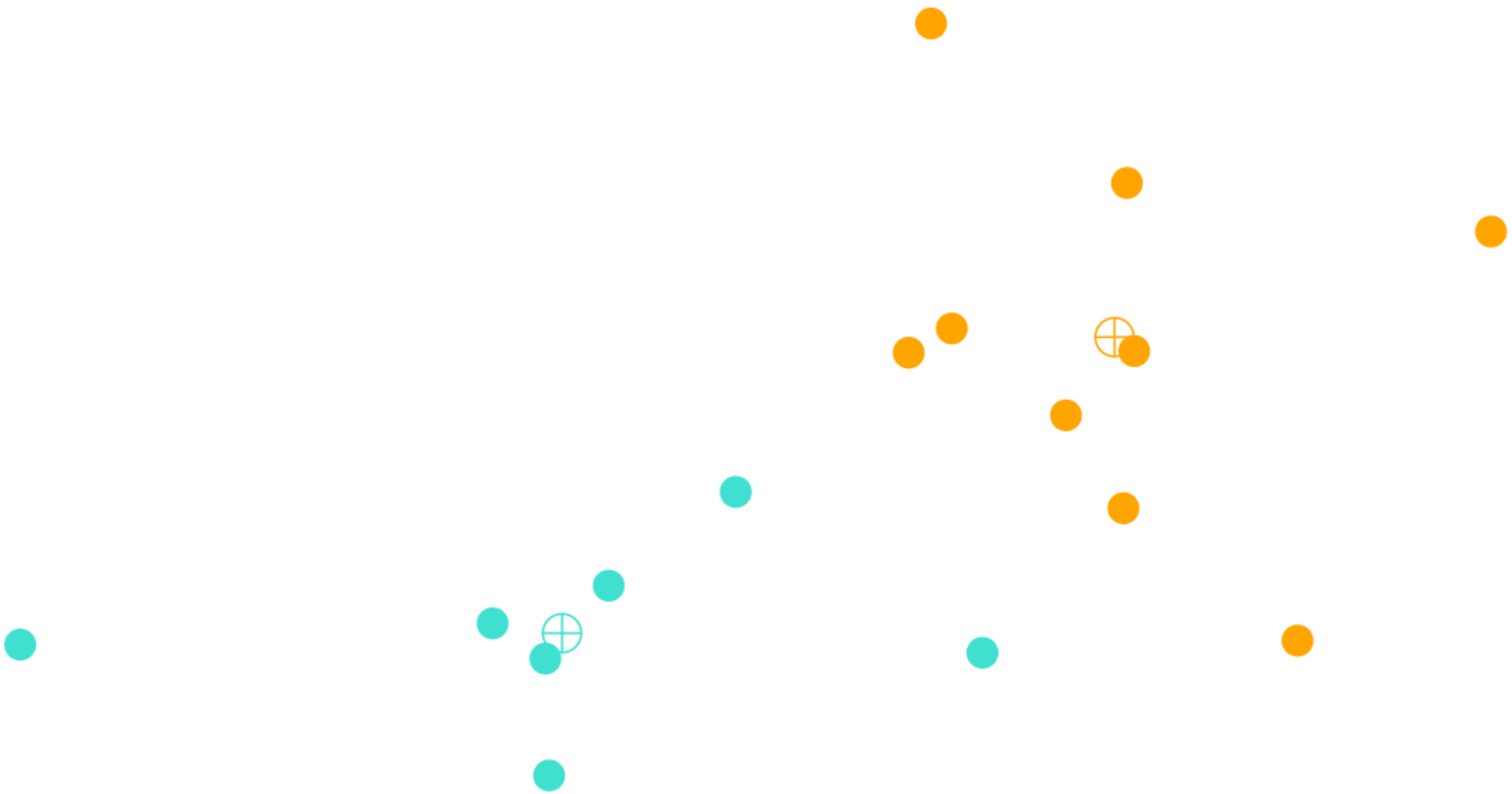
Mise à jour des centres (on itère . . .)

Illustration de l'algorithme ($K = 2$)



Affectation de chaque point au centre le plus proche, rien ne change !

Illustration de l'algorithme ($K = 2$)



Classification finale des données en 2 groupes

Cas des centres explicites

Un cas particulier est celui où les **deux conditions suivantes** sont réunies :

- la **similitude** entre deux objets se mesure à l'aide d'une fonction $s : \mathcal{E}^2 \to \mathbb{R}$ (*e.g.* distance, variance, corrélation, ...),
- les centres $c_1, \ldots, c_K$ des groupes respectifs $G_1, \ldots, G_K$ peuvent être calculés **explicitement** tels que

$$\forall m \in \{1, \ldots, K\},\ c_m \text{ minimise } c \in \mathcal{E} \mapsto \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c)$$

où n_m désigne la taille du groupe G_m.

Dans ce cadre, l'algorithme précédent est une **méthode de minimisation d'un critère de variabilité intra-groupe**.

$$(c_1, \ldots, c_K) \in \mathcal{E}^K \longmapsto \frac{1}{K} \sum_{m=1}^{K} \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c_m).$$

Cas des centres explicites

Un cas particulier est celui où les **deux conditions suivantes** sont réunies :

- la **similitude** entre deux objets se mesure à l'aide d'une fonction $s : \mathcal{E}^2 \to \mathbb{R}$ (*e.g.* distance, variance, corrélation, ...),
- les centres $c_1, \ldots, c_K$ des groupes respectifs $G_1, \ldots, G_K$ peuvent être calculés **explicitement** tels que

$$\forall m \in \{1, \ldots, K\},\ c_m \text{ minimise } c \in \mathcal{E} \mapsto \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c)$$

où n_m désigne la taille du groupe G_m.

Dans ce cadre, l'algorithme précédent est une **méthode de minimisation d'un critère de variabilité intra-groupe**.

Ce critère admet généralement des **minima locaux**. Comme dans le cas général, la solution trouvée peut **ne pas être optimale** et dépendre **fortement** de l'initialisation.

Initialisation de l'algorithme

Exemple de l'influence du **tirage aléatoire** des centres initiaux.

Initialisation de l'algorithme

Lorsque les centres initiaux sont tirés **au hasard**, des exécutions successives de l'algorithme peuvent conduire à des **classification différentes**. Pour minimiser l'impact de cette initialisation aléatoire, nous pouvons :

- relancer la procédure plusieurs fois et affecter les objets à une classe selon un principe de **vote majoritaire**,
- imposer un choix **non aléatoire** des centres initiaux (nous en reparlerons bientôt),
- renforcer la procédure en **imposant** à certains objets d'être toujours dans le **même groupe**.

Agrégation autour de centres mobiles (variantes)

L'algorithme des centres mobiles est simple à mettre en œuvre en pratique et il en existe plusieurs variantes :

- les centres peuvent être recalculés **après chaque affectation** d'un objet à un groupe, il s'agit des **nuées dynamiques**. Cette variante se stabilise plus rapidement mais accroît le **risque d'une solution sous optimale**.
- lorsque nous ne disposons pas des objets eux-mêmes mais seulement de la **matrice de similitude** S (*i.e.* les mesures $S_{kk'}$ des similitudes entre toutes les paires d'objets $(x_k, x_{k'}) \in \mathcal{E}^2$), alors les centres doivent être définis comme les **objets les plus « centraux »** des groupes selon un **critère de variabilité intra-groupe approché**,

$$\forall m \in \{1, \ldots, K\},\ c_m = x_{k_m} \text{ où } k_m \text{ minimise } k' \in G_m \mapsto \frac{1}{n_m} \sum_{k \in G_m} S_{kk'}.$$

- ...

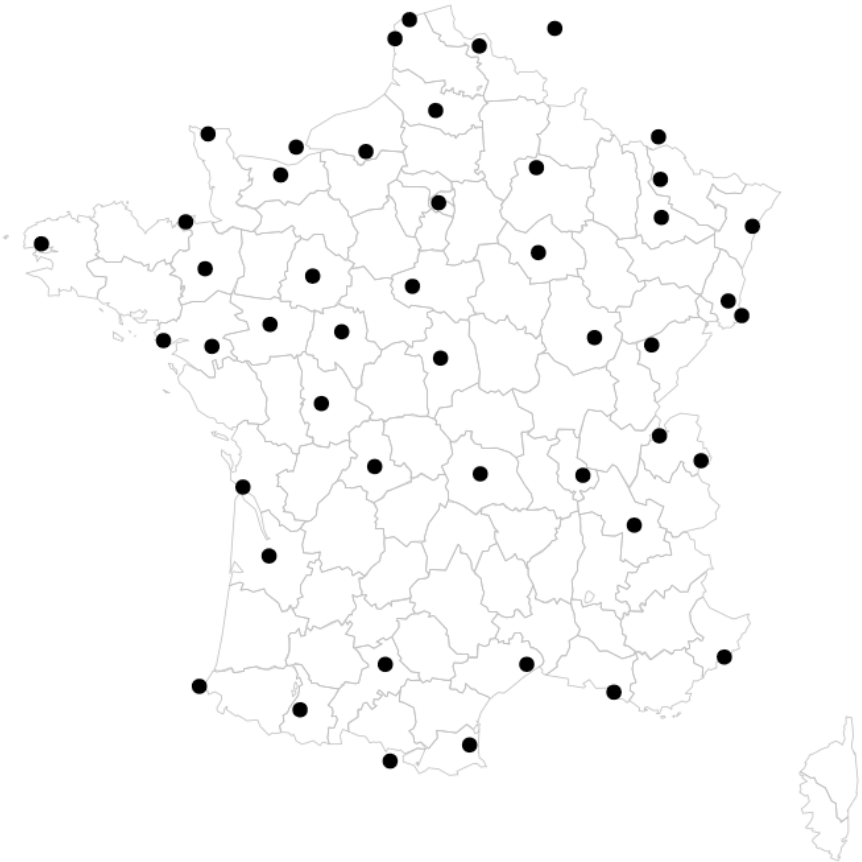
Exemple géographique (distances IGN)

Pour $n = 47$ villes de France ou frontalières, nous mesurons la « similitude » entre deux villes avec la distance IGN. Les données brutes correspondent donc à la matrice de similitude suivante,

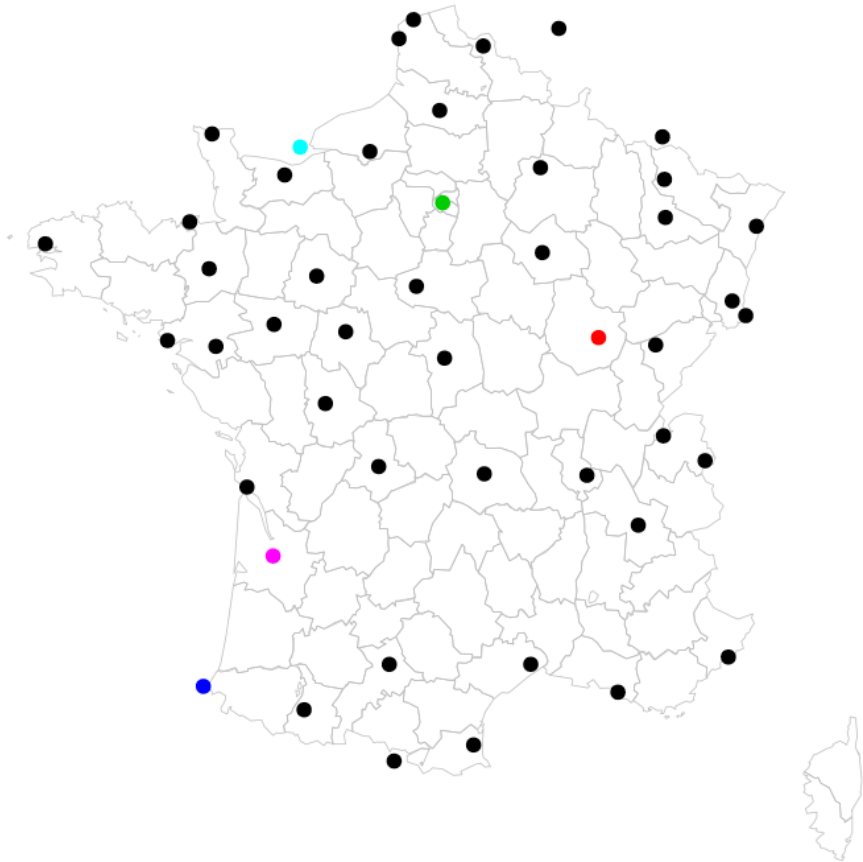
	Amiens	Andorre	Angers	Bâle	LaBaule	Besançon	Bordeaux	Boulogne	Bourges	Brest	Bruxelles	Caen	...
Amiens	0	1020	440	560	590	560	730	120	380	610	210	240	...
Andorre	1020	0	760	1130	830	970	430	1020	680	1130	1200	950	...
Angers	440	760	0	770	160	620	340	480	260	380	600	220	...
Bâle	560	1130	770	0	940	160	840	690	500	1090	560	800	...
LaBaule	590	830	160	940	0	770	400	550	430	270	760	350	...
Besançon	560	970	620	160	770	0	700	610	350	960	550	640	...
Bordeaux	730	430	340	840	400	700	0	830	400	620	890	580	...
Boulogne	120	1020	480	690	550	610	830	0	480	690	260	300	...
Bourges	380	680	260	500	430	350	400	480	0	630	550	360	...
Brest	610	1130	380	1090	270	960	620	690	630	0	910	370	...
Bruxelles	210	1200	600	560	760	550	890	260	550	910	0	450	...
Caen	240	950	220	800	350	640	580	300	360	370	450	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	

Selon les variantes envisagées, nous utiliserons uniquement ces distances ou les données GPS de chaque agglomération.

Exemple géographique ($K = 5$)

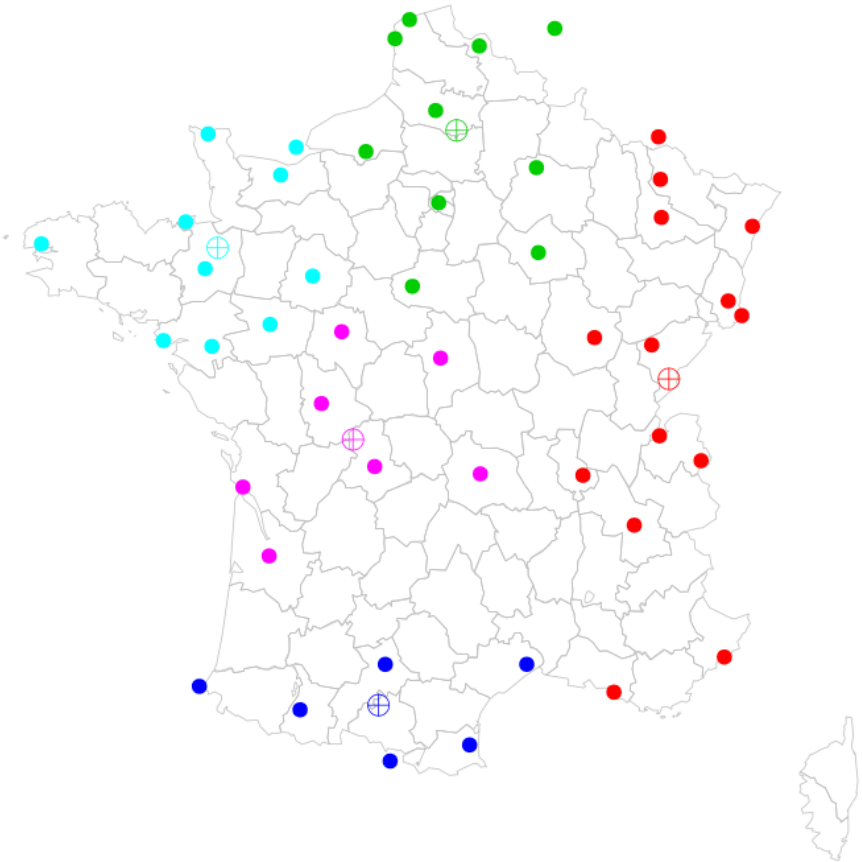


Positions des 47 villes

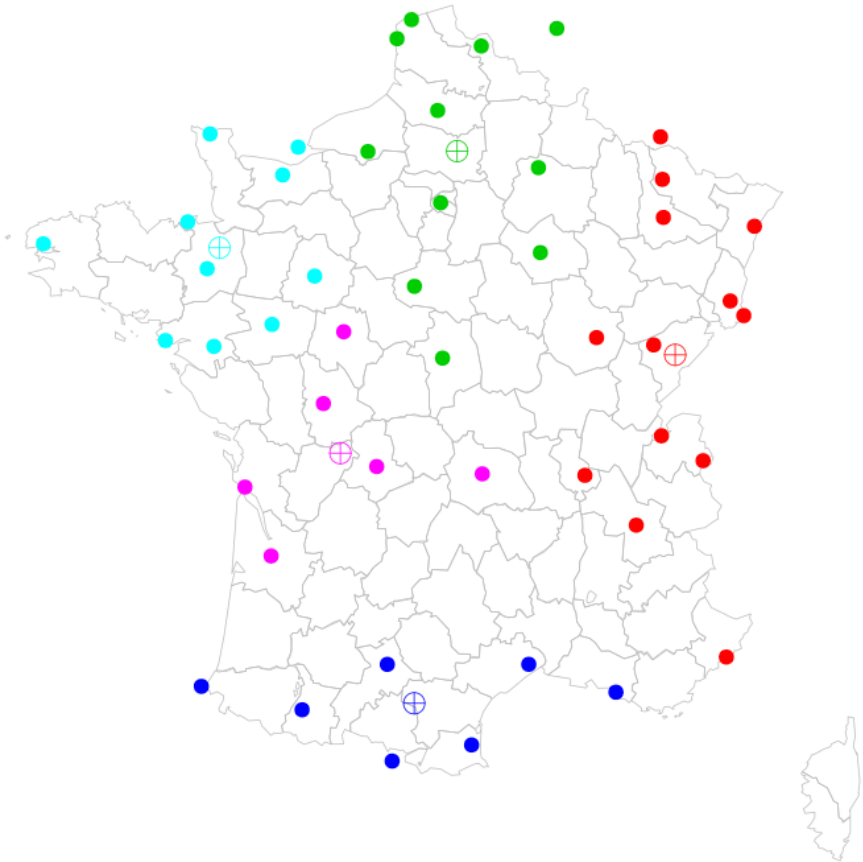


Initialisation de 5 centres aléatoires
(Dijon, Paris, Hendaye, Le Havre, Bordeaux)

Exemple géographique ($K = 5$)

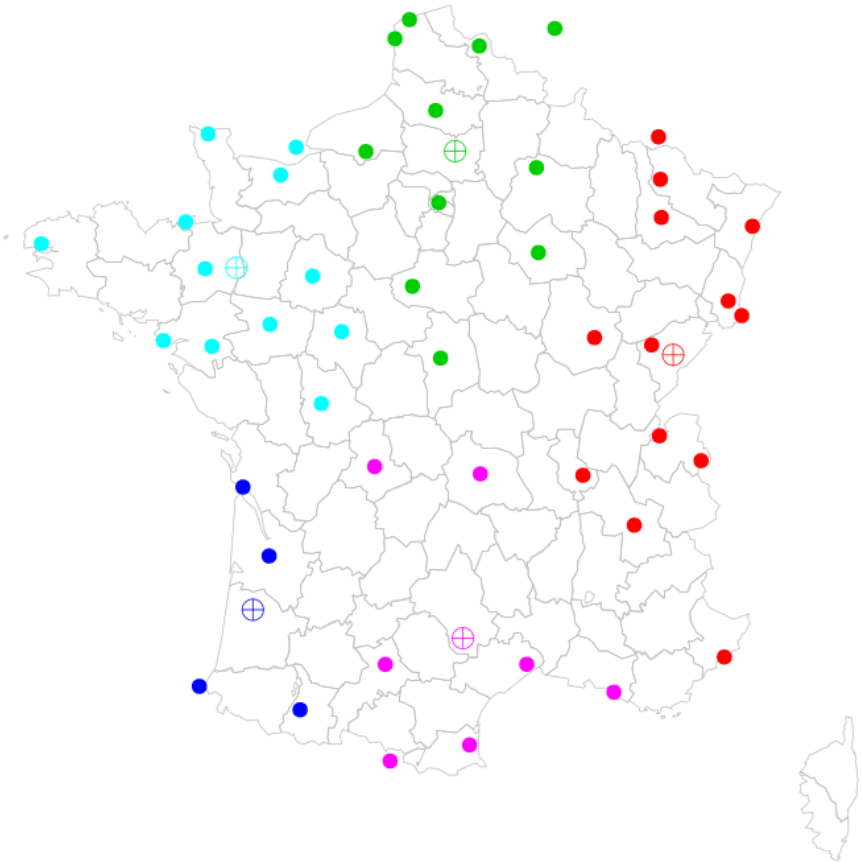


Centres mobiles classiques
(distance L^2, 6 itérations)

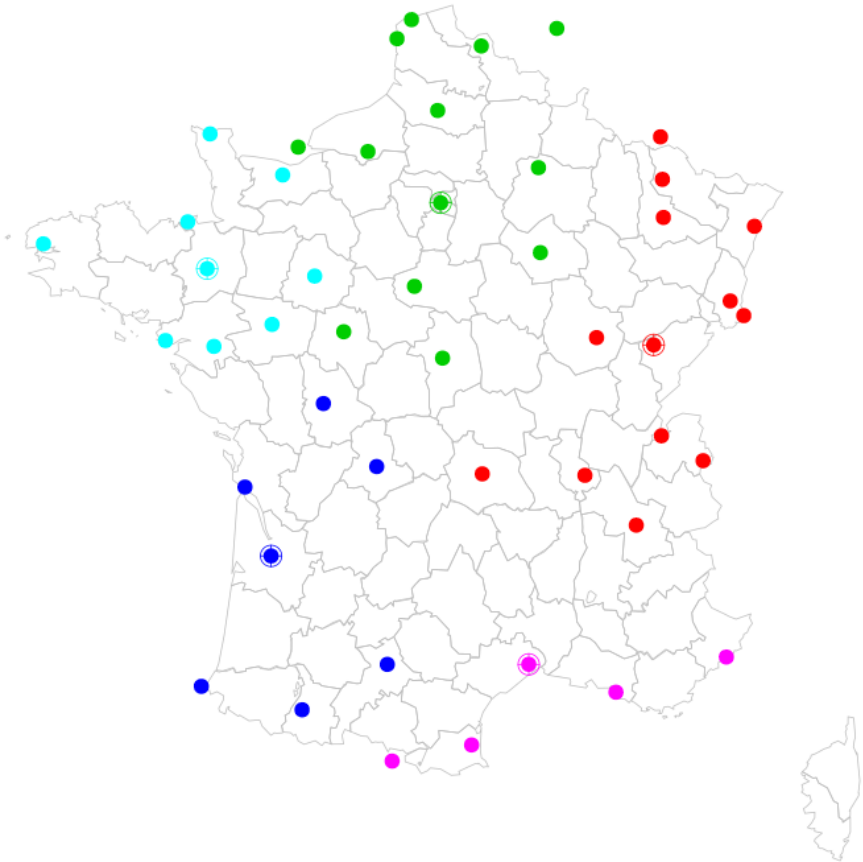


Centres mobiles classiques
(distance L^1, 4 itérations)

Exemple géographique ($K = 5$)



Nuées dynamiques
(distance L^2, 3 itérations)



Matrice de similitude uniquement
(5 itérations)

Classification ascendante hiérarchique (CAH)

Prérequis : choisir un **critère d'agglomération** pour donner un sens à la notion de **similitude entre des groupes d'objets**.

Algorithme

1. Initialiser n groupes « singletons » contenant chacun un objet
2. Répéter : regrouper les deux groupes les plus proches au sens du critère d'agglomération
3. Terminer lorsque il n'y a plus qu'un seul groupe contenant les n objets

À l'issue de cet algorithme, nous obtenons un diagramme appelé **dendrogramme** qui décrit les agglomérations effectuées.

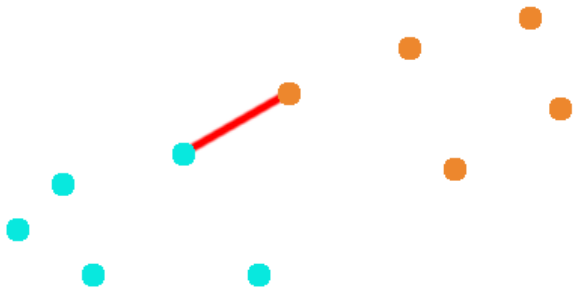
Critères d'agglomération (*linkage*)

Objectif : pour deux ensembles $A, B \subset \{1, \dots, n\}$ **disjoints**, nous voulons donner un sens à la **similitude entre les groupes d'objets**

$$\{x_k,\ k \in A\} \qquad \text{et} \qquad \{x_k,\ k \in B\}.$$

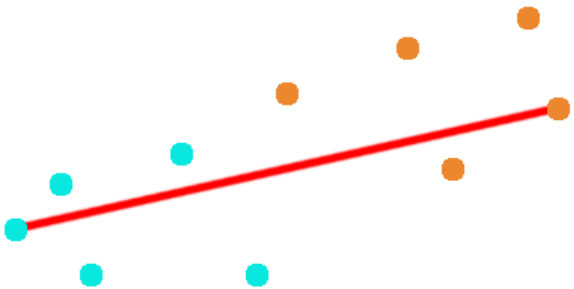
Si nous ne disposons que de la matrice S des similitudes $S_{kk'} = s(x_k, x_{k'})$ entre les objets x_k et $x_{k'}$ pour tout $k, k' \in \{1, \dots, n\}$, alors nous pouvons définir les critères suivants :

Minimum (*single*)



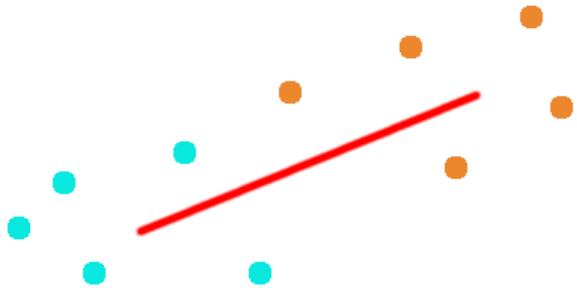
$$\min_{k \in A, k' \in B} S_{kk'}$$

Maximum (*complete*)



$$\max_{k \in A, k' \in B} S_{kk'}$$

Moyen (*average*)



$$\frac{1}{n_A n_B} \sum_{k \in A, k' \in B} S_{kk'}$$

Critères d'agglomération (*linkage*)

Objectif : pour deux ensembles $A, B \subset \{1, \dots, n\}$ **disjoints**, nous voulons donner un sens à la **similitude entre les groupes d'objets**

$$\{x_k,\ k \in A\} \qquad \text{et} \qquad \{x_k,\ k \in B\}.$$

Si il est possible de manipuler la **fonction de similitude** s et de **calculer explicitement** les « objets moyens » $c_A, c_B \in \mathcal{E}$ des deux groupes d'objets, alors la similitude entre ces centres peut être utilisée,

$$s(c_A, c_B).$$

Ce critère ne tient pas compte des tailles n_A et n_B des groupes, ce qui rend son interprétation difficile. Cette similitude peut être renormalisée de façon à correspondre à la **perte d'inertie inter-groupe** associée au regroupement de A et B. Cette méthode est très utilisée en pratique et s'appelle le **critère de Ward**,

$$\frac{n_A n_B}{n_A + n_B} s(c_A, c_B).$$

Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Nous disposons des similitudes entre $n = 5$ objets (données brutes ou calcul avec une fonction s).

	I1	I2	I3	I4	I5
I1	·				
I2	3.61	·			
I3	5.10	2.24	·		
I4	10.34	8.12	7.28	·	
I5	3.00	2.00	4.12	8.60	·

Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Les deux objets les plus proches sont I2 et I5.

	I1	I2	I3	I4	I5
I1	·				
I2	3.61	·			
I3	5.10	2.24	·		
I4	10.34	8.12	7.28	·	
I5	3.00	2.00	4.12	8.60	·

Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Nous regroupons les objets I2 et I5 dans un **nœud** N1 et nous recalculons les similitudes avec ce nouveau groupe à l'aide du critère d'agglomération.

	I1	I3	I4	N1
I1	·			
I3	5.10	·		
I4	10.34	7.28	·	
N1	3.00	2.24	8.12	·

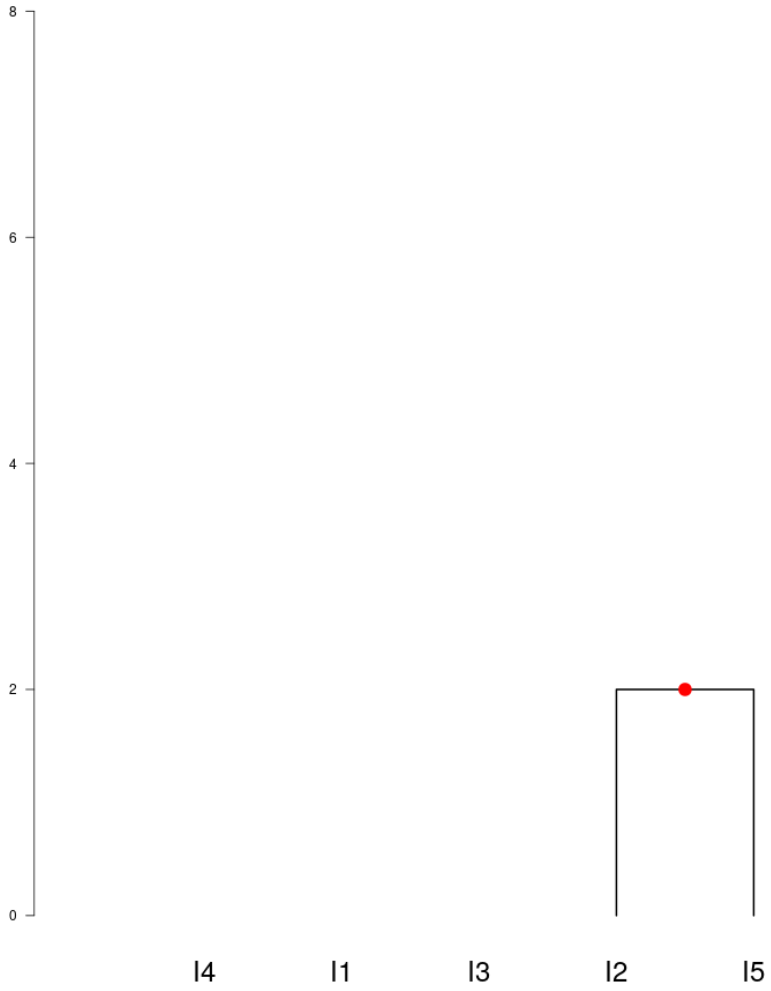


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La plus faible similitude est maintenant celle entre I3 et N1.

	I1	I3	I4	N1
I1	·			
I3	5.10	·		
I4	10.34	7.28	·	
N1	3.00	2.24	8.12	·

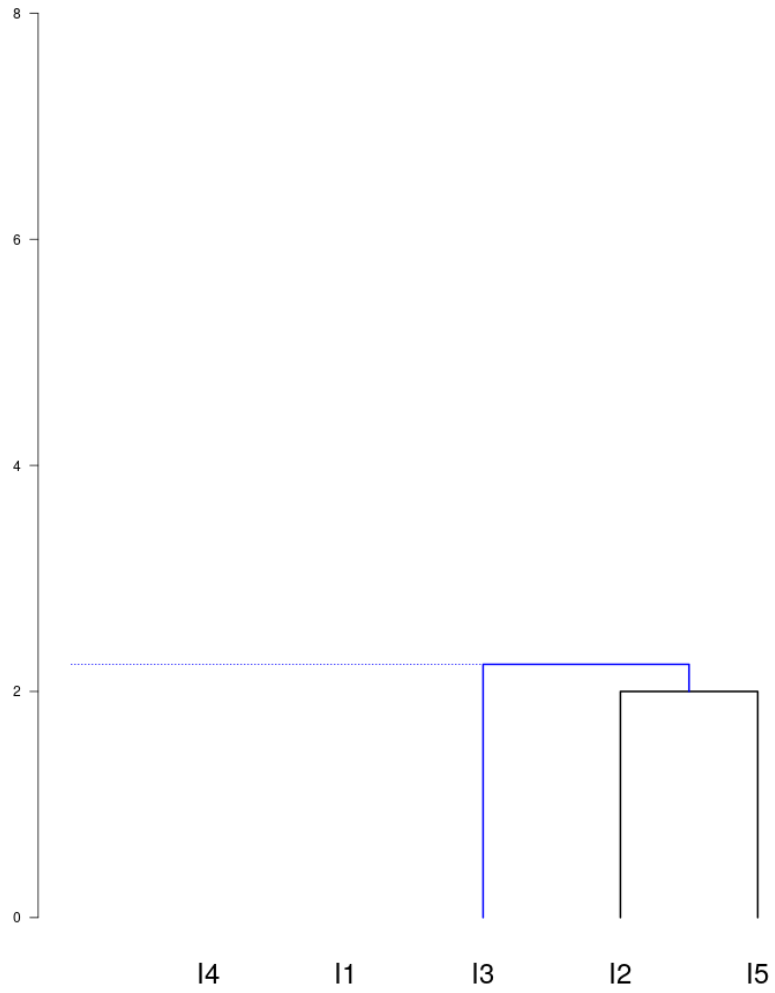


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Le groupe singleton {I3} et le groupe N1 sont regroupés dans un nœud N2 et les similitudes sont mises à jour.

	I1	I4	N2
I1	·		
I4	10.34	·	
N2	3.00	7.28	·

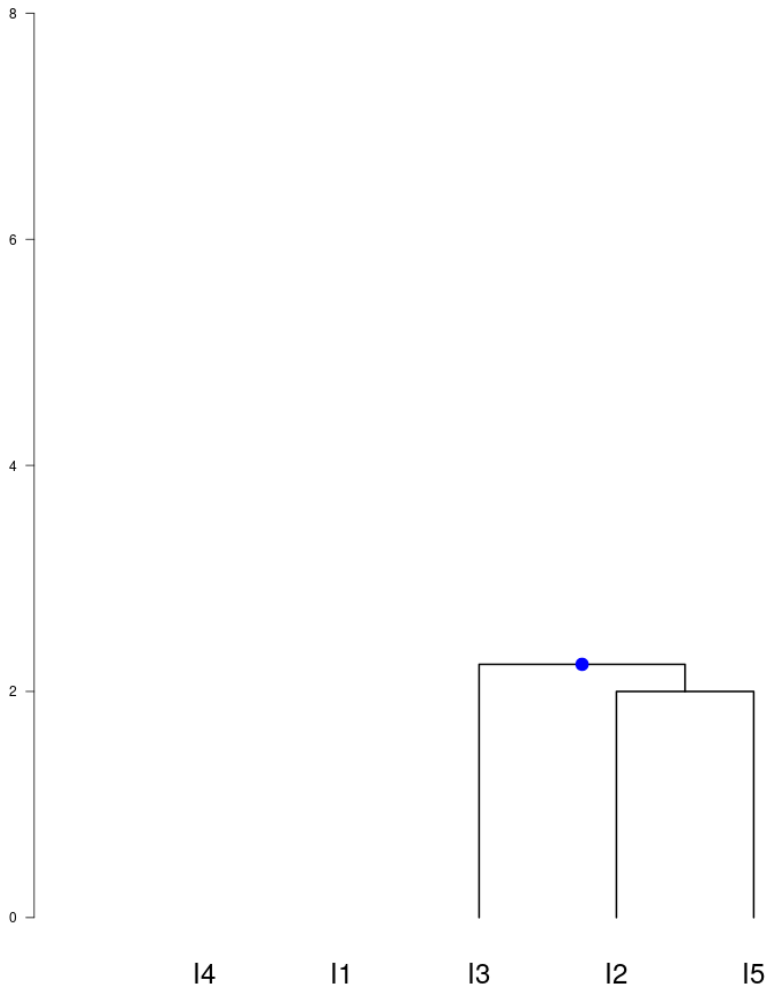


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La plus petite similitude est observée pour I1 et N2.

	I1	I4	N2
I1	·		
I4	10.34	·	
N2	3.00	7.28	·

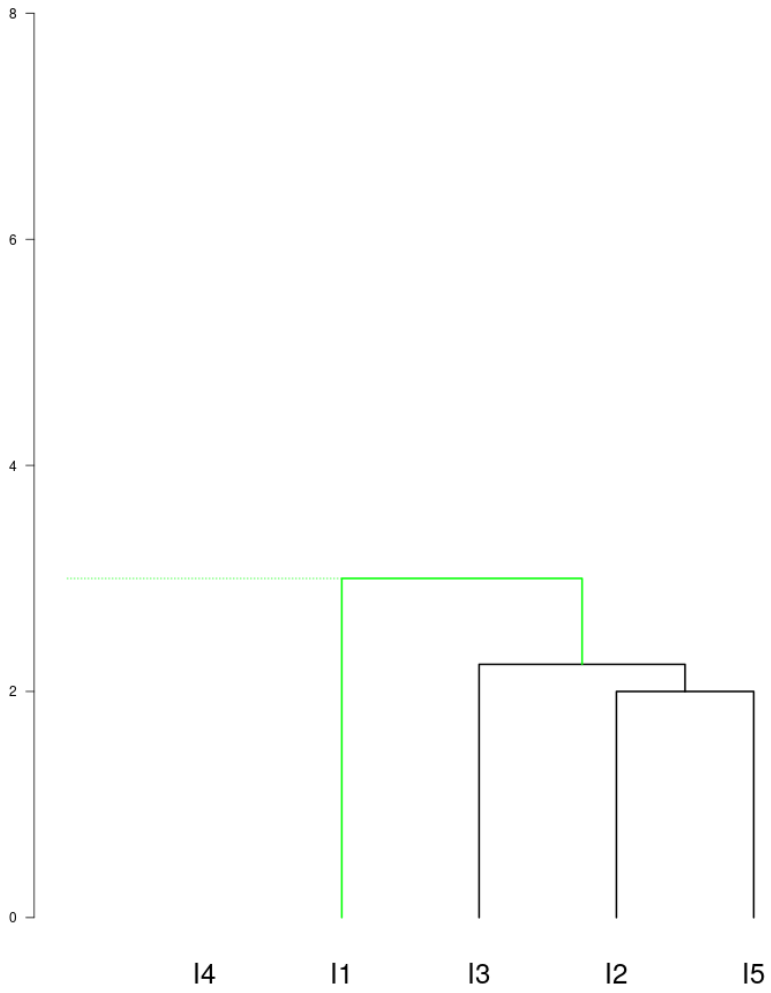


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La dernière similitude est celle entre l'objet I4 et le groupe N3 formé par les objets I1, I2, I3 et I5.

	I4	N3
I4	.	
N3	7.28	.

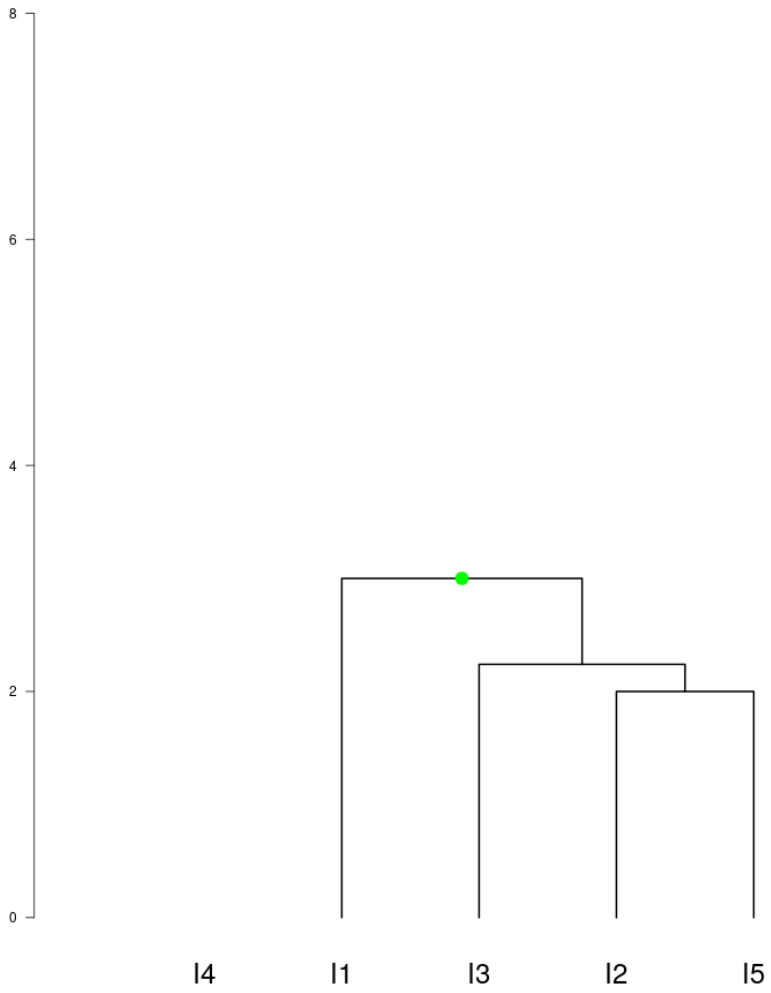


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Pour terminer l'algorithme, il suffit de regrouper tous les objets dans un dernier nœud N4 à la hauteur 7.28.

	I4	N3
I4	·	
N3	7.28	·

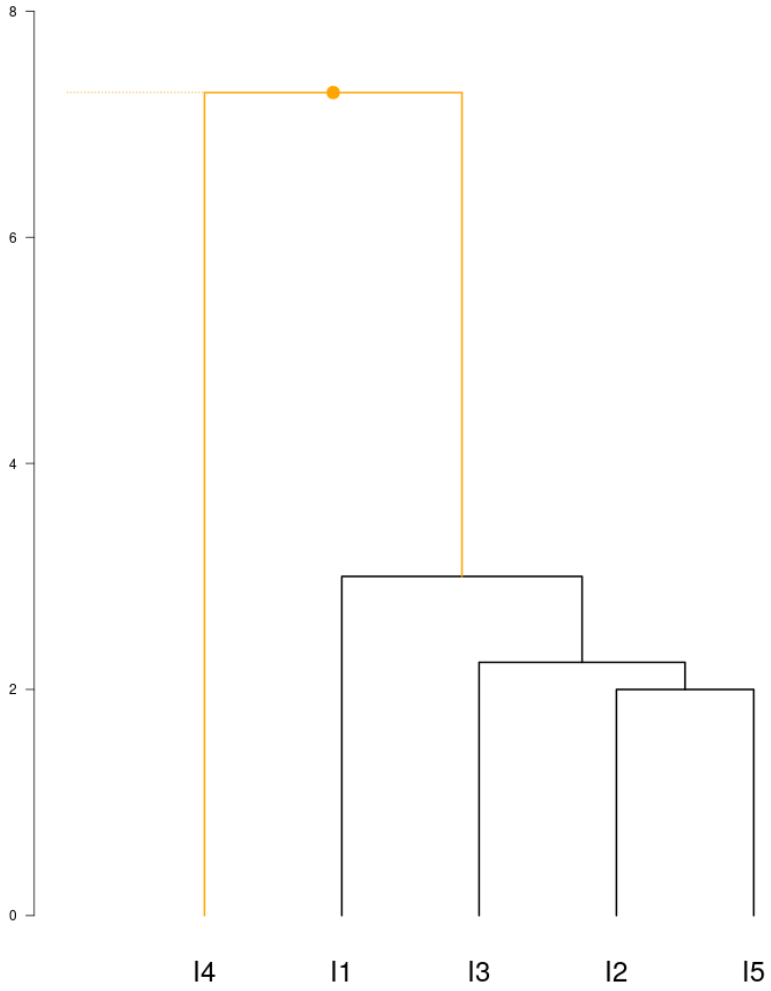
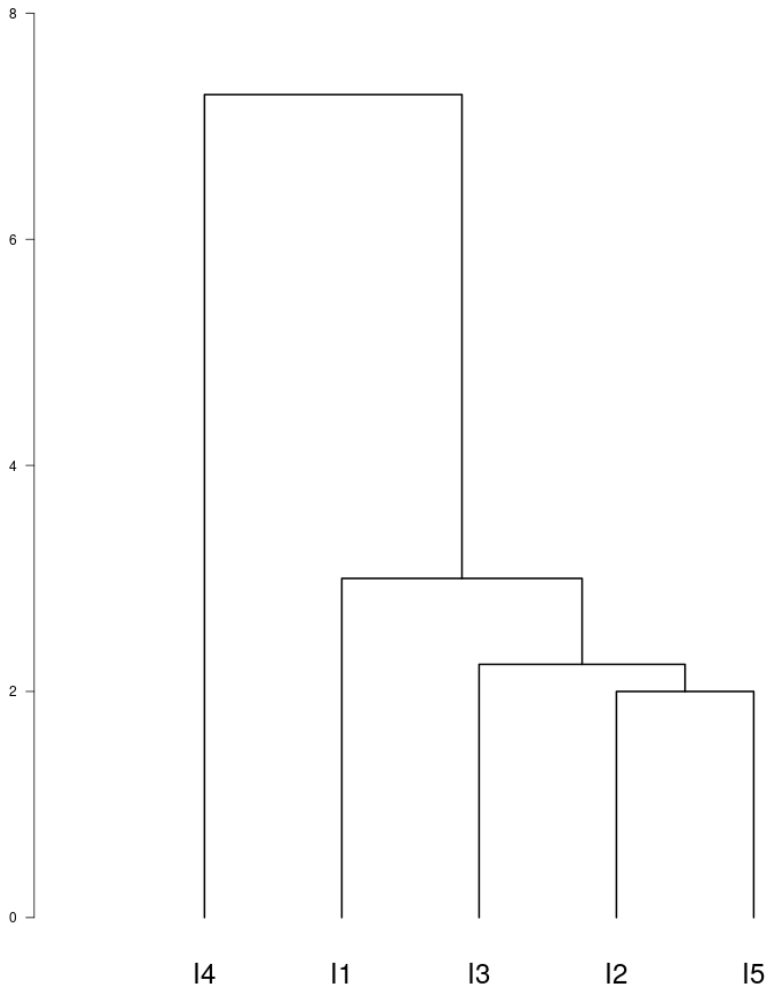


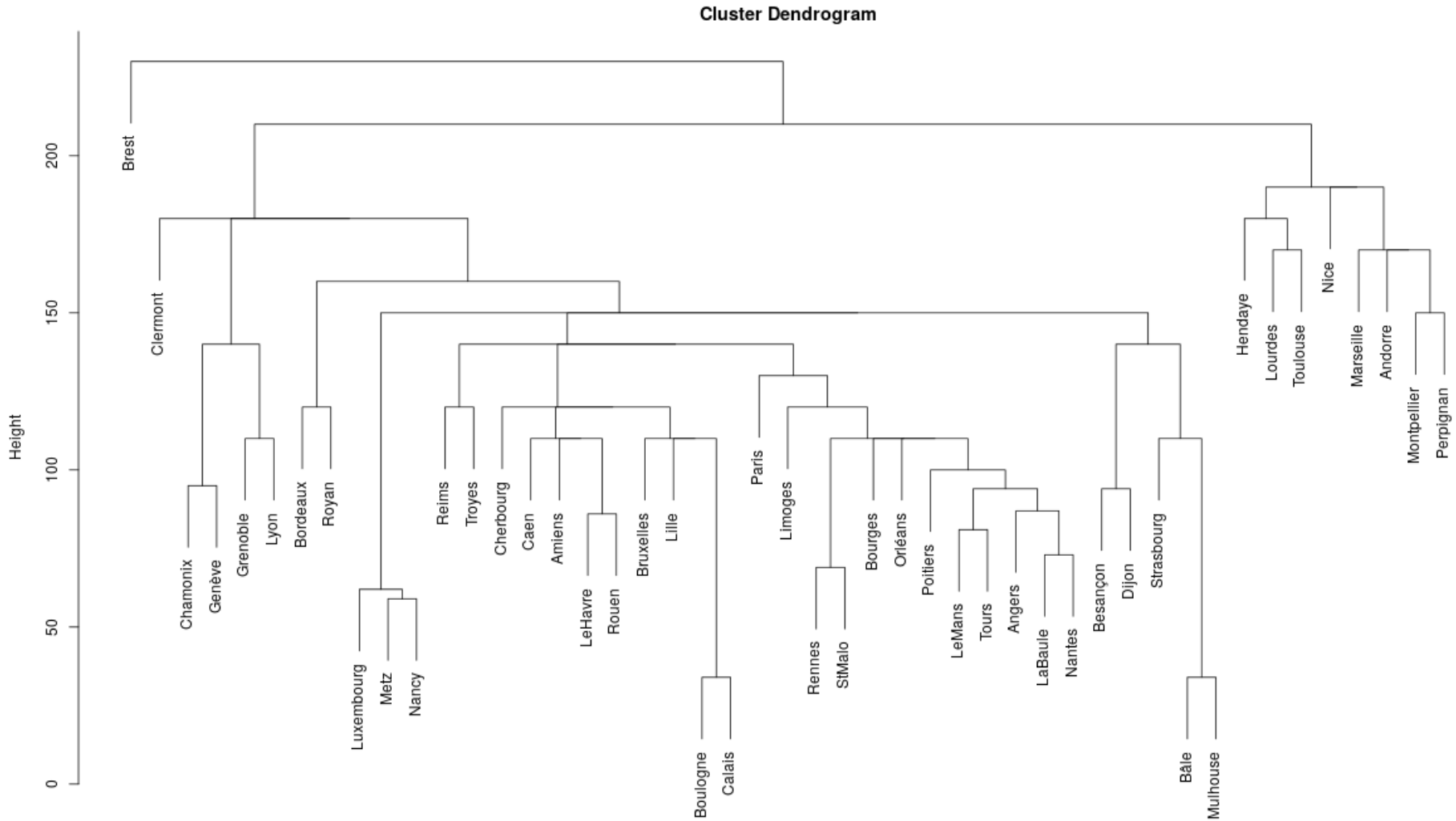
Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Le **dendrogramme** ainsi obtenu rend compte des différentes étapes de regroupement ainsi que des hauteurs des **sauts de similitudes** effectués.

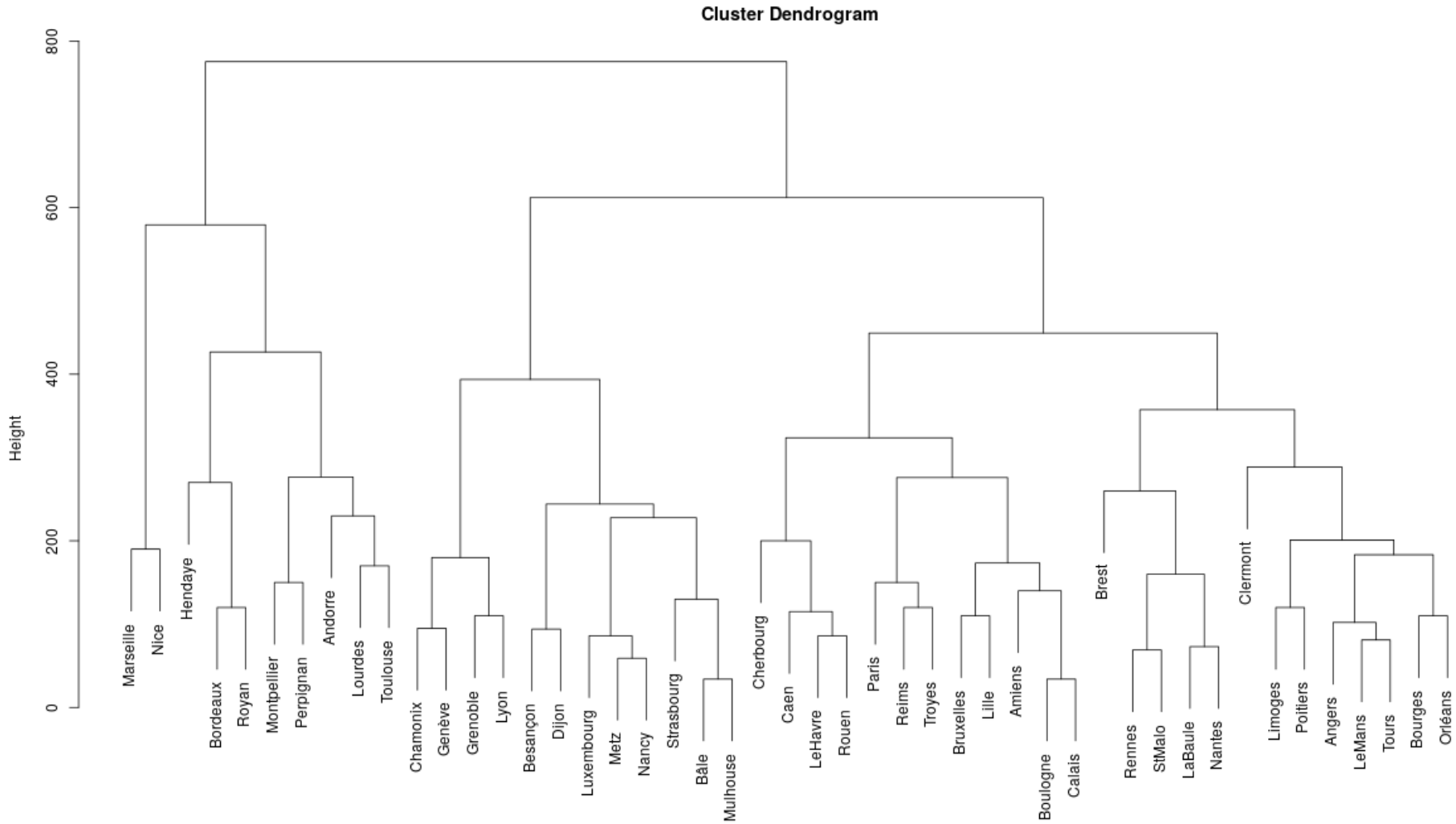


Exemple géographique (distances IGN)



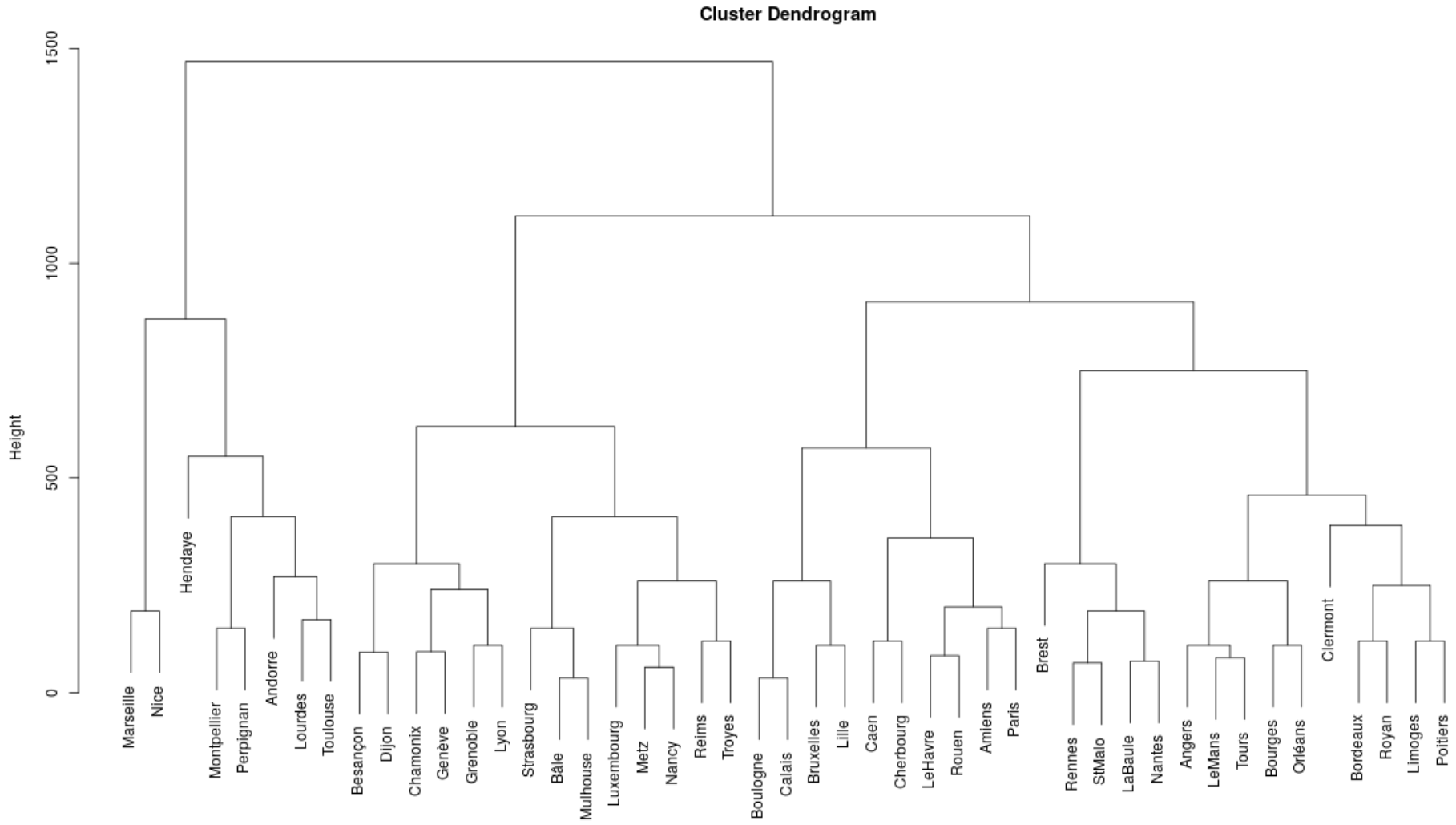
Critère d'agglomération minimum (*single linkage*)

Exemple géographique (distances IGN)



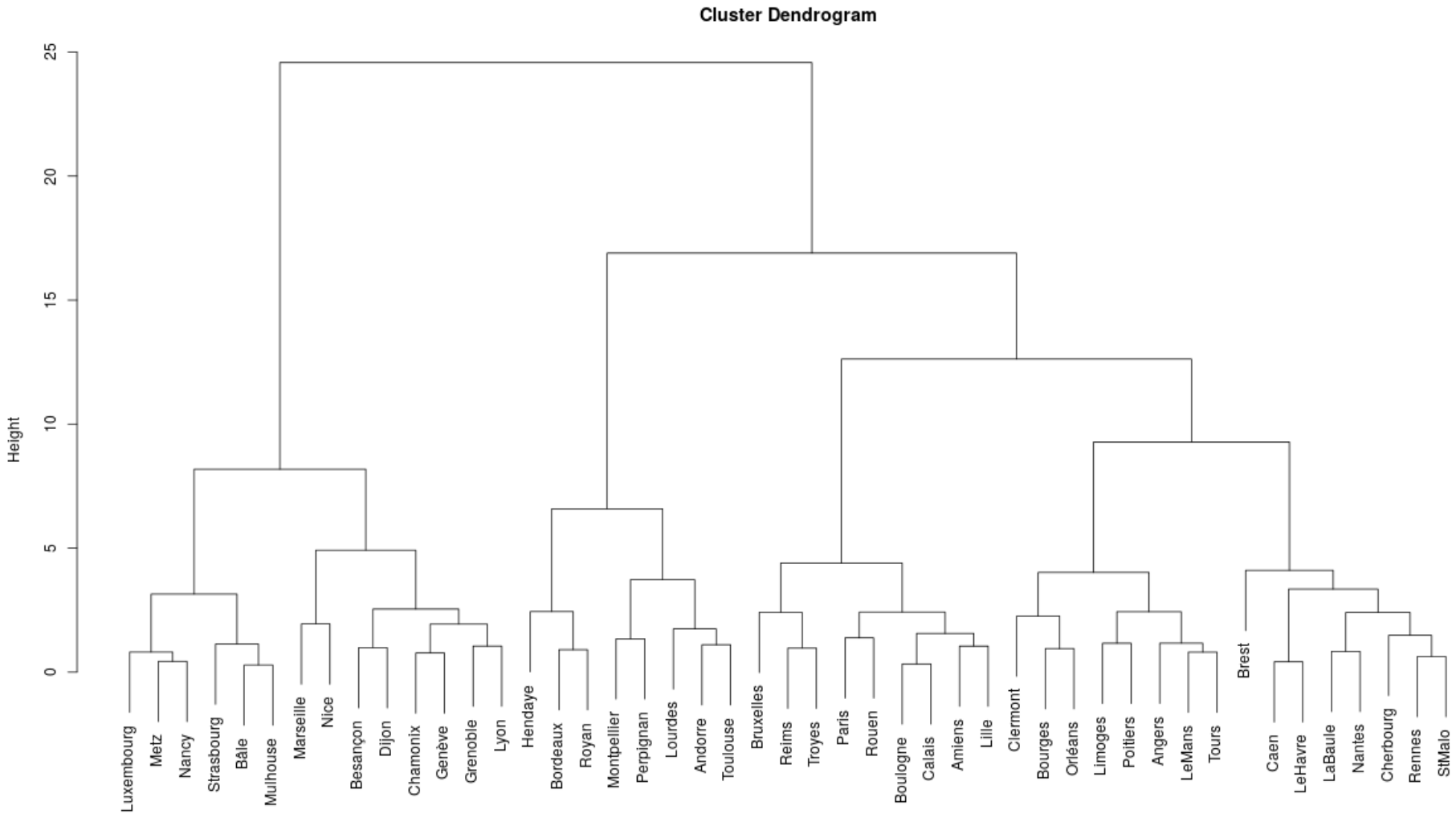
Critère d'agglomération moyen (*average linkage*)

Exemple géographique (distances IGN)



Critère d'agglomération maximum (*complete linkage*)

Exemple géographique (distances GPS)



Critère d'agglomération de Ward

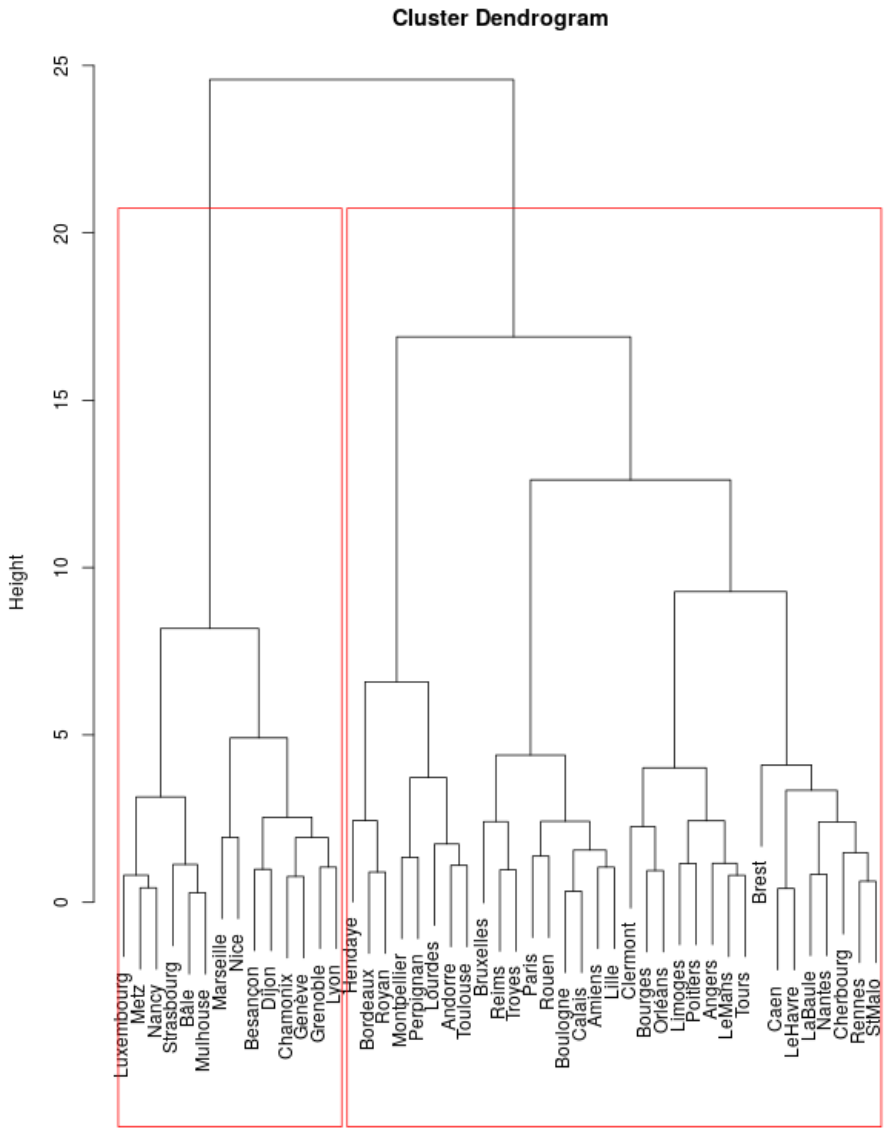
Utilisation d'un dendrogramme

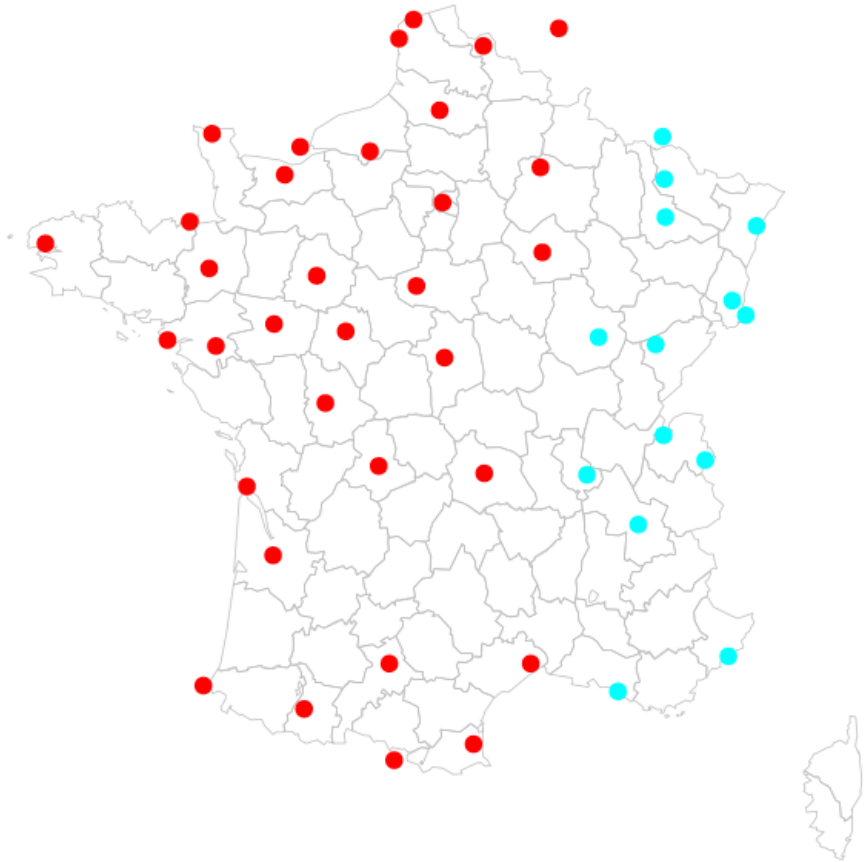
Afin de déduire une classification des objets initiaux à partir d'un dendrogramme, il faut se donner une **hauteur de coupe**. Les groupes objets donnés par les « branches » obtenues forment les classes.

Plus le dendrogramme est coupé haut, plus la classification est grossière, *i.e.* peu de classes voire même une seule contenant tous les objets.

Une hauteur de coupe est pertinente si elle se trouve entre deux nœuds séparés par une hauteur relativement « grande ». Avec le critère de Ward, cela s'interprète comme une **part d'inertie inter-groupe expliquée** similaire à ce que nous avons manipulé dans le cadre de l'ACP.

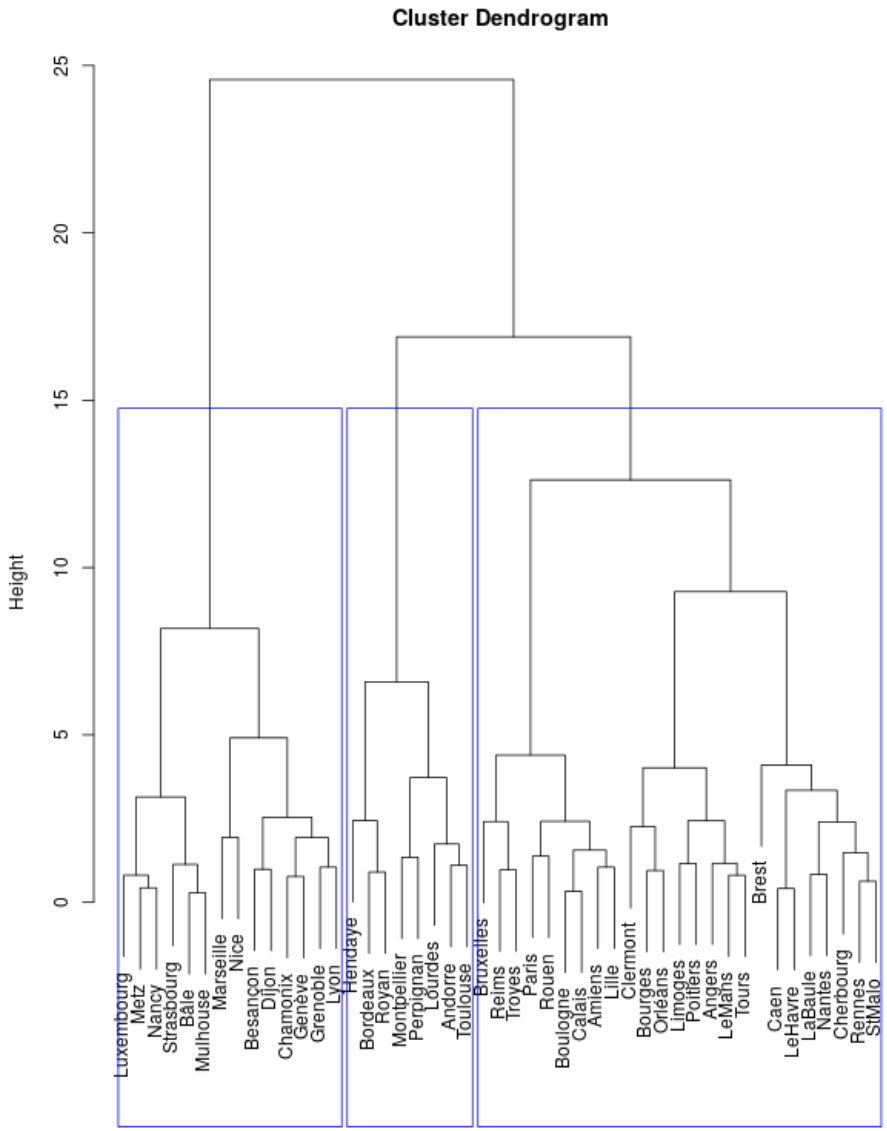
Exemple géographique (distances GPS, critère de Ward)

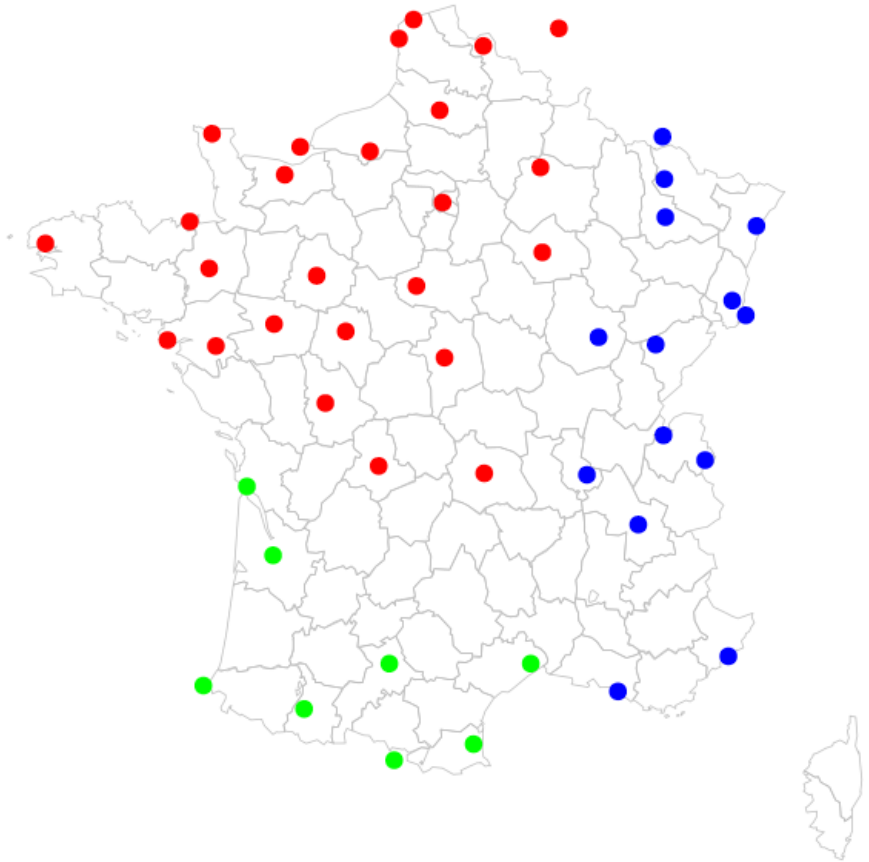




2 classes

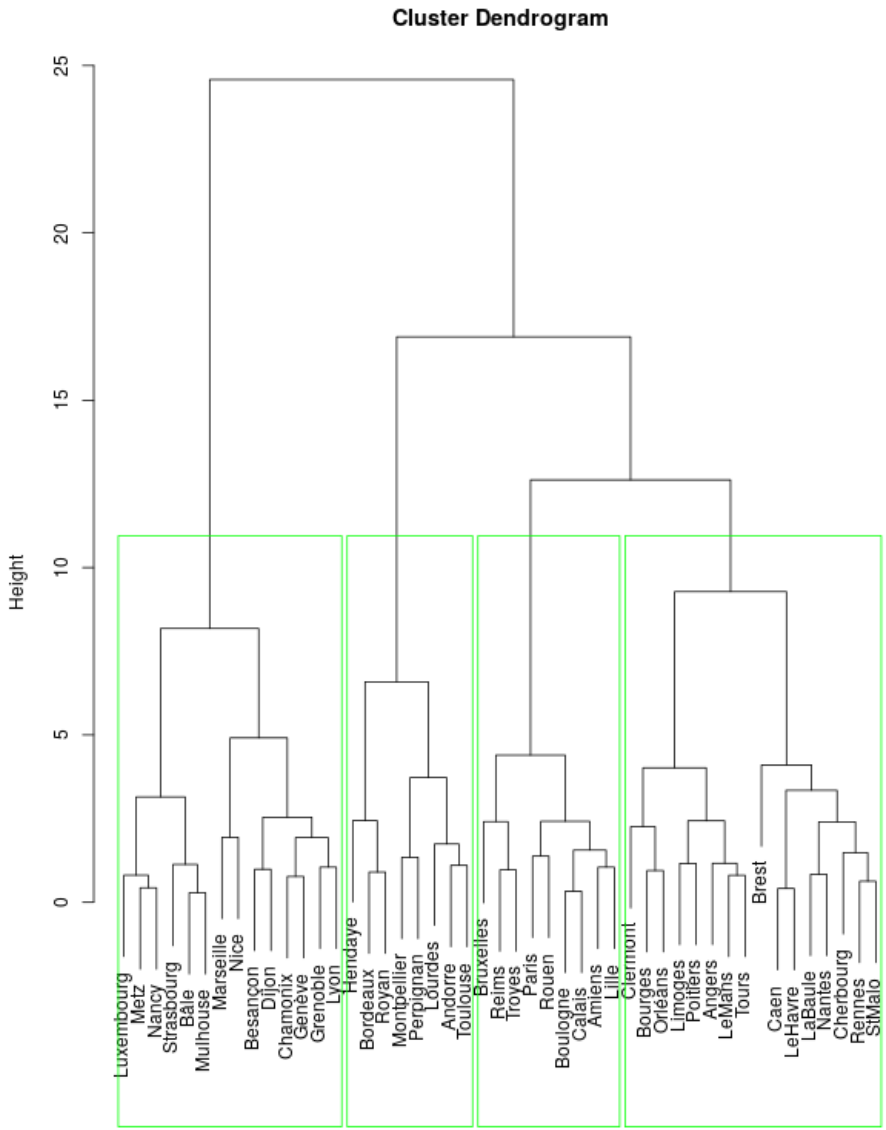
Exemple géographique (distances GPS, critère de Ward)

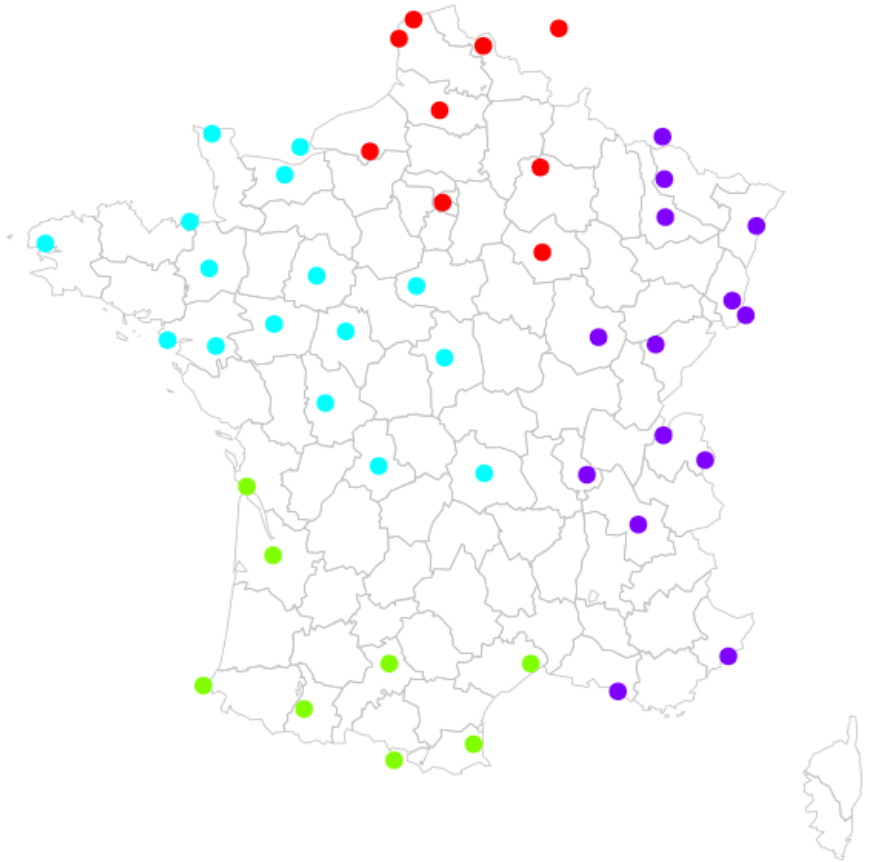




3 classes

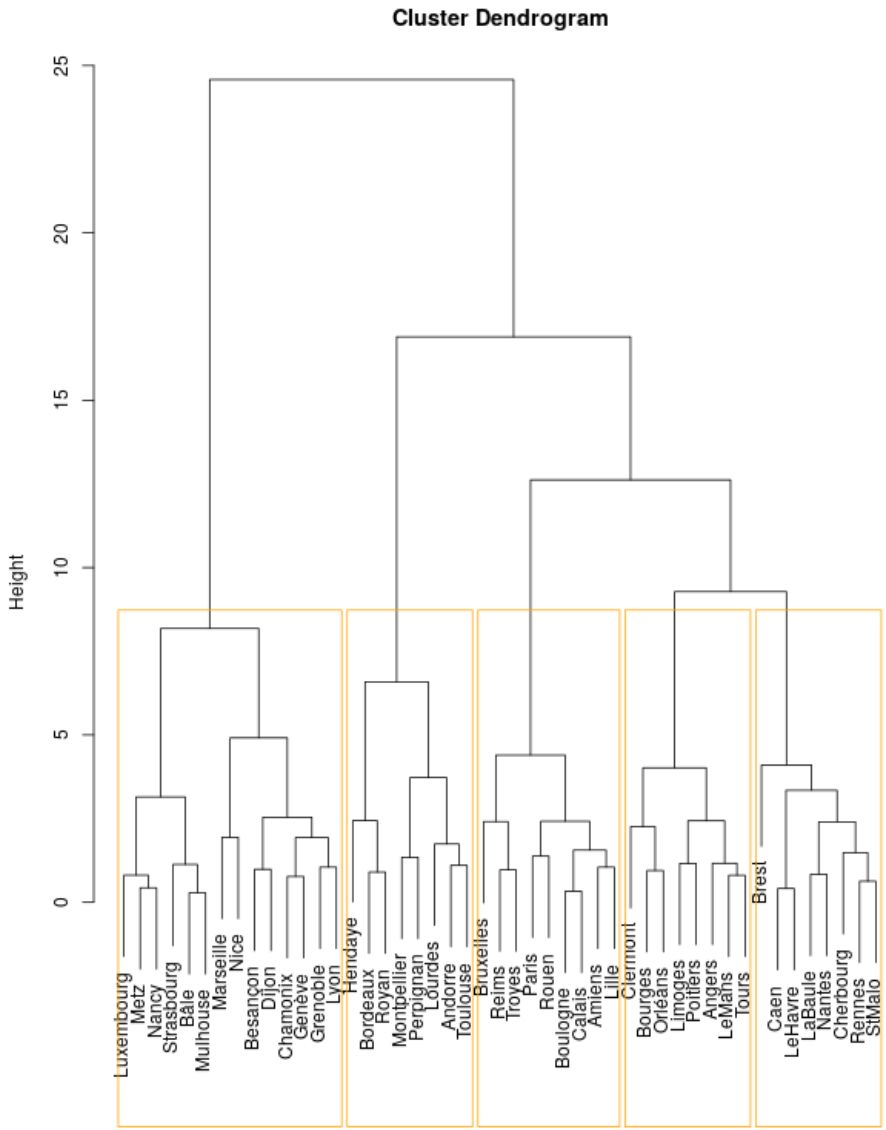
Exemple géographique (distances GPS, critère de Ward)

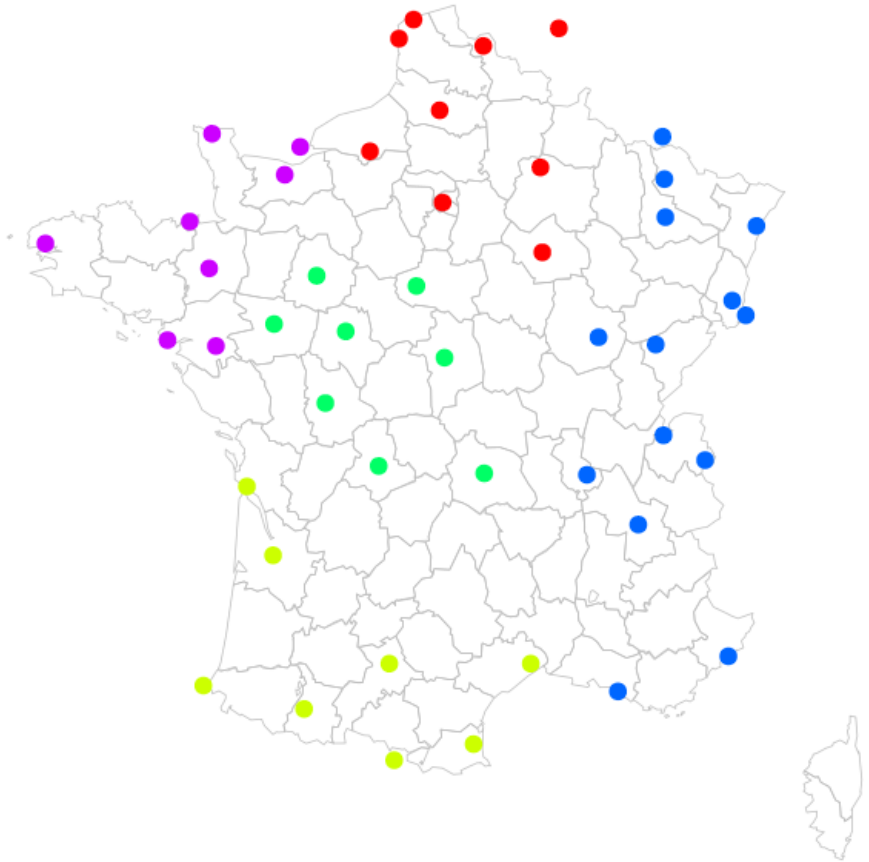




4 classes

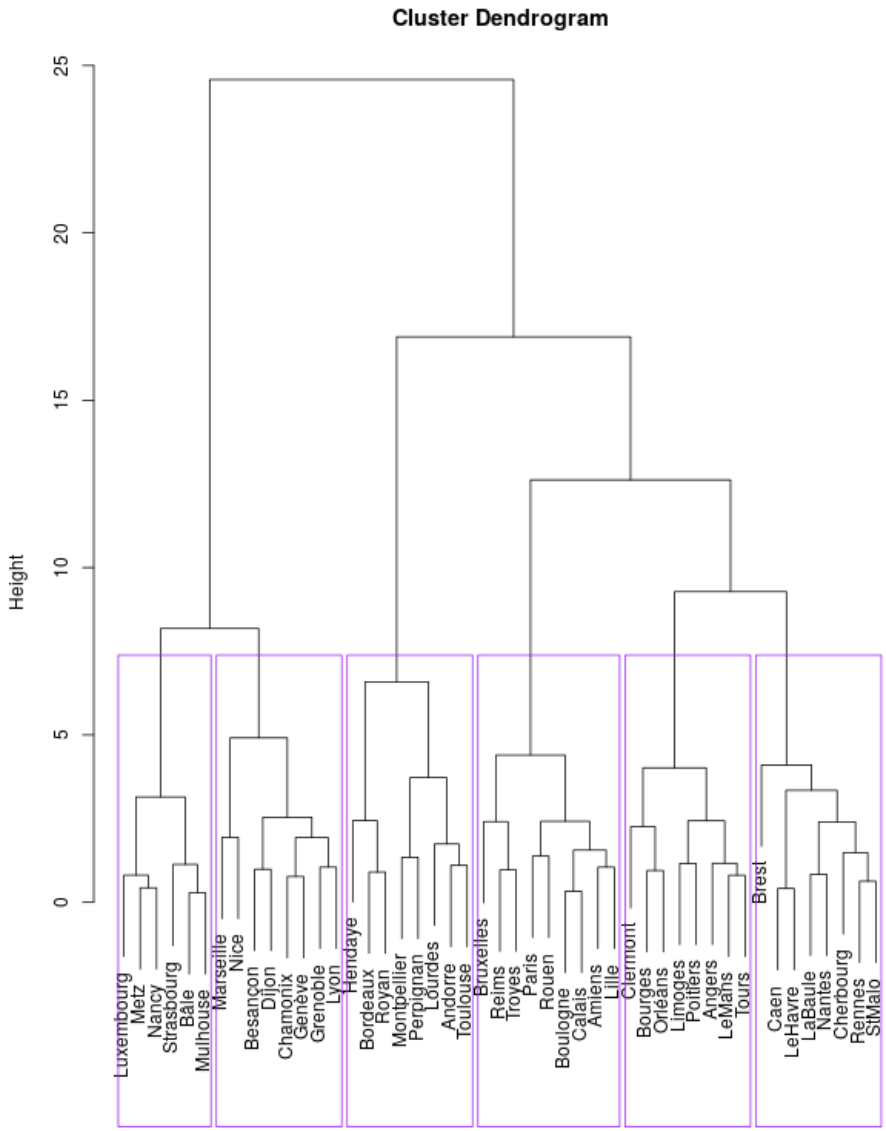
Exemple géographique (distances GPS, critère de Ward)

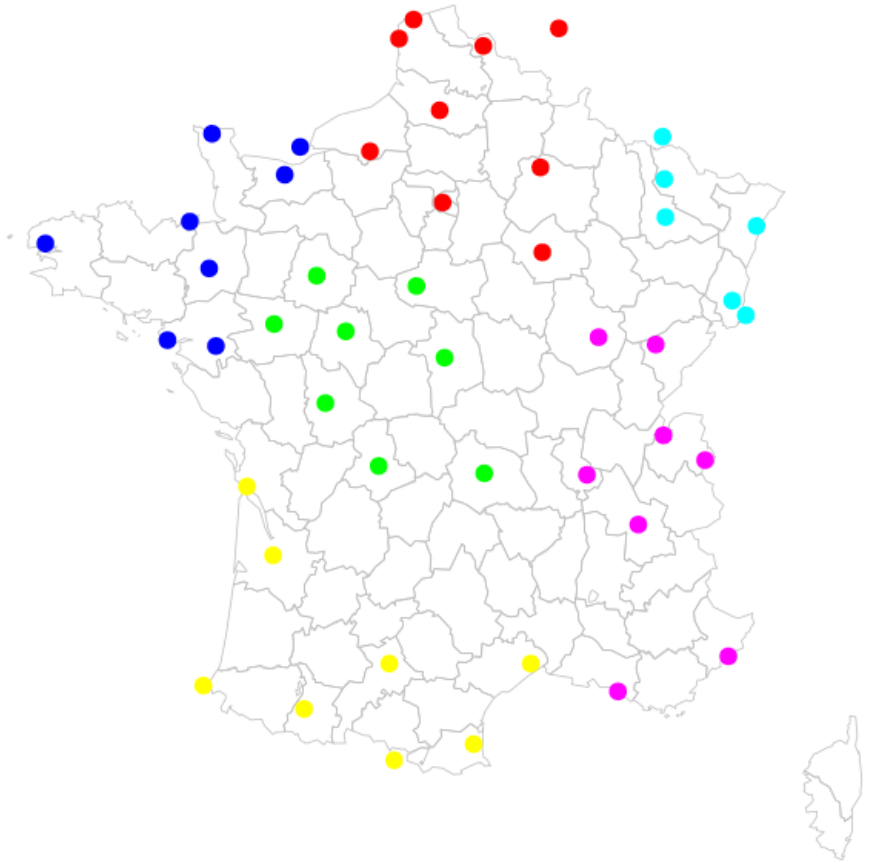




5 classes

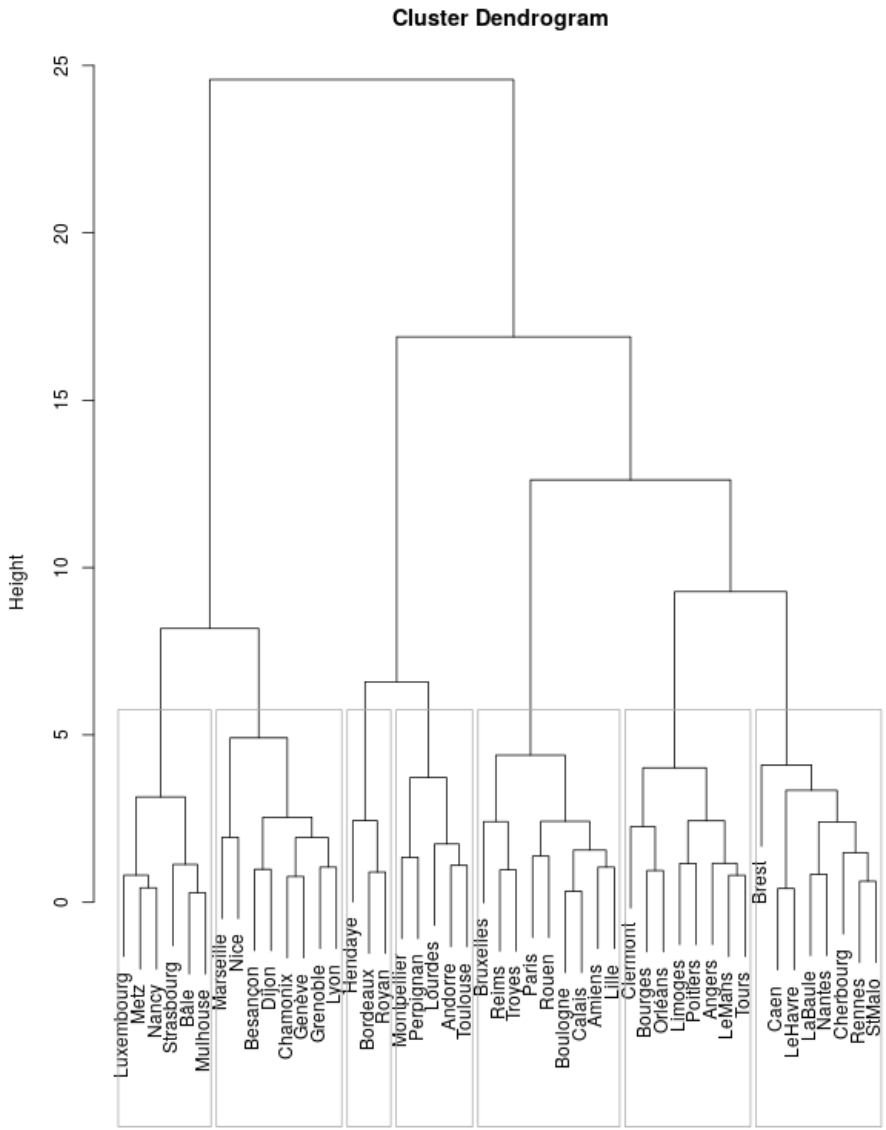
Exemple géographique (distances GPS, critère de Ward)

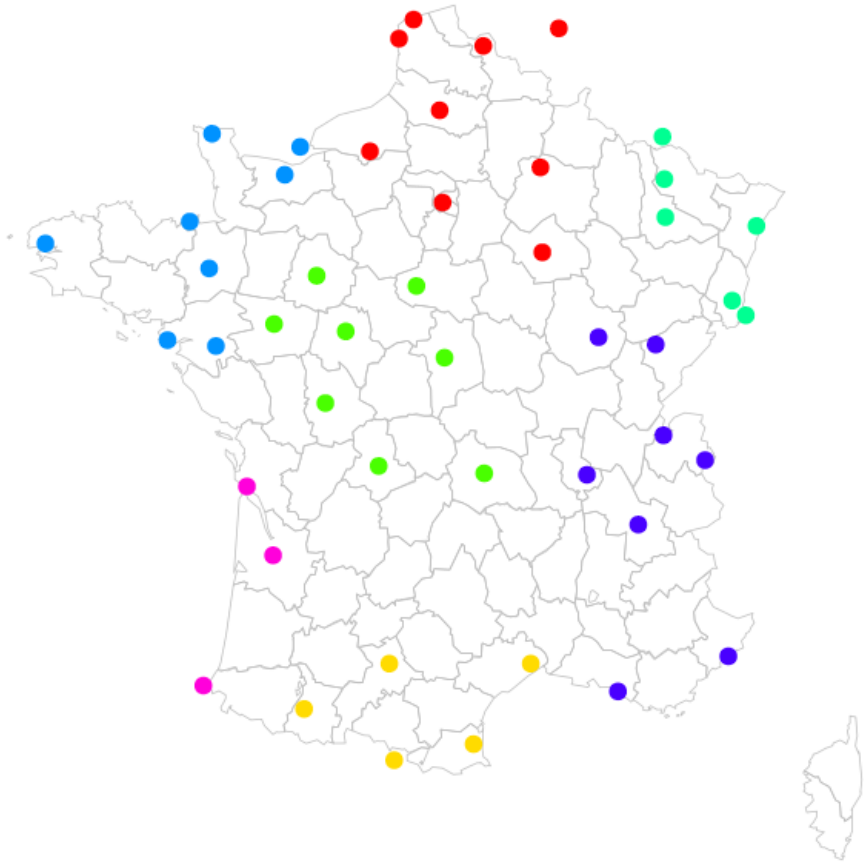




6 classes

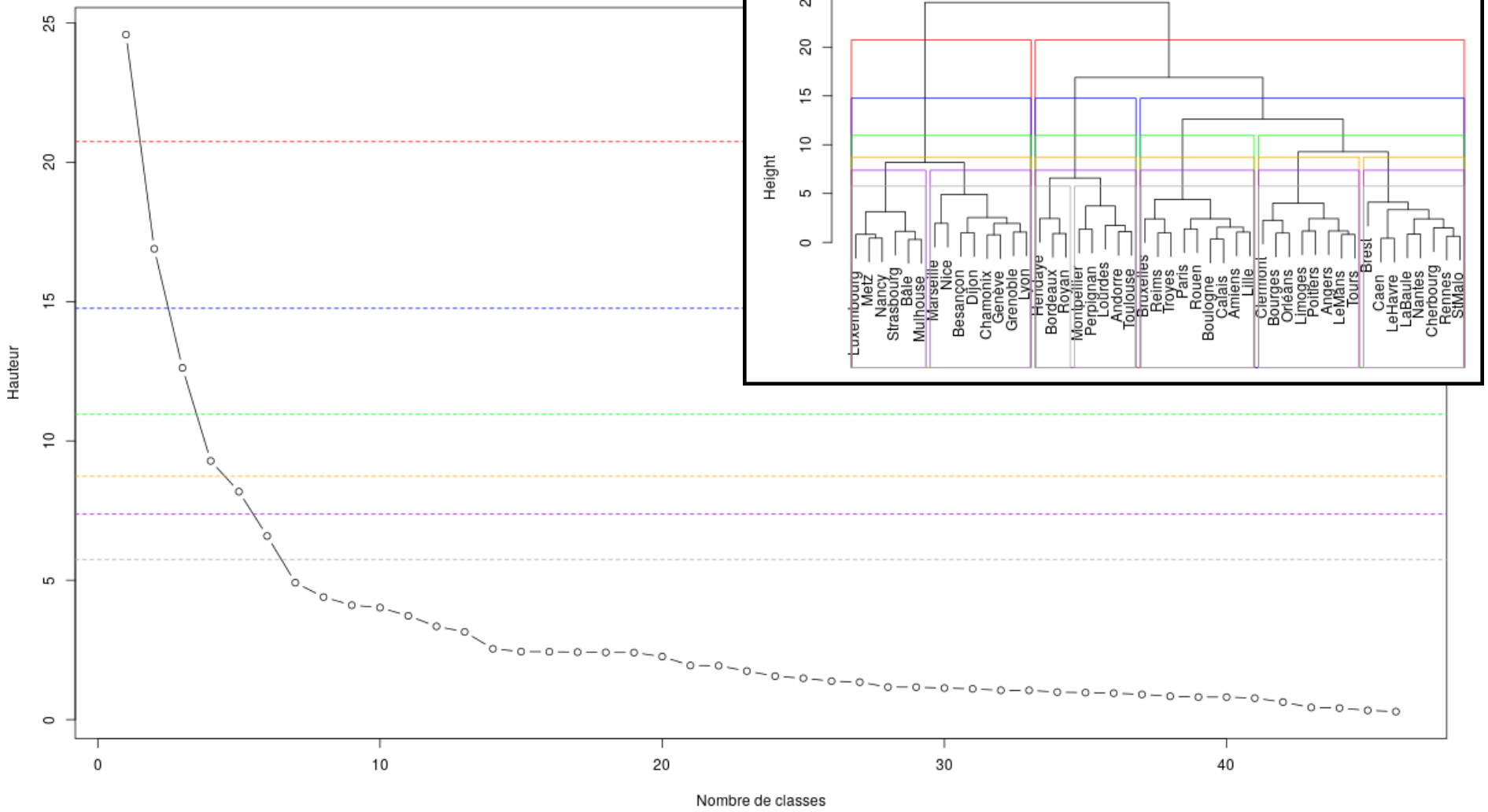
Exemple géographique (distances GPS, critère de Ward)





7 classes

Exemple géographique (distances GPS, critère de Ward)



Centres mobiles versus CAH

Centres mobiles

Avantages :

- Classification **robuste** car minimum d'un critère
- Algorithme simple à mettre en œuvre et rapide

Inconvénients :

- Demande de connaître le nombre K de classes

CAH

Avantages :

- Ne nécessite pas de connaître le nombre K de classes

Inconvénients :

- Classification **sensible aux données** car la topologie du dendrogramme dépend fortement des premiers regroupements
- Algorithme lourd en temps de calcul

Stabilisation

Pour tirer parti des avantages des deux méthodes, il est possible de les enchaîner en utilisant le résultat de l'une comme initialisation de la suivante.

- **Étape optionnelle :** si le nombre n d'objets à classer est trop important pour envisager une CAH, nous pouvons appliquer les centres mobiles avec un nombre $K_0 \ll n$ de classes **grand** et nous restreindre à ces K_0 groupes comme des objets élémentaires dans les étapes suivantes.
- **Étape CAH :** le dendrogramme permet de **déterminer un nombre de classes** K_1 « pertinent » d'après les données (critère d'inertie, ...) et de fournir une première classification **peu robuste**.
- **Étape Centres Mobiles :** pour **stabiliser** la classification obtenue à l'étape précédente, nous utilisons une agrégation autour de K_1 centres mobiles avec la classification de la CAH comme initialisation.