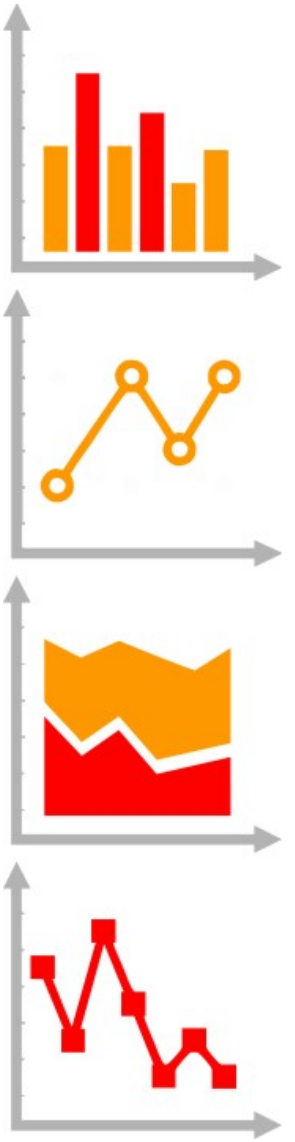
Identification des problèmes de différenciation géographique

Journées de méthodologie statistique, jeudi 14 juin 2018



Vianney Costemalle, *Département des méthodes statistiques, Division des méthodes et référentiels géographiques, Insee*

Insee

Mesurer pour comprendre

Rappel : Le secret statistique

- Ne pas diffuser de données révélant l'identité d'un individu ou des caractéristiques **sensibles**

- Protection des données : agréger les données avec un **seuil minimal** d'observations

- Données **géographiques** : les zones de diffusion doivent être suffisamment grandes (dépend de la densité des observations)

Secret statistique primaire et secondaire

- Secret statistique primaire : concerne les informations **directement** diffusées
  - Par exemple, pour les données carroyées de Filosofi, constitution de carrés comportant tous au moins 11 ménages fiscaux

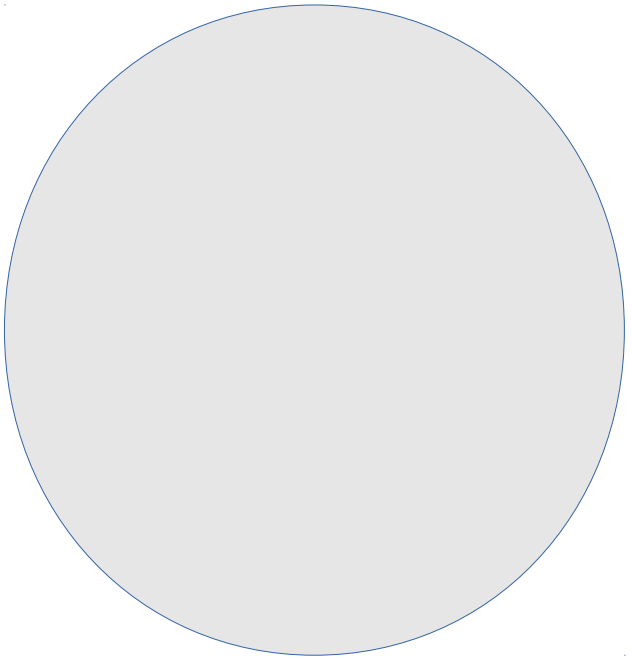
- Secret statistique secondaire : concerne les informations **indirectement** disponibles.
  - Par combinaison et recoupement des différentes données, on peut déduire de nouvelles informations

La différenciation géographique

- Information diffusée sur zones géographiques (communes, carreaux, unités urbaines, etc.)
- Déduire des informations sur de **plus petites** zones
- Variables de nature **additive**
  - Ex : somme des revenus, nombre de personnes âgées de plus de 65 ans, etc.
  - Contre ex : médiane des revenus, âge moyen, etc.

La différenciation en géographie

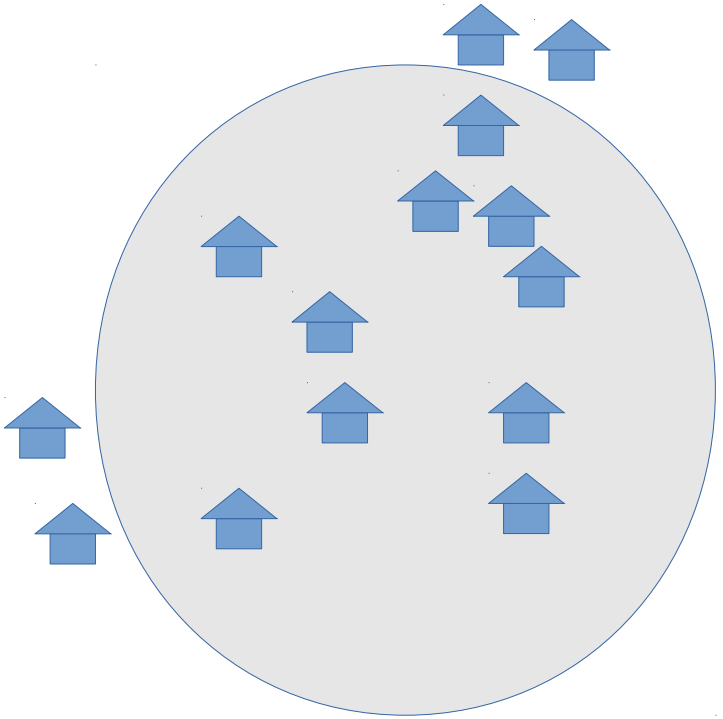
- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux



Commune

La différenciation en géographie

- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux

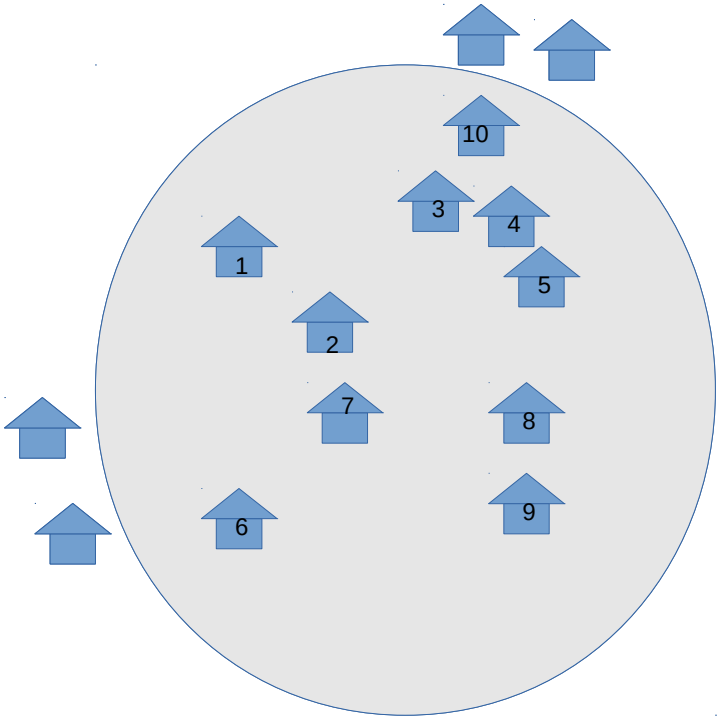


Commune : 10 ménages

Insee

La différenciation en géographie

- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux

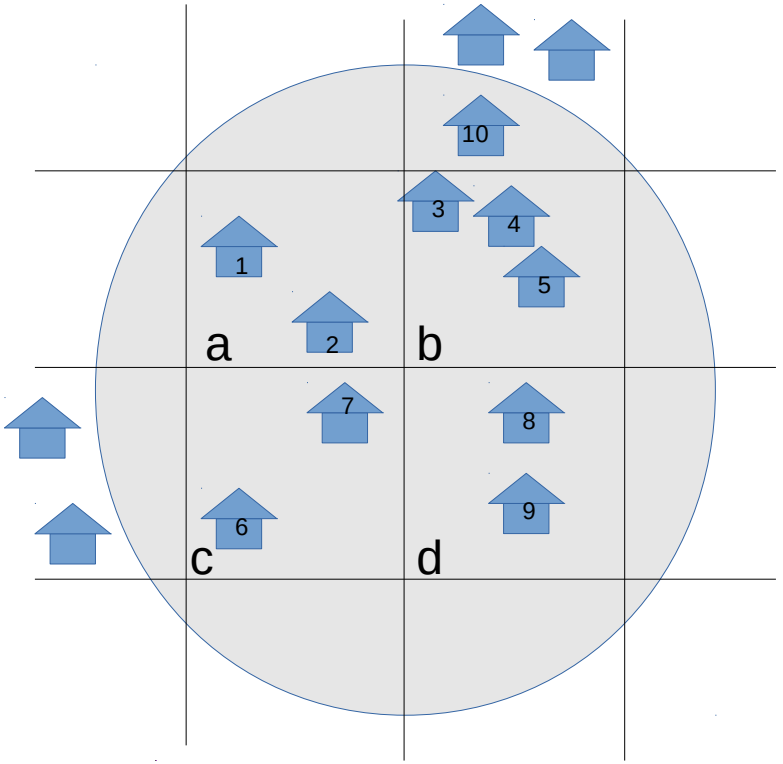


Commune : 10 ménages

zone	info diffusée
commune	Σ Xi (i=1 … 10)

La différenciation en géographie

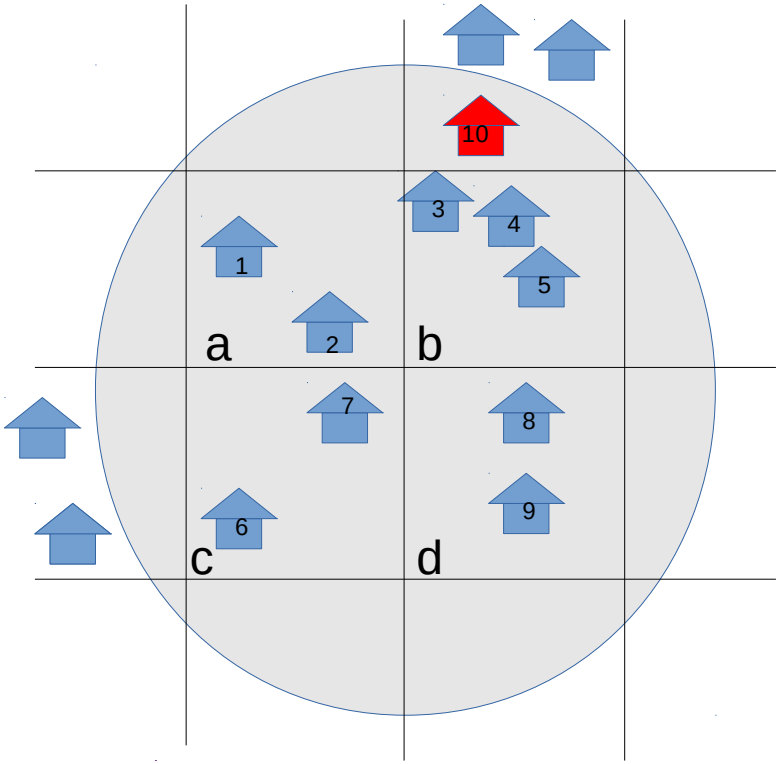
- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux



zone	info diffusée
commune	Σ Xi (i=1 … 10)
Carreau a	X1 + X2
Carreau b	X3 + X4 + X5
Carreau c	X6 + X7
Carreau d	X8 + X9

La différenciation en géographie

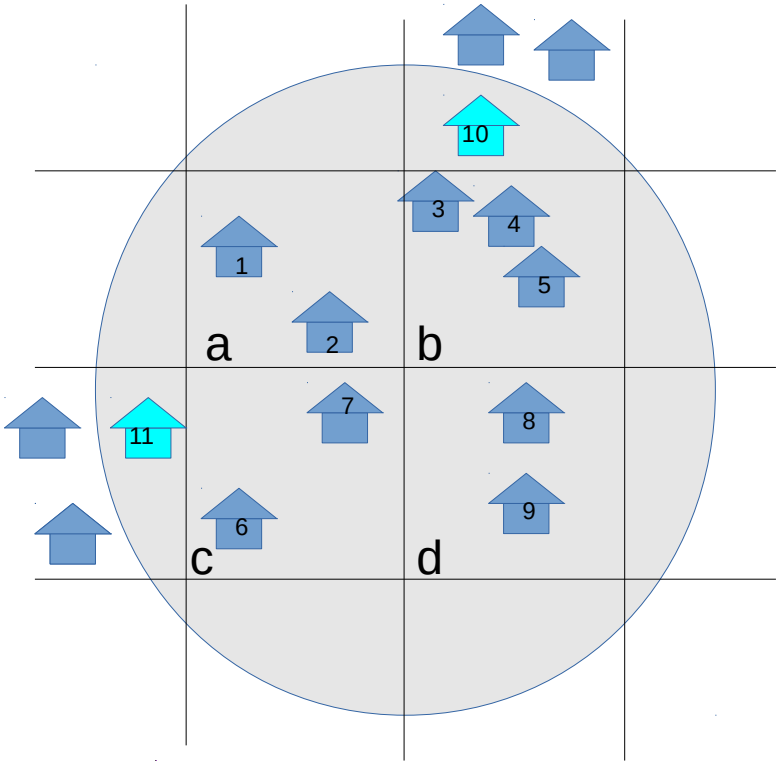
- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux



zone	info diffusée
commune	Σ Xi (i=1 … 10)
Carreau a	X1 + X2
Carreau b	X3 + X4 + X5
Carreau c	X6 + X7
Carreau d	X8 + X9
Commune – (a+b+c+d)	X10

La différenciation en géographie

- Informations diffusées sur plusieurs zonages différents
  - exemple : communes et grille de carreaux



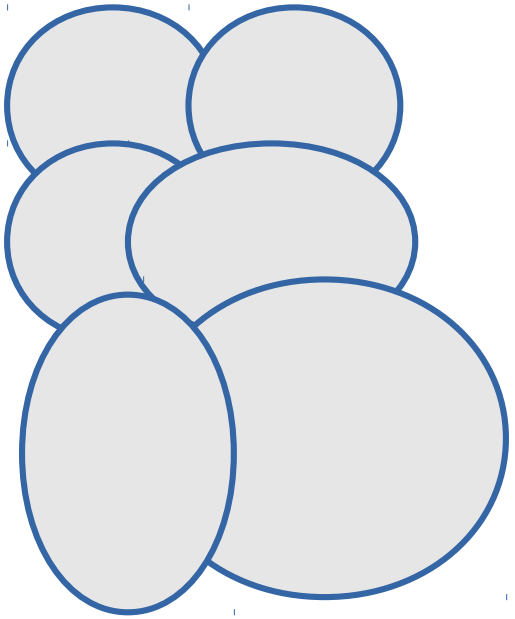
zone	info diffusée
commune	Σ Xi (i=1 … 10)
Carreau a	X1 + X2
Carreau b	X3 + X4 + X5
Carreau c	X6 + X7
Carreau d	X8 + X9
Commune – (a+b+c+d)	X10 + X11

La différenciation en géographie

- Zone englobante – zone englobée = zone frontière
  - Zone englobante/englobée :
    - 1 commune ou plusieurs communes
    - 1 carreau ou plusieurs carreaux
  - Zone frontière
    - composée de **briques élémentaire** que sont les intersections carreaux-communes
- Toutes les intersections carreaux-communes sous le seuil, ne sont pas forcément à risque
  - l'information locale n'est pas suffisante : il faut une information **globale**

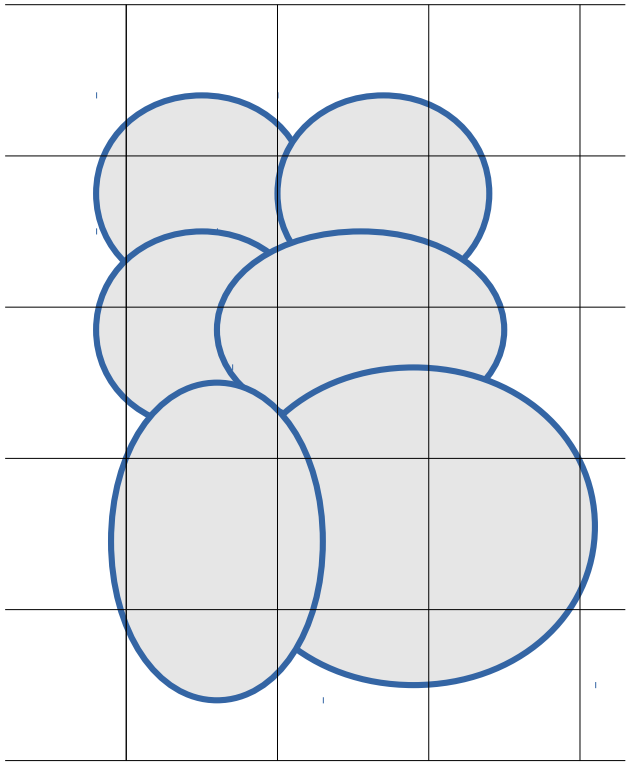
Différenciation « interne » et « externe »

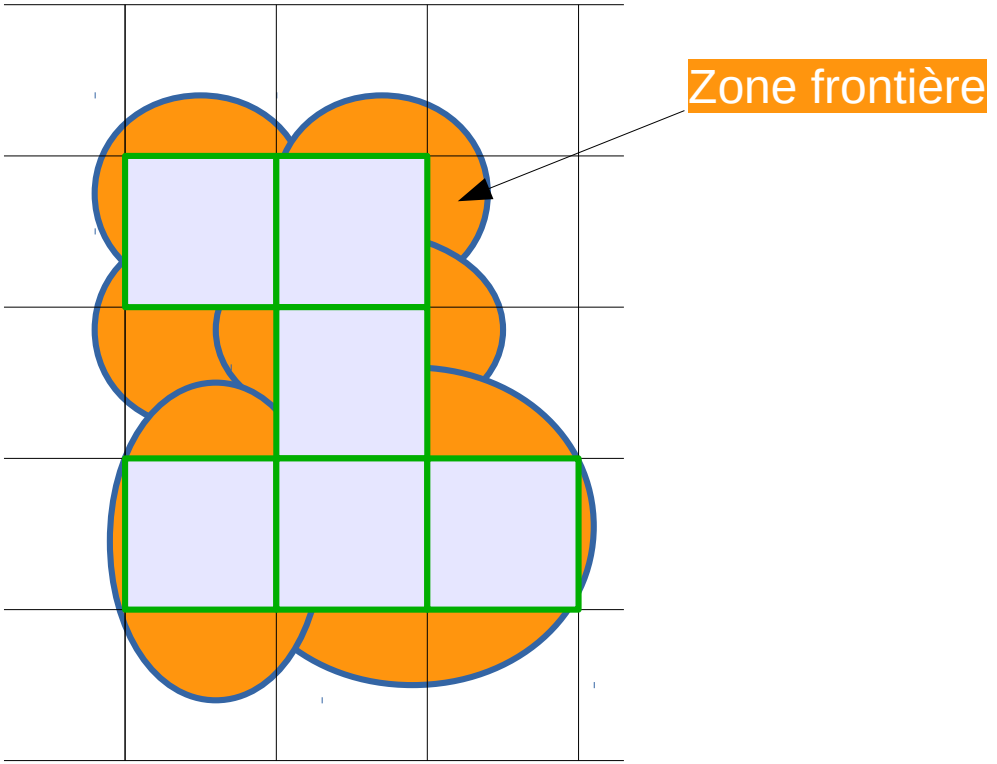
- On se donne un groupe de communes



Différenciation « interne » et « externe »

- On se donne un groupe de communes



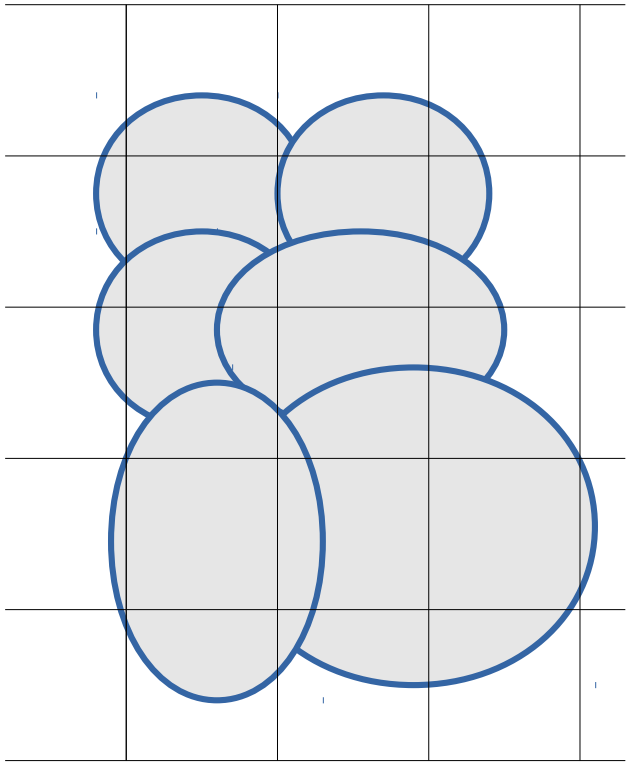


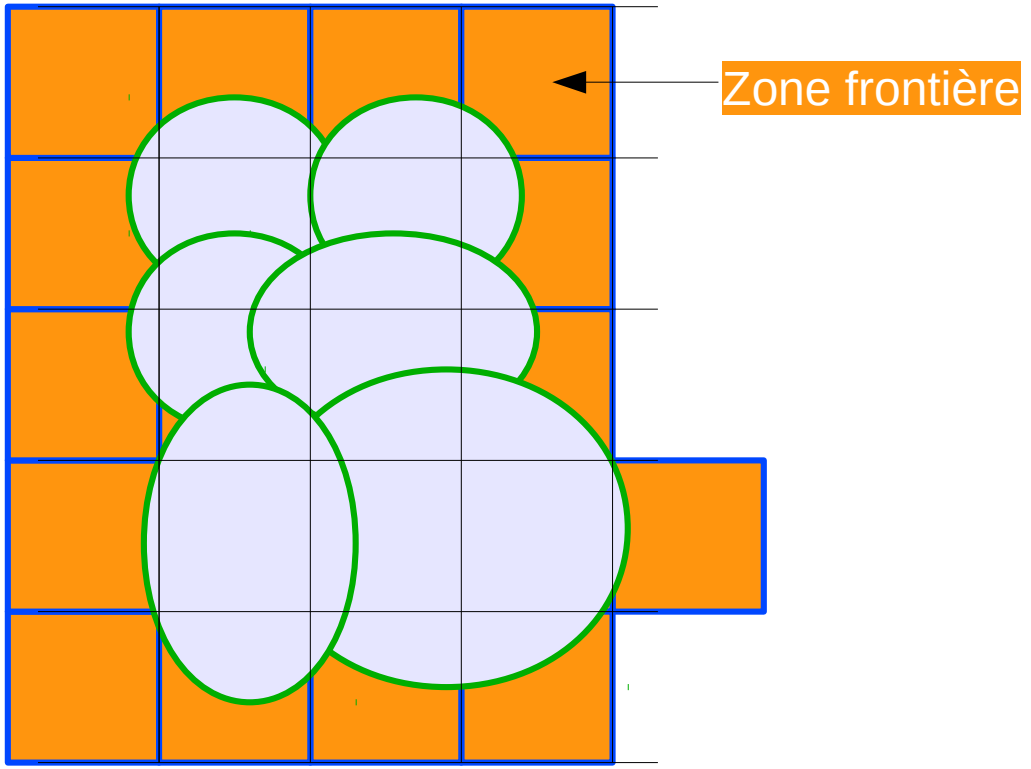
Différenciation « interne »

Insee

Différenciation « interne » et « externe »

- On se donne un groupe de communes





Différenciation « externe »

De très nombreuses combinaisons

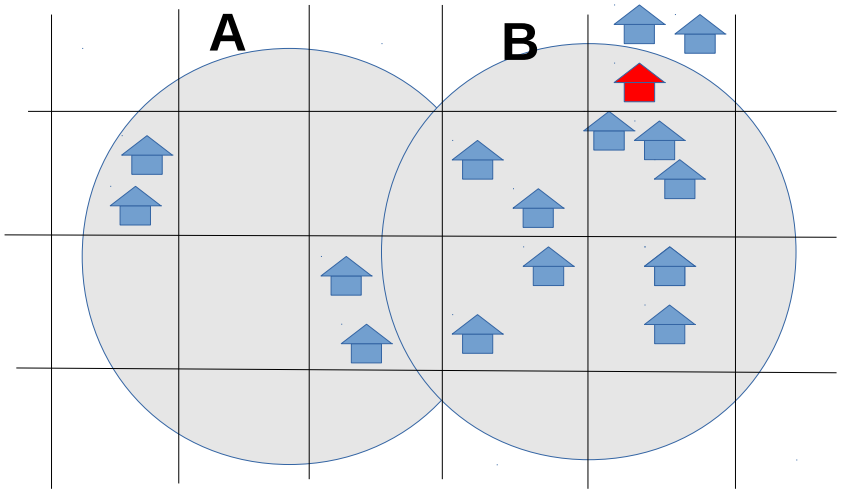
- Environ 36 000 communes en France
  - 2^36000 groupes de communes possibles
  - **Impossible** de tous les tester
  - Avoir une approche plus **sélective**

- Idées :
  - 1 – se restreindre au groupes de communes **contiguës**
  - 2 – regrouper certaines communes qui **ne peuvent être considérées séparément**

Communes contiguës

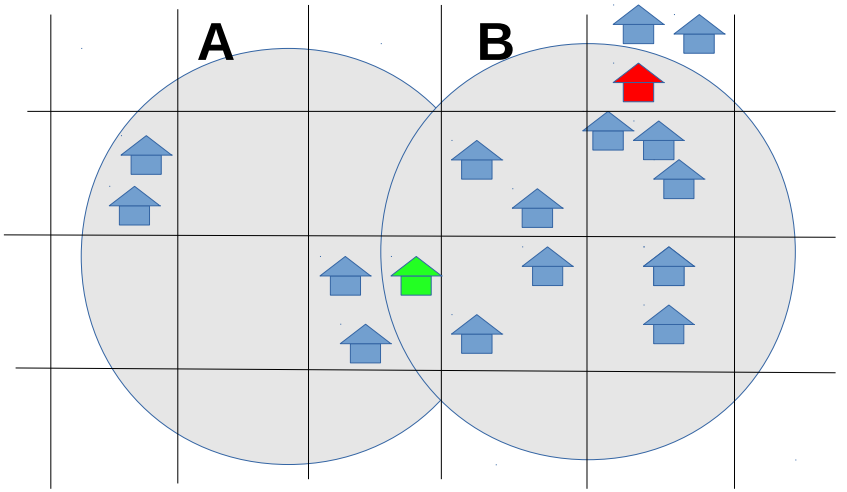
- 2 commune sont contiguës s'il existe au moins un carreau dont les observations sont réparties sur ces deux communes

Non contiguës



Pbm sur (A+B)
==> pbm sur A ou sur B

Contiguës



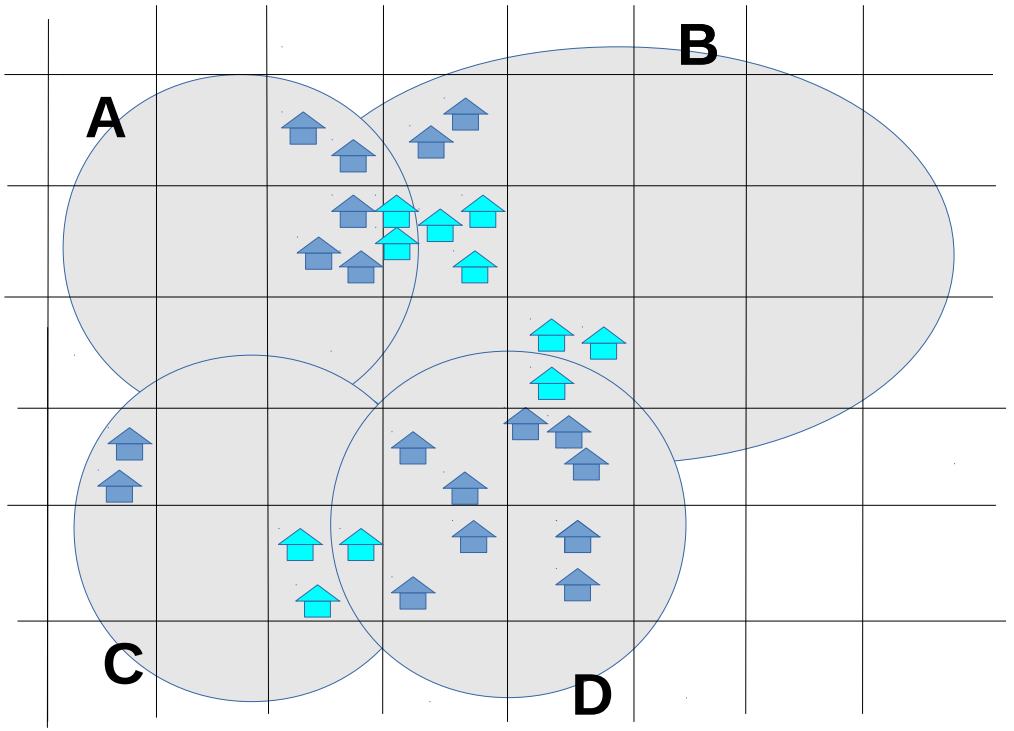
Pbm sur (A+B)
=X=> pbm sur A ou sur B

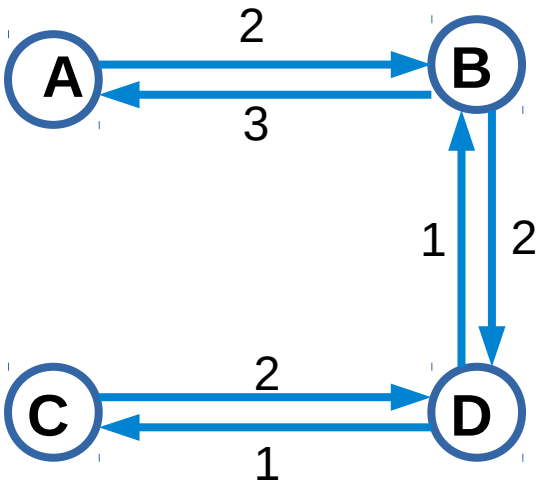
Représentation sous forme de graphe

- Graphe dont :
  - Les sommets sont les communes
  - Deux sommets sont reliés par une arête si les deux communes correspondantes sont contiguës
  - Les arêtes sont pondérées par le nombre d'observations « à la frontière »

- Le graphe comporte toute l'information utile pour déceler les problèmes de différenciation

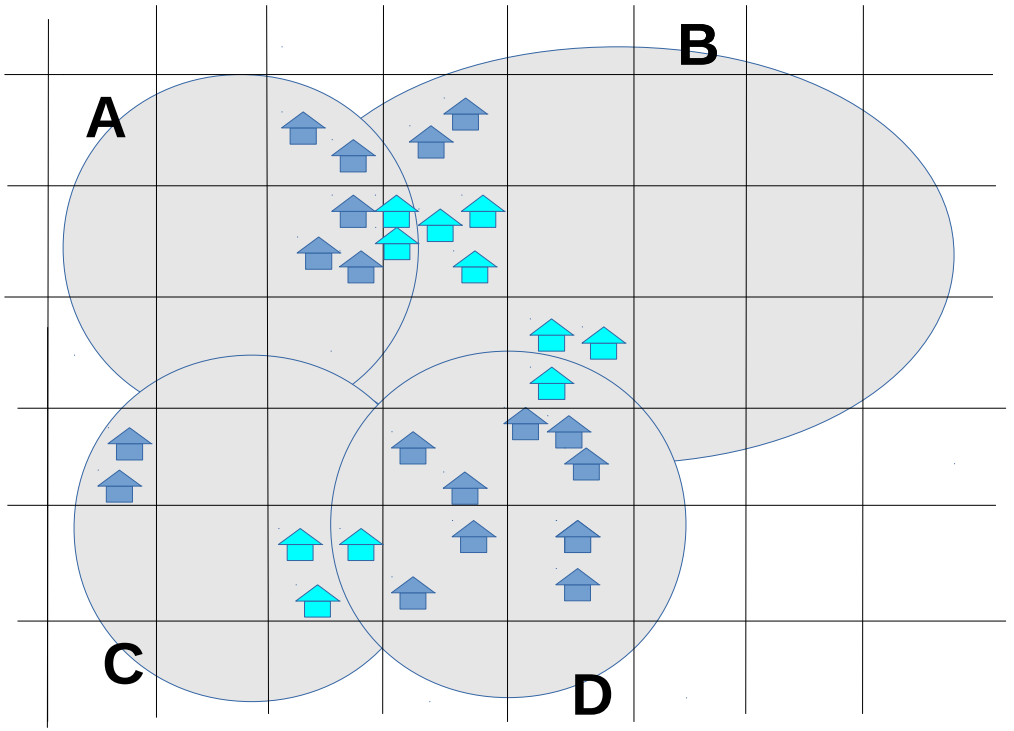
Exemple

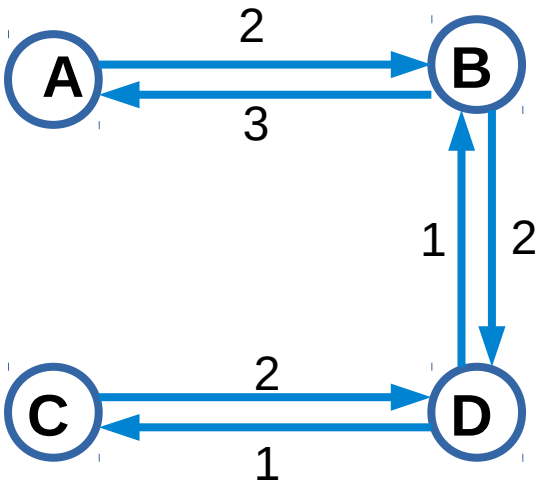




Les arêtes sont pondérées par le nombre d'observations à la frontière

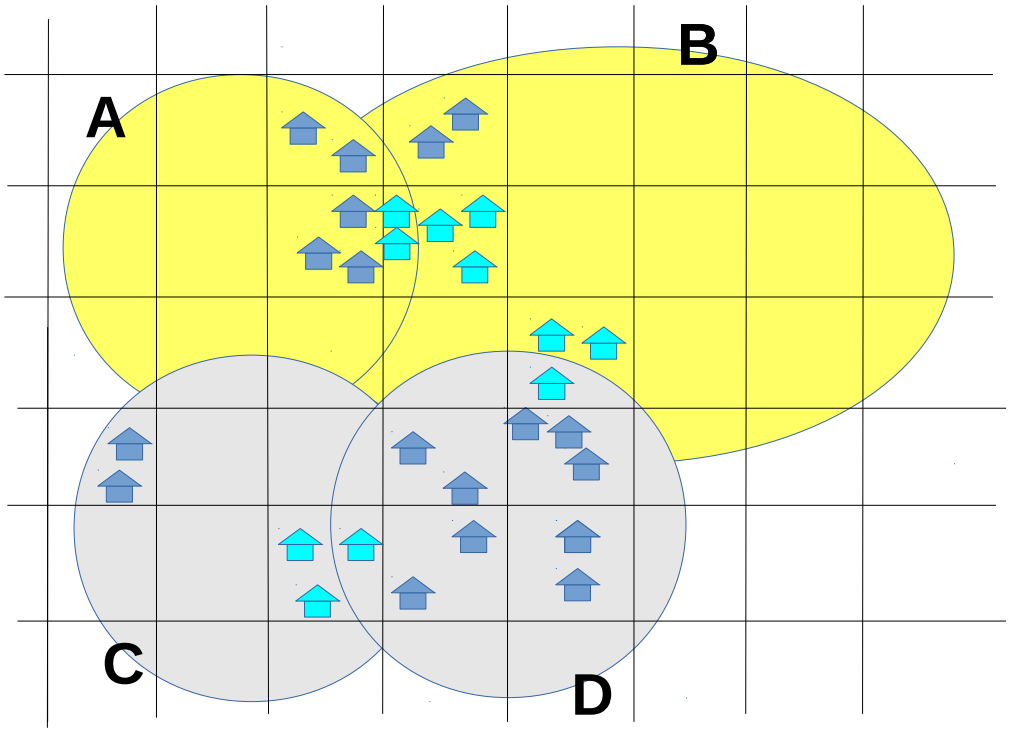
Exemple

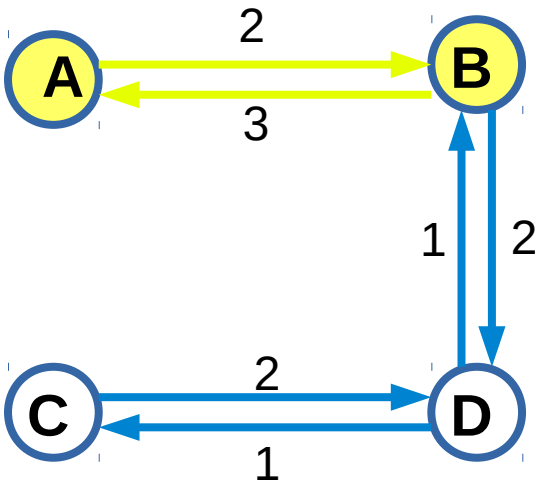




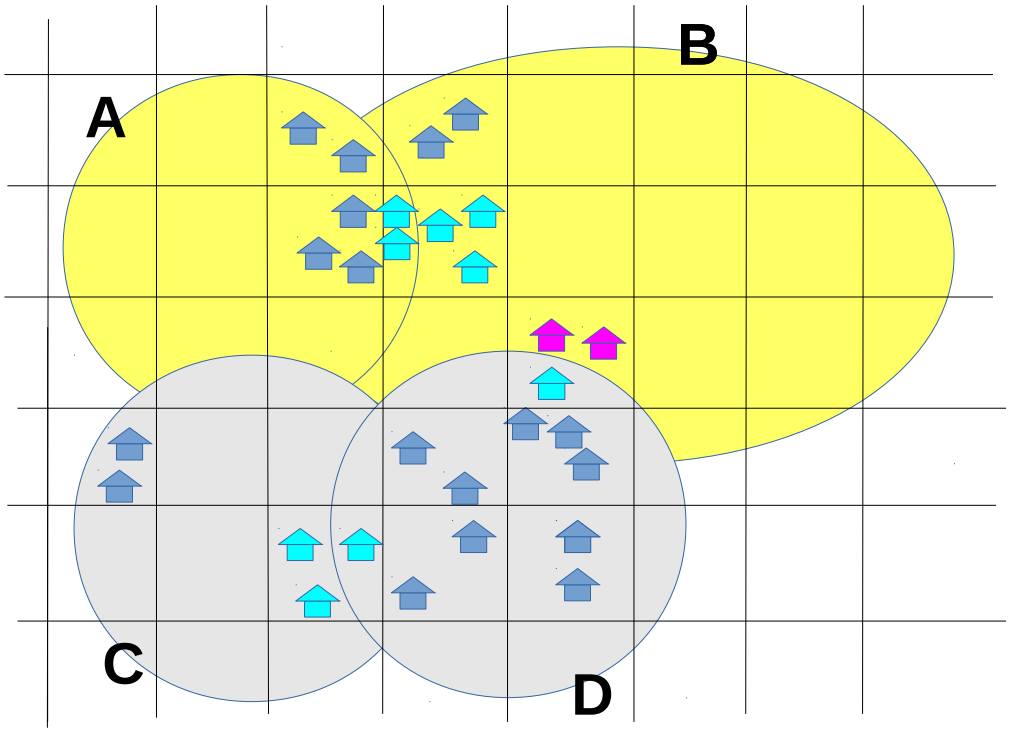
- 15 sous-graphes possibles
- seuls 9 sont connexes (i.e. dont les sommets sont reliés par des arêtes)

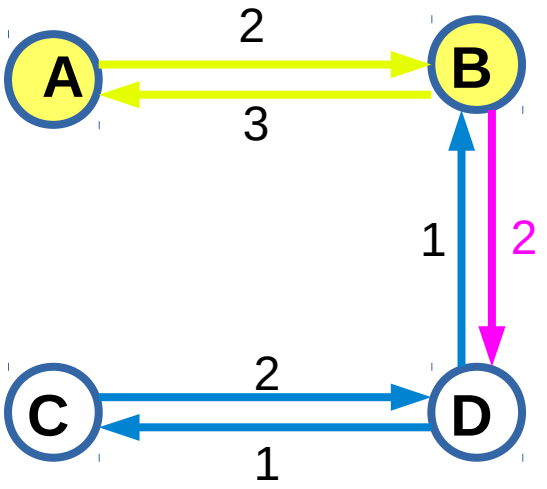
Exemple : sous-graphe A+B





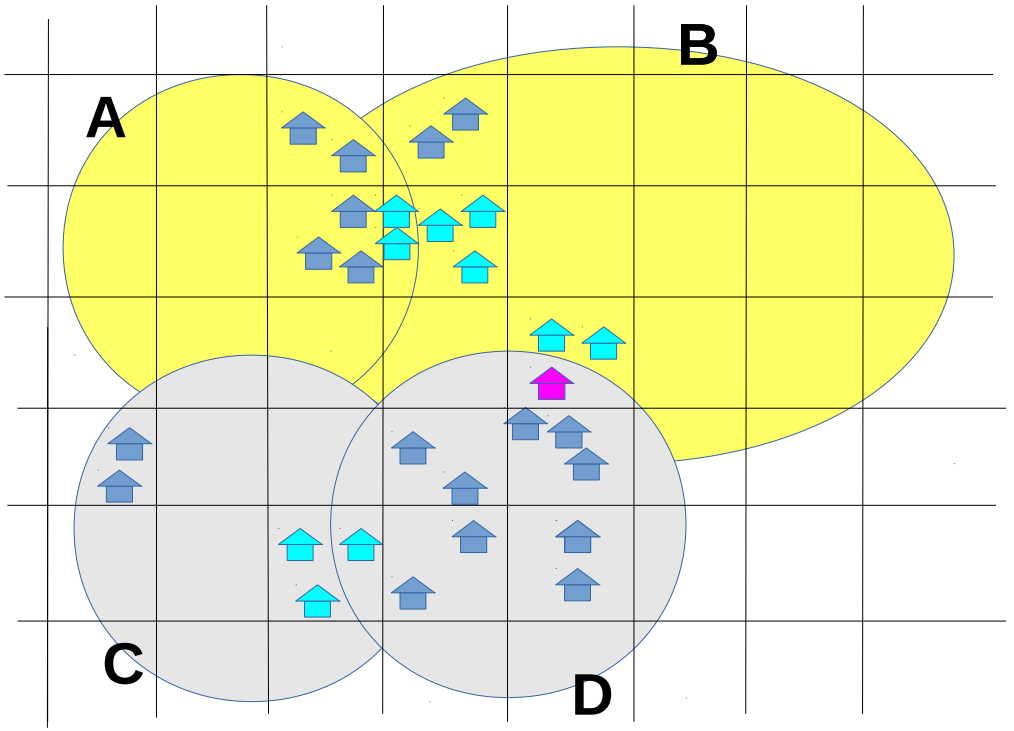
Exemple : sous-graphe A+B

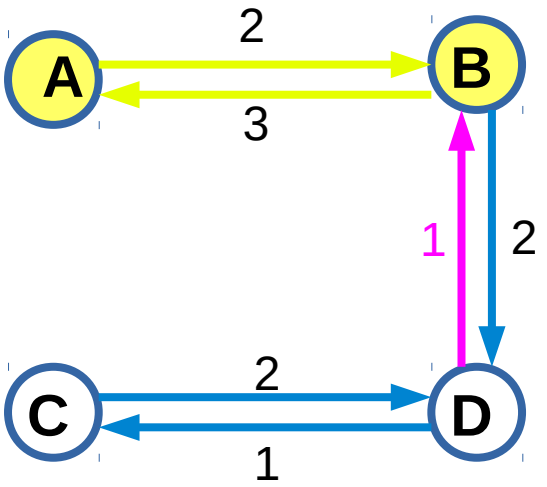




Somme des arêtes sortantes = différenciation « interne »

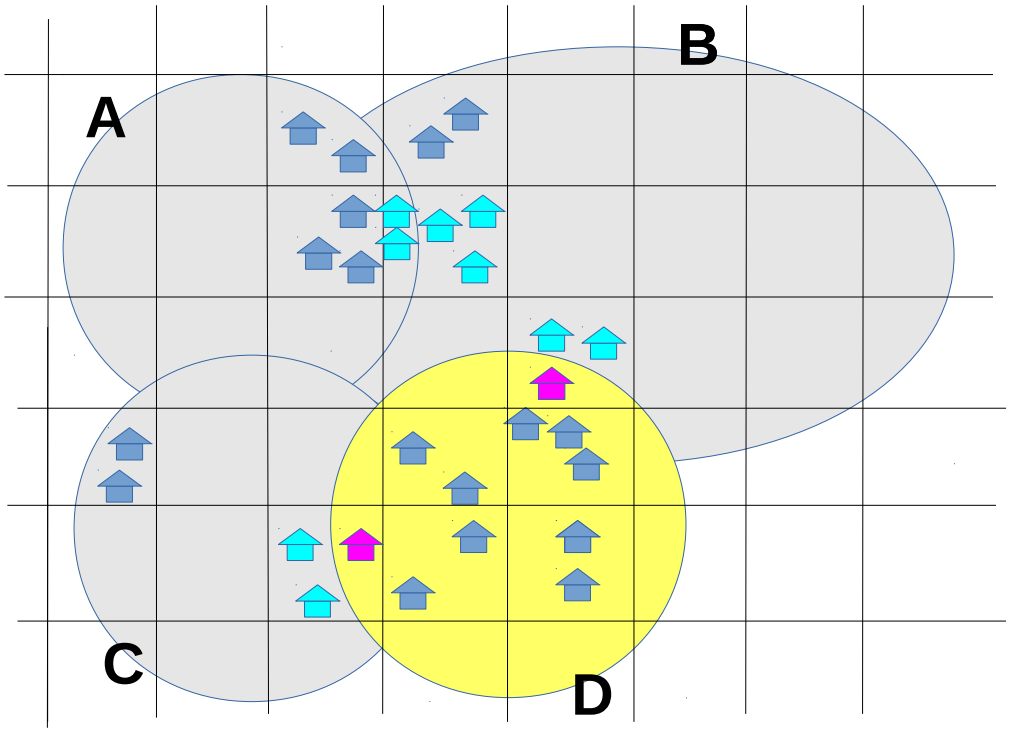
Exemple : sous-graphe A+B

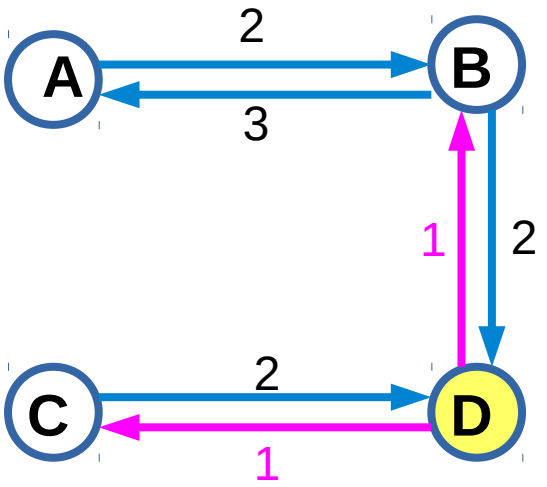




Somme des arêtes entrantes = différenciation « externe »

Exemple : sous-graphe D





Somme des arêtes sortantes = différenciation « interne »

Un nombre toujours très élevé de combinaisons

- Graphe pour la France :
  - 36 000 sommets
  - Chaque sommet a environ 6 voisins
  - ==> le nbr de sous-graphes connexes est très élevé : impossible de tous les tester

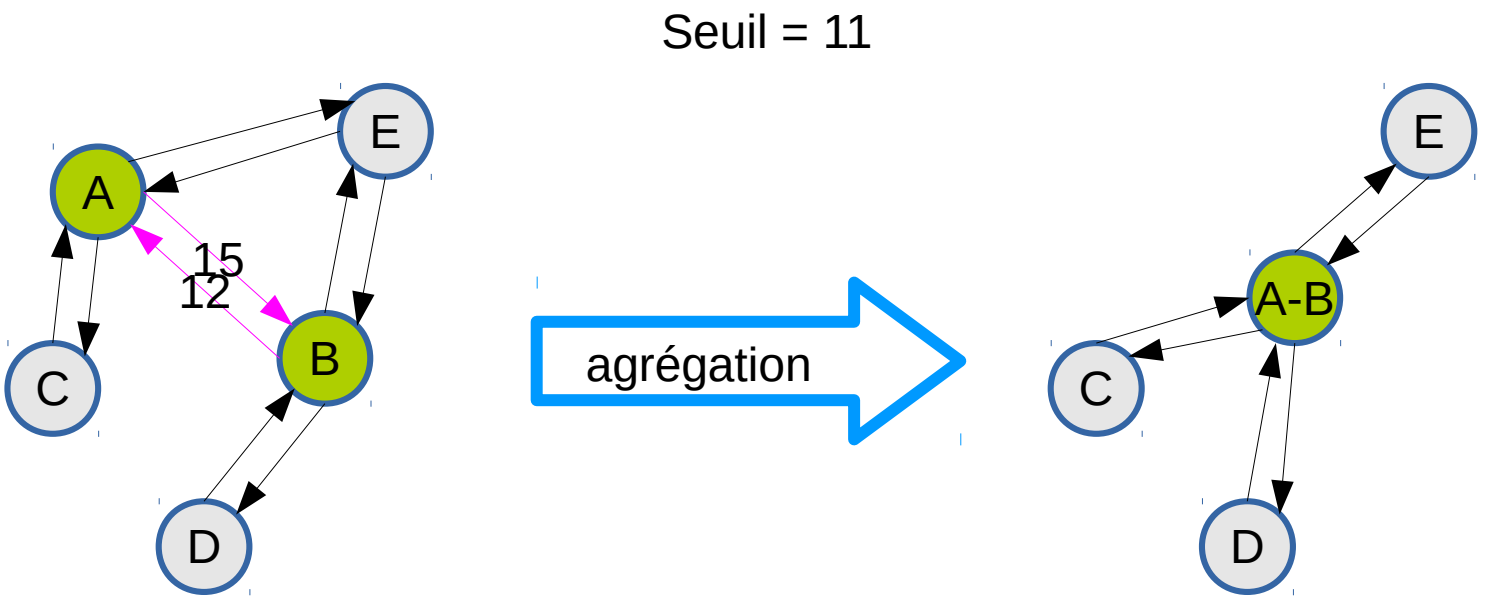
- Idée : simplifier le graphe en agrégeant des sommets entre eux

Simplification du graphe par agrégation

- Soient **A** et **B** deux sommets voisins
- Si pour tout sous-graphe connexe **G** contenant **A** et ne contenant pas **B** :
  - La somme des arêtes entrantes >= seuil
  - ET la somme des arêtes sortantes >= seuil
- Alors il ne peut pas y avoir de problème de différenciation en considérant A et B séparément
- On **agrège A et B en un seul et même sommet**

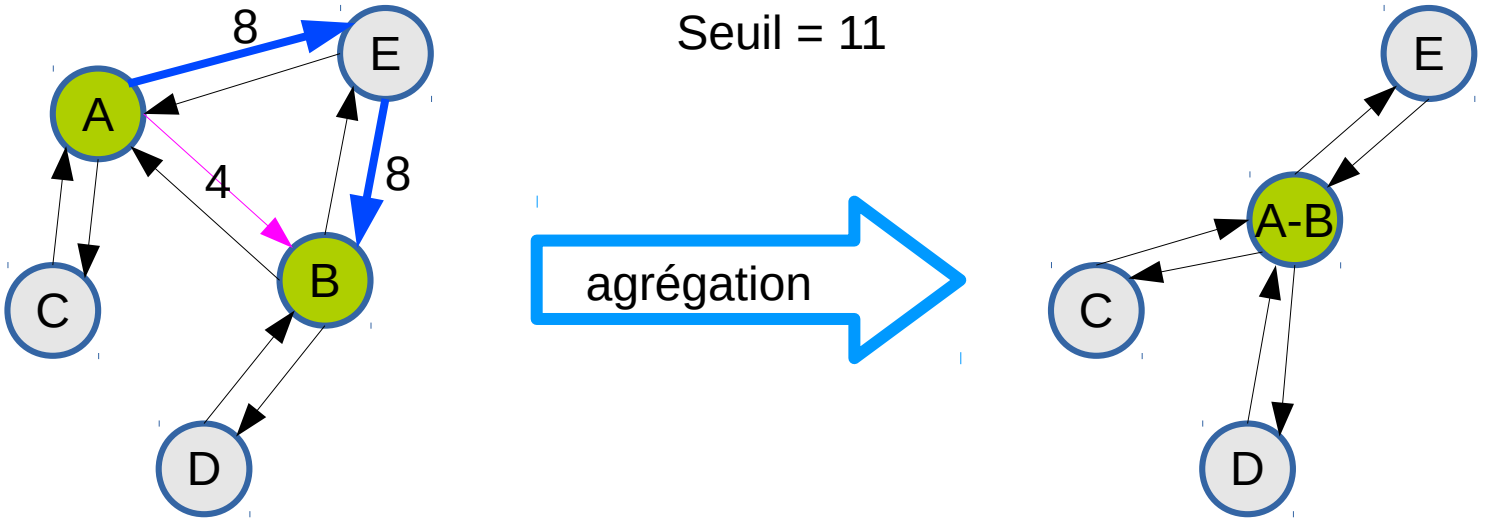
Agrégation du graphe - méthode 1

- On agrège les sommets A et B si
  - l'arête de A vers B >= seuil
  - l'arête de B vers A >= seuil



Agrégation du graphe – méthode 2

- On agrège les sommets A et B si :
  - l'arête e_A de A vers B < seuil
  - Il existe un autre chemin, de A vers B, (e_1, … , e_k) tel que e_A+e_i >=seuil pour tout i
  - Et inversement de B vers A



Recherche exhaustive sur le graphe simplifié

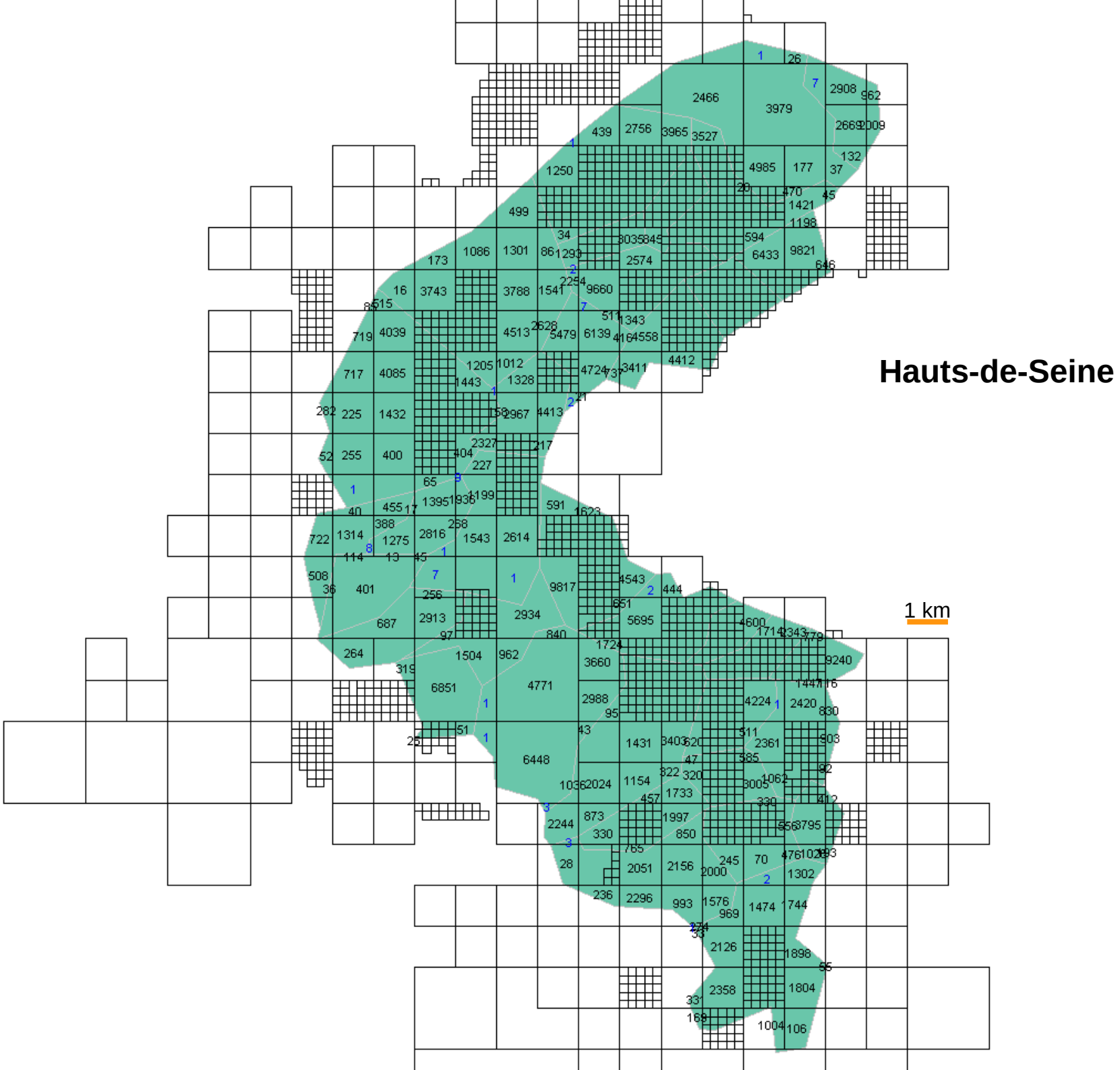
- Tester tous les sous-graphes connexes
- Sinon, tous les sous-graphes connexes de taille N donnée (par exemple, pour N de 1 à 15)
  - Programme en C++ de recherche exhaustive développé par la DSAU (Arlindo Dos Santos et François Sémécurbe)

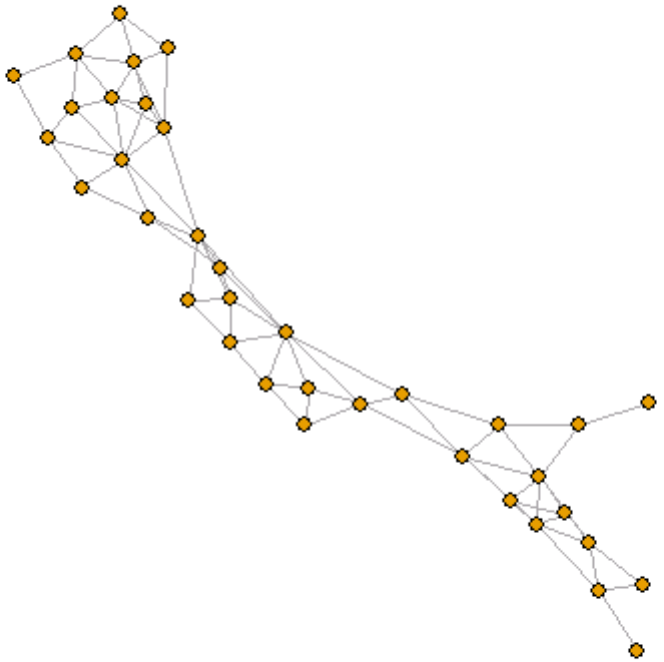
Résultats

- Source : Filosofi 2014, France métropolitaine

  - 27 millions de ménages diffusés sur
    - 36 000 communes
    - Et 144 000 carreaux

  - 244 000 croisements carreaux-communes dont
    - 68 000 sous le seuil de 11 ménages
    - A priori, 285 000 ménages à risque de différenciation

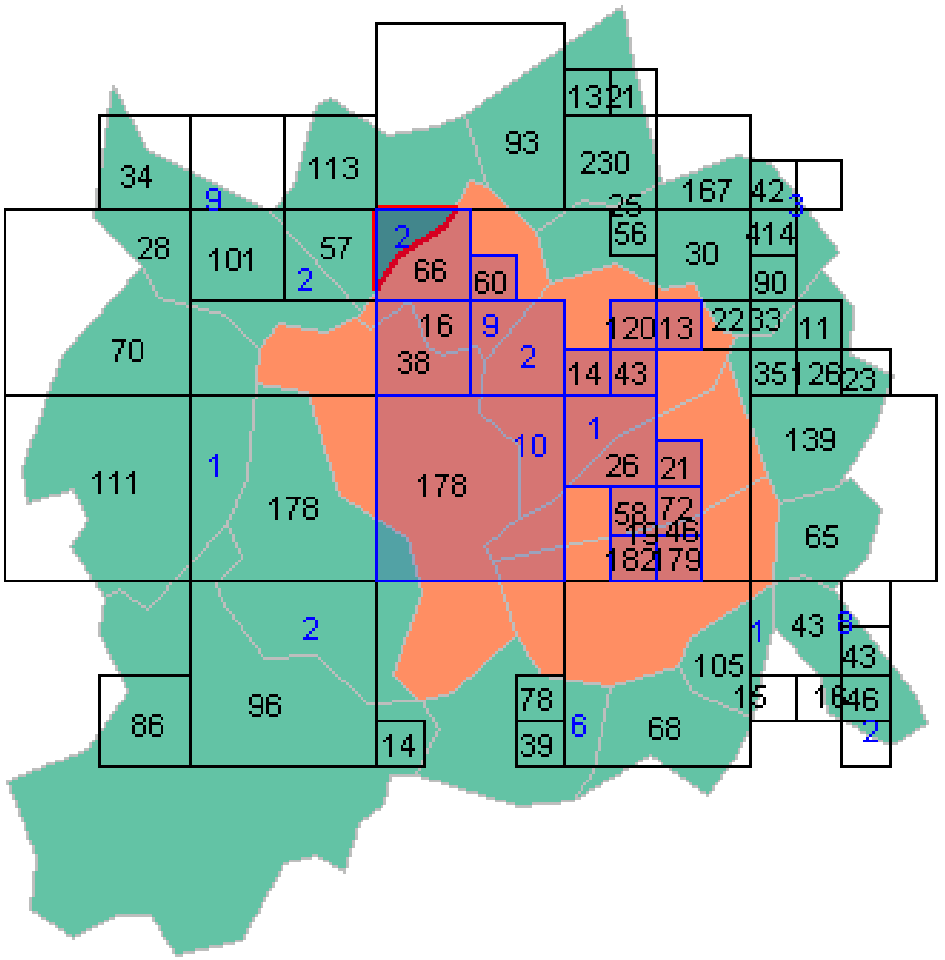

Hauts-de-Seine
1 km



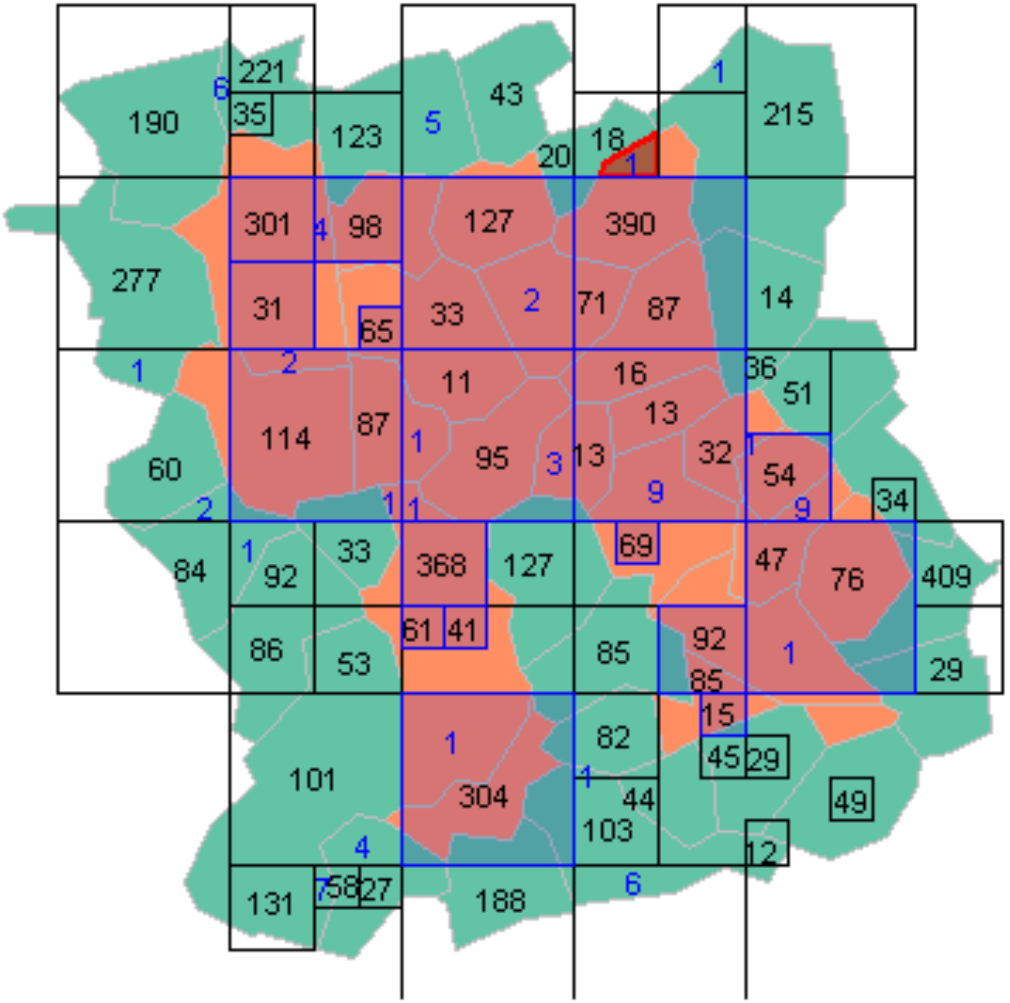
agrégation

Insee

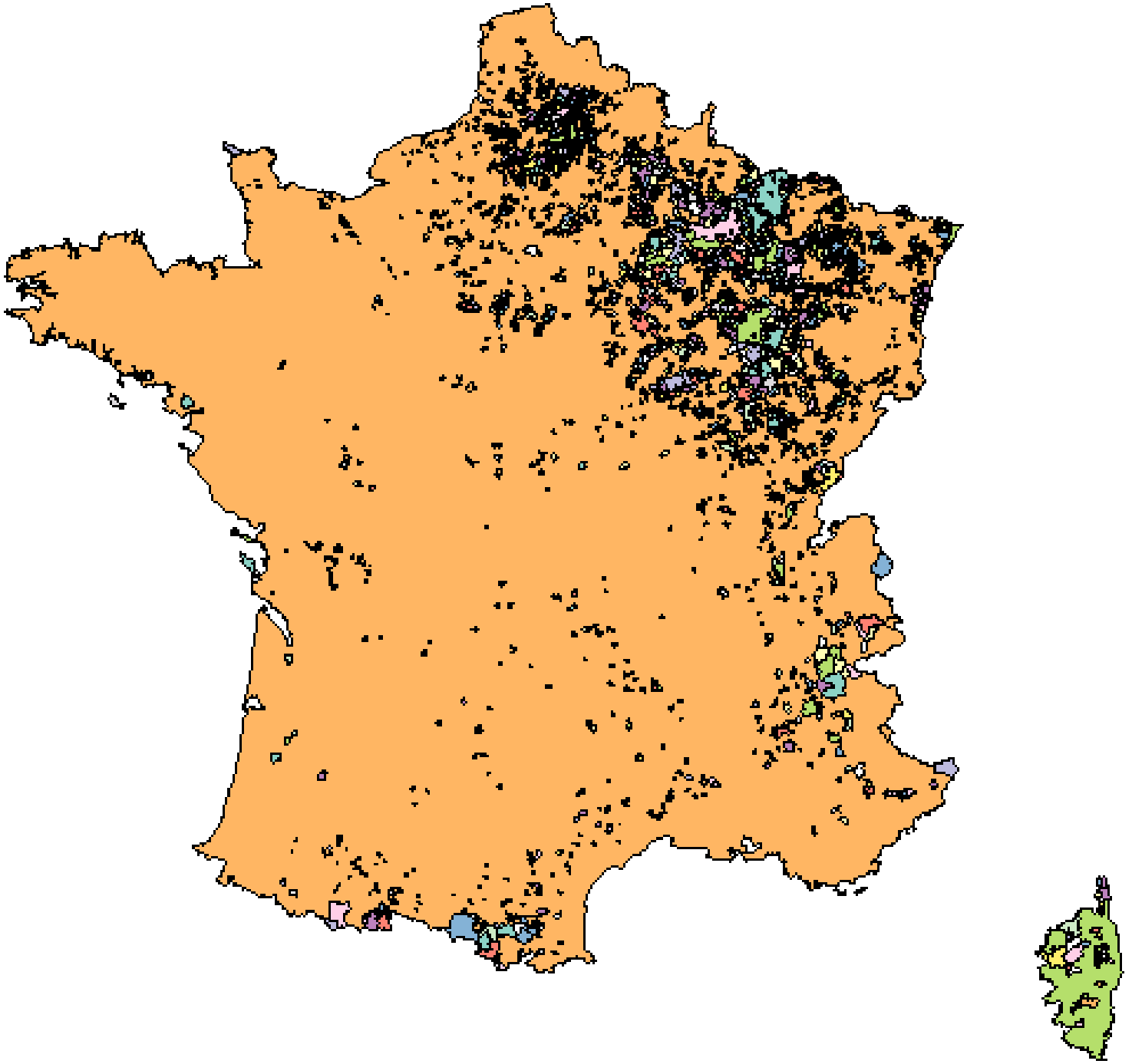
Exemples différenciation



Exemples différenciation



Communes agrégées

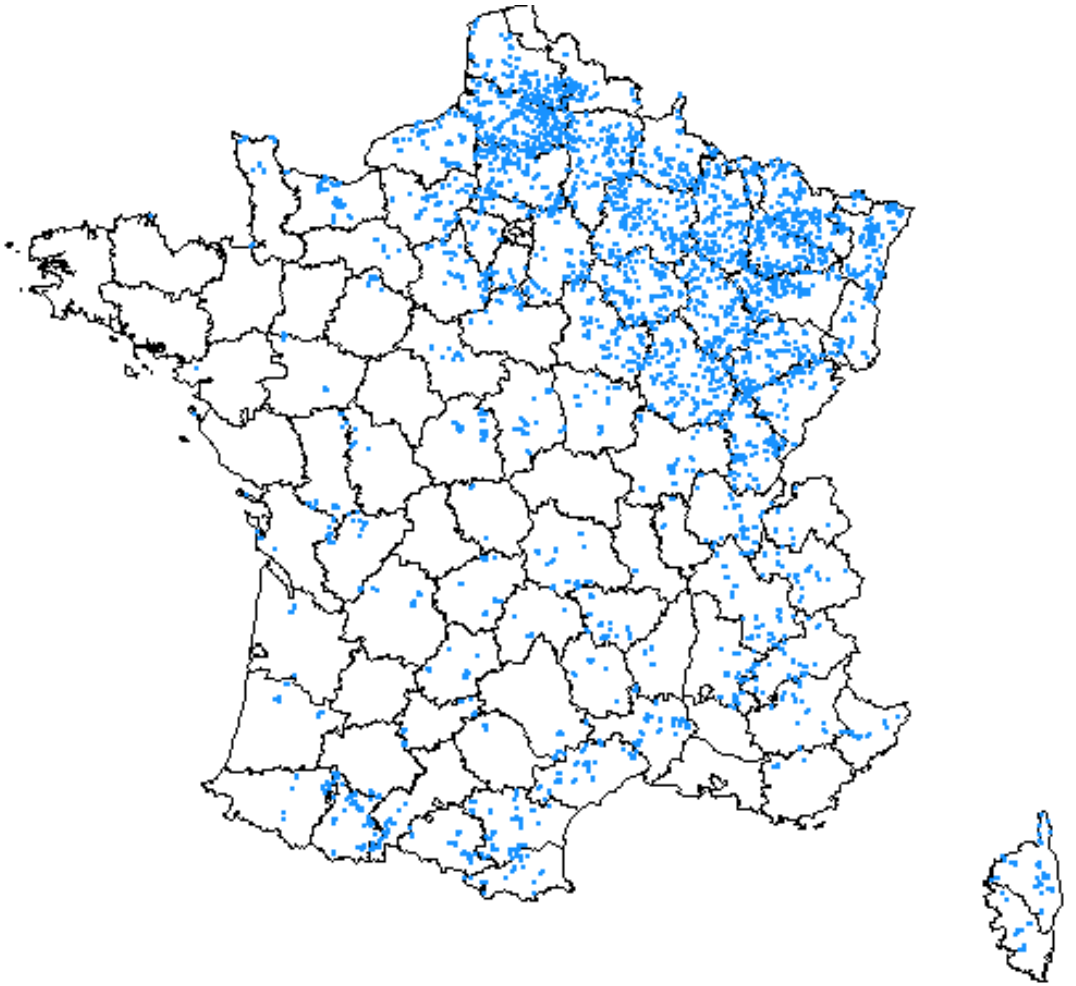


Résultats, France métropolitaine

	N_A	N_B	Nb. intersections sous le seuil	Nb. ménages sous le seuil	Nb. composantes connexes	Taille moyenne composantes connexes
État initial	36 671	144 706	68 418	285 727		
Premières simplifications	35 851	52 235	53 557	222 833	1 147	32, 0
Première agrégation	4 674	4 886	5 977	21 259	327	14, 3
Deuxième agrégation	2 740	2 822	2 937	10 884	327	8, 4

Résultats, France métropolitaine

Taille agrégat	1	2	3	4	5	6	7	8	9	10	Total
Nb. ménages	719	836	888	989	1074	1052	968	1232	1288	1439	10485
Prop. (en %)	6,9	8,0	8,5	9,4	10,2	10,0	9,2	11,8	12,3	13,7	100



Insee

Merci de votre attention

Vianney Costemalle
Responsable de la section Analyse Spatiale, DMRG
vianney.costemalle@insee.fr

Insee

www.insee.fr

@InseeFr