Aurait-on pu construire des régions selon des critères statistiques ?

Marc CHRISTINE, Insee, DMCSI

Michel ISNARD, Insee, Secrétariat général

Plan de la présentation

1. Introduction : objectif, outils
2. Quelques rappels historiques
3. Le cadre général de base
4. Extension au cas de plusieurs variables
5. Mise en œuvre de la méthode
6. Résultats
7. Conclusion

1. Objectif

- Résumer la construction de régions

- En prenant en compte des critères statistiques :
  - Assurer une ***homogénéité*** ou au contraire une ***hétérogénéité*** au vu de différentes variables descriptives
  - En nombre fixé
  - Assujetties à des contraintes de tailles
  - *Connexes*
  - Par agrégation d'unités élémentaires (régions actuelles, départements, cantons..)
  - .. entre lesquelles est définie une distance à partir des variables descriptives

- ***Pas de considérations politiciennes*** : on ne cherche pas à minimiser ou maximiser le nombre d'élus de tel ou tel parti.

- Comparer au travail du Parlement

Outils

- Une phase d'agrégation proche d'une ***classification ascendante hiérarchique***
- Sous différentes contraintes :
  - *Contiguïté*
  - Taille
  - Nombre d'unités créées

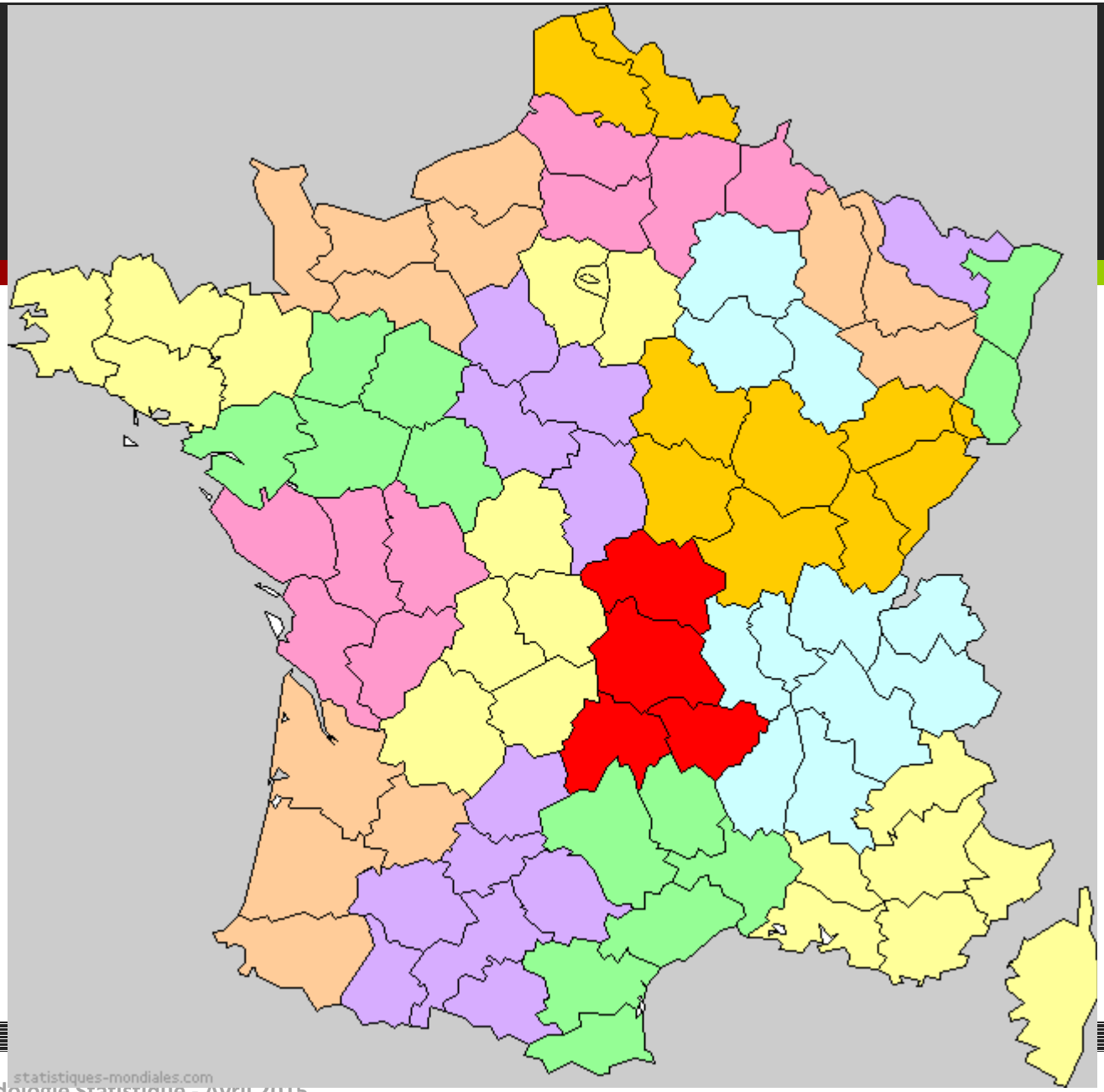
- Une seconde phase d'échange entre les classes créées précédemment afin de respecter les critères de taille et d'optimiser le critère d'homogénéité / hétérogénéité intra-classe.

- C'est l'occasion de revisiter quelques questions théoriques.

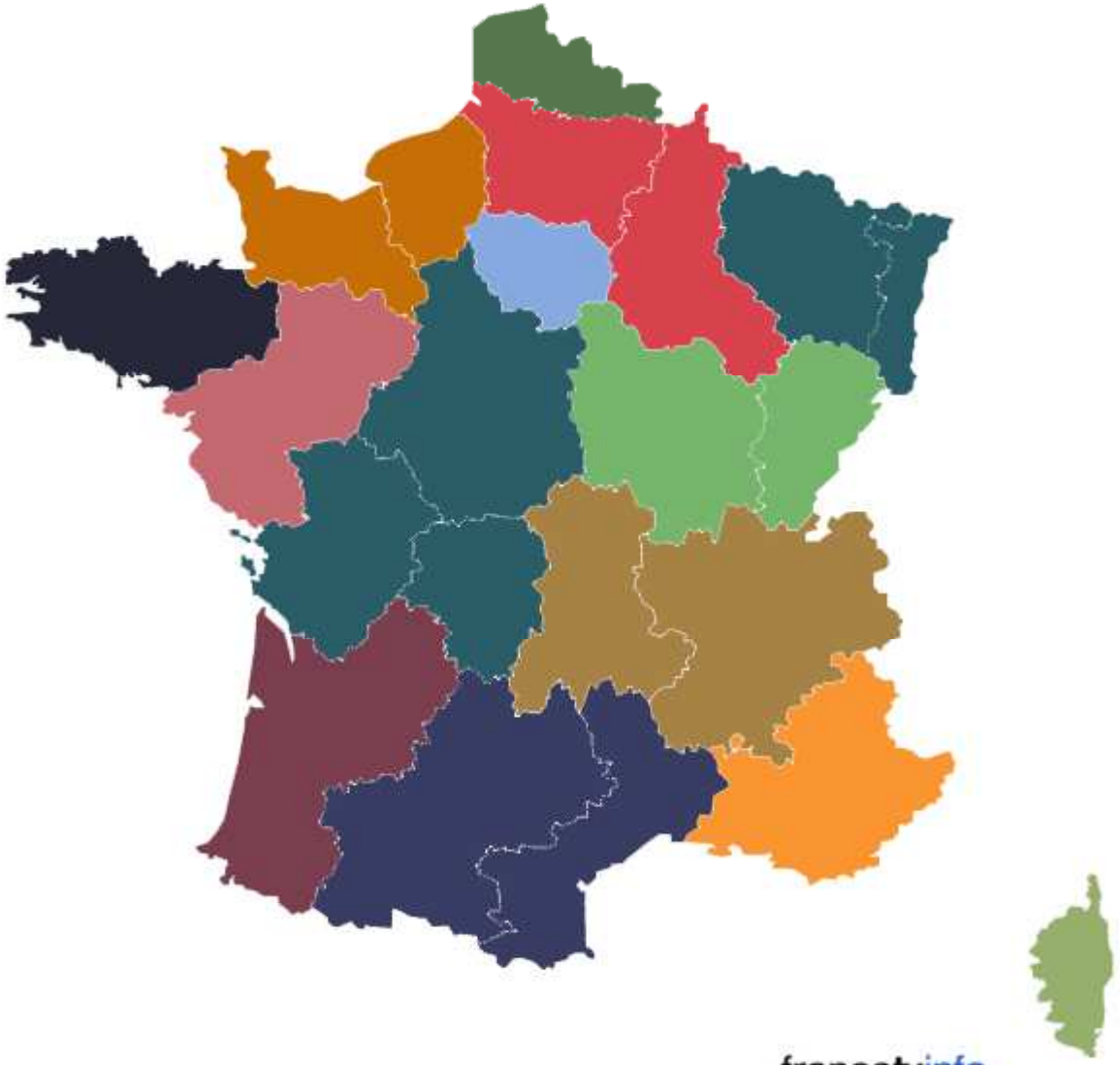
2. Un peu d'histoire de la région

- L'entre-deux guerres
- 1941 : régions, préfets régionaux
- 1948 : les Igames
- 1960 : circonscriptions d'action régionale
- 1969 : échec à la régionalisation !
- 1972 : établissements publics régionaux
- 1982 : vive la décentralisation !
- 1986 : 1es élections régionales
- 2015 : les nouvelles régions !

1941



Proposition de François Hollande
2 juin



Proposition du rapporteur à l'Assemblée nationale
7 juillet

13 régions (version députés)
17 juillet

15 régions (version sénateurs)
22 octobre

3. Le cadre général de base

- Pour chaque unité *i* de la population de référence $\mathcal{P}$:

- Une variable d'intérêt numérique ou vectorielle : χ_i
- Un poids : $\alpha_i > 0$
- Une taille numérique : T_i

- Centre de gravité de la population :

$$g = \frac{\sum_{i \in P} \alpha_i x_i}{\sum_{i \in P} \alpha_i}$$

- Variance ou *inertie* de la population :

$$I = \frac{\sum_{i \in P} \alpha_i (x_i - g)^2}{\sum_{i \in P} \alpha_i}$$

- Equation d'***analyse de la variance*** pour un partitionnement de $\mathcal{P}$ en K classes :

$$I = \sum_{k=1}^{K} \left[\frac{\omega_k}{\omega} [I_k + (g_k - g)^2] \right]$$

Avec : $\omega = \sum_{i \in P} \alpha_i$ et $\omega_k = \sum_{i \in P_k} \alpha_i$

I_k = inertie propre à chacune des classes $\mathcal{P}_k$

- On cherchera à *minimiser* ou *maximiser* l'inertie intra-classe :

$$I^a = \sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k$$

- … qui s'écrit aussi :

$$I^a = \frac{1}{2\omega} \sum_{k=1}^{K} \frac{1}{\omega_k} \left[\sum_{i,j \in P_k} \alpha_i \alpha_j (x_i - x_j)^2 \right]$$

- Agrégation de deux classes (***cas euclidien***) :
  - La variation d'inertie intra-classe résultante est :

$$\Delta I^a = \frac{(g_{k_1} - g_{k_2})^2}{\omega} \frac{\omega_{k_1}\omega_{k_2}}{\omega_{k_1} + \omega_{k_2}}$$

  - Elle est positive ou nulle.
  - Si l'on agrège une unité à une classe, la variation devient

$$\Delta I^a = \frac{(g_k - x_0)^2}{\omega} \frac{\omega_k \alpha_0}{\omega_k + \alpha_0}$$

- Cette variation d'inertie intra-classe peut s'interpréter :
  - Comme une distance entre deux classes...
  - … ou comme une distance entre une unité et une classe,
- … calculées à partir des centres de gravité de ces classes et des poids afférents.
- La minimisation de l'inertie intra-classe peut se faire au moyen d'un algorithme qui, à chaque étape, ***agrège les unités les plus proches au sens de cette distance***.

Agrégation en cascade

- On peut partir d'une population de référence déjà constituée à partir d'une agrégation préliminaire d'unités élémentaires
- Exemple : ménages => communes.
- Cela conduit à repartir d'une inertie définie par le 1er terme de l'équation d'analyse de la variance :

$$I = \sum_{k=1}^{K} \frac{\omega_k}{\omega} (g_k - g)^2 + \underbrace{\sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k}_{\substack{\textit{terme omis,} \\ \textit{lié à l'agrégation} \\ \textit{initiale}}}$$

- 3 types d’unités statistiques :
  - Unités statistiques **élémentaires**
  - **Agrégats de base** : communes, cantons…
  - **Agrégats composites**, constitués par regroupement des agrégats de base lors de la procédure d‘agrégation.

Réflexion sur les poids

- Il doit y avoir cohérence entre les poids et les variables définies sur les agrégats de base..
- .. pour que l'agrégation et le recalcul, sur l'agrégat constitué, d'un centre de gravité au moyen de ces poids, ait un sens.
- Ceci peut poser problème si l'on travaille avec plusieurs variables descriptives simultanément.

Unité statistique élémentaire (i)	Variable x_i	Agrégat de base (k)	Variable g_k	Poids ω_k
Ménage	Revenu	Commune	Revenu moyen des ménages de[1] la commune	Nombre de ménages de la commune
Individu	Age	Commune	Age moyen des individus de la commune	Nombre d'individus de la commune
Individu	1 si l'individu i a plus de 50 ans, 0 sinon	Commune	Proportion des individus de plus de 50 ans résidant dans la commune	Nombre d'individus de la commune
Électeur	1 si l'individu i a voté pour le candidat A aux élections présidentielles, 0 sinon	Commune (d'inscription sur les listes électorales)	Proportion des individus inscrits sur les listes électorales de la commune k (ou ayant voté et exprimé un suffrage) ayant voté pour le candidat A	Nombre d'individus inscrits (ou ayant voté et exprimé un suffrage) sur les listes électorales de la commune
Individu en emploi	Catégorie socio-professionnelle	Commune du lieu de travail	Répartition des personnes en emploi sur une commune par catégorie socio-professionnelle	Nombre d'individus en emploi dans la commune
Individu de 15 ans ou plus.	Statut d'activité (actif occupé / chômage / en formation / autre inactif)	Commune	Répartition des personnes de 15 ans ou plus par statut d'activité	Nombre de personnes de 15 ans ou plus dans la commune
Femme	Age	Commune	Répartition des femmes par âge	Nombre de femmes dans la commune

4. Extension au cas de plusieurs variables

- Plusieurs variables numériques, mêmes poids
- Plusieurs variables numériques, poids différents
- Variables catégorielles
- Mélange

A/ Plusieurs variables numériques, mêmes poids

- Méthode utilisée :
  - Normaliser chaque variable en divisant par l'écart-type
  - Additionner les inerties relatives à chacune de ces variables. Avec *p* variables :

$$I = \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{K}\frac{\omega_k}{\omega}(g'_{i,k} - g'_i)^2 \qquad g'_{i,k} = \frac{g_{i,k}}{\sqrt{\sum_{l=1}^{K}\frac{\omega_l}{\omega}(g_{i,l} - g_i)^2}} = \frac{g_{i,k}}{\sigma_i}$$

- D'autres méthodes possibles pour tenir compte et annihiler l'effet des corrélations entre variables (réduction générale, ACP …)

B/ Plusieurs variables numériques, poids différents

- On additionne les inerties résultant de chacune des variables, chacune avec son système de poids

- L'inertie de chaque variable est normalisée à 1

- Les valeurs des variables relatives à l'agrégat constitué sont calculées comme barycentres affectés des poids correspondant à chaque variable.

C/ Variables catégorielles

- Variable catégorielle à *Q* modalités

$n_{k,q}$: nombre d'unités élémentaires de l'agrégat k appartenant à la modalité q de la variable

$p_{k,q} = \frac{n_{k,q}}{n_k}$: proportion d'unités de l'agrégat correspondant à la modalité q de la variable

Distance du χ^2 entre les agrégats k et l : $$d_{k,l}^2 = \sum_{q=1}^{Q} \frac{(p_{k,q} - p_{l,q})^2}{\pi_q}$$

$$\pi_q = \frac{1}{N} \sum_{k=1}^{K} n_{k,q} = \frac{n_{\bullet,q}}{N}$$

On peut écrire cette distance sous la forme : $$d_{k,l}^2 = \sum_{q=1}^{Q} \left(\frac{p_{k,q}}{\sqrt{\pi_q}} - \frac{p_{l,q}}{\sqrt{\pi_q}} \right)^2$$

Cela équivaut à la distance euclidienne relative aux variables : $$Z_{k,q} = \frac{p_{k,q}}{\sqrt{\pi_q}} = \frac{n_{k,q}}{n_k} \frac{\sqrt{N}}{\sqrt{n_{\bullet,q}}}$$

Les poids utilisés seront : $\alpha_k = \frac{n_k}{N}$

- On traitera les variables catégorielles comme des quantitatives au moyen de cette conversion.

D/ Mélange

- Les variables catégorielles sont converties en quantitatives

- L'inertie de chaque groupe de quantitatives issues d'une même variable catégorielle est normalisée à 1

- On additionne l'inertie de chaque variable quantitative, également normalisée.

5. Mise en œuvre de la méthode

- Application à la **France continentale**
- **12 macro-régions à construire**…
- … à partir d'une agrégation de cantons, de zones d'emploi, d'arrondissements, de départements ou des régions actuelles
- **contraintes** :
  - de taille : ***[4%, 19%]*** de la population totale
  - de connexité
- **simulations en faisant varier** :
  - le critère (minimisation / maximisation)
  - les variables utilisées : revenu / 6 tranches d'âge
  - l'agrégat statistique de base

La première phase

- Bâtie sur le même algorithme qu'une CAH
- Mais on limite l'agrégation aux groupes *contigus*, c'est-à-dire aux classes dont au moins une unité est contiguë à une unité de l'autre classe.
- A chaque étape, on choisit les classes à agréger : ceux dont l'agrégation minimise ou maximise la variation d'inertie intra-classe
- **A l'issue de cette phase, les classes formées sont connexes, mais ne respectent pas forcément les contraintes de taille**

La seconde phase

- La seconde phase tente de vérifier les contraintes de taille fixées par l'utilisateur et procède à des échanges entre les classes constituées, ***tout en respectant la connexité***.

- Elle se base sur un critère prenant en compte, pour chaque classe, le ***carré de l'écart entre sa taille et la taille minimale ou maximale***. Il vaut 0 si la taille vérifie les contraintes.

- Après la seconde phase, diminution de ce critère **mais** variation de l'inertie intra : ***on peut perdre en optimalité***.

6. Résultats

A/ Effet de la seconde phase

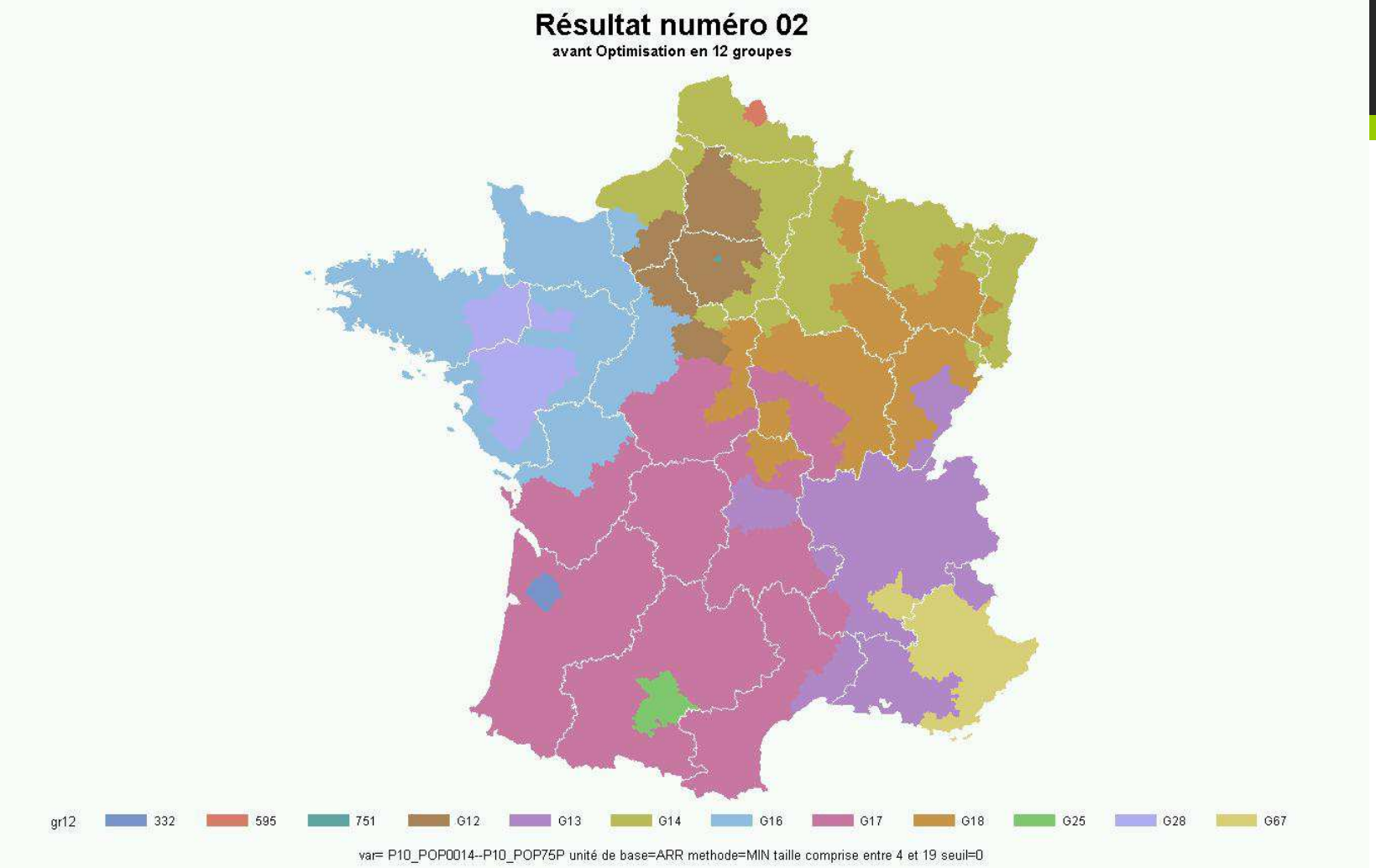
- L'objectif est de mettre en évidence l'effet de la seconde phase sur l'inertie et le respect des seuils de taille.

- Variables utilisées : Population en 6 tranches d'âge, agrégats de base = ARR, minimisation de l'inertie intra et pas de seuil d'étendue géographique des régions créées.

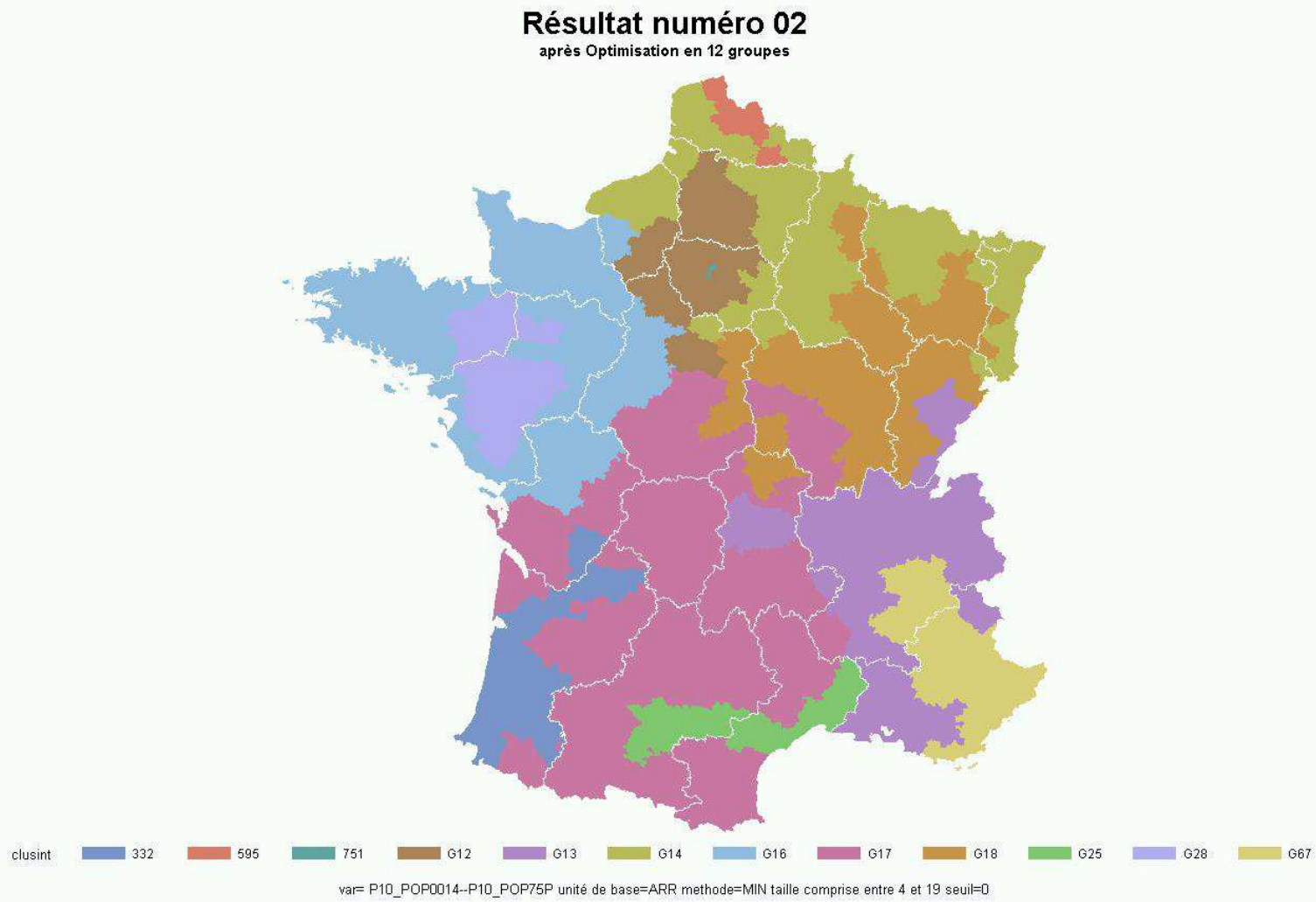
POP – ARR – MIN – 0

Indicateur	Avant	Après
Critère	15.92	0
In. Intra	0.34	0.41
T. Min	1.39% (332)	4.01% (332)
T. Max	18.53% (G12)	17.87% (G12)

POP – ARR – MIN – 0 – Avant



POP – ARR – MIN – 0 – Après



B/ Minimisation ou Maximisation ?

- Deux manières différentes de concevoir l'agrégation :
  - MIN : on crée des groupements ***homogènes***, donc très différents les uns des autres ;
  - MAX : on crée des groupements ***hétérogènes***, donc semblables, c'est-à-dire des « petites France »

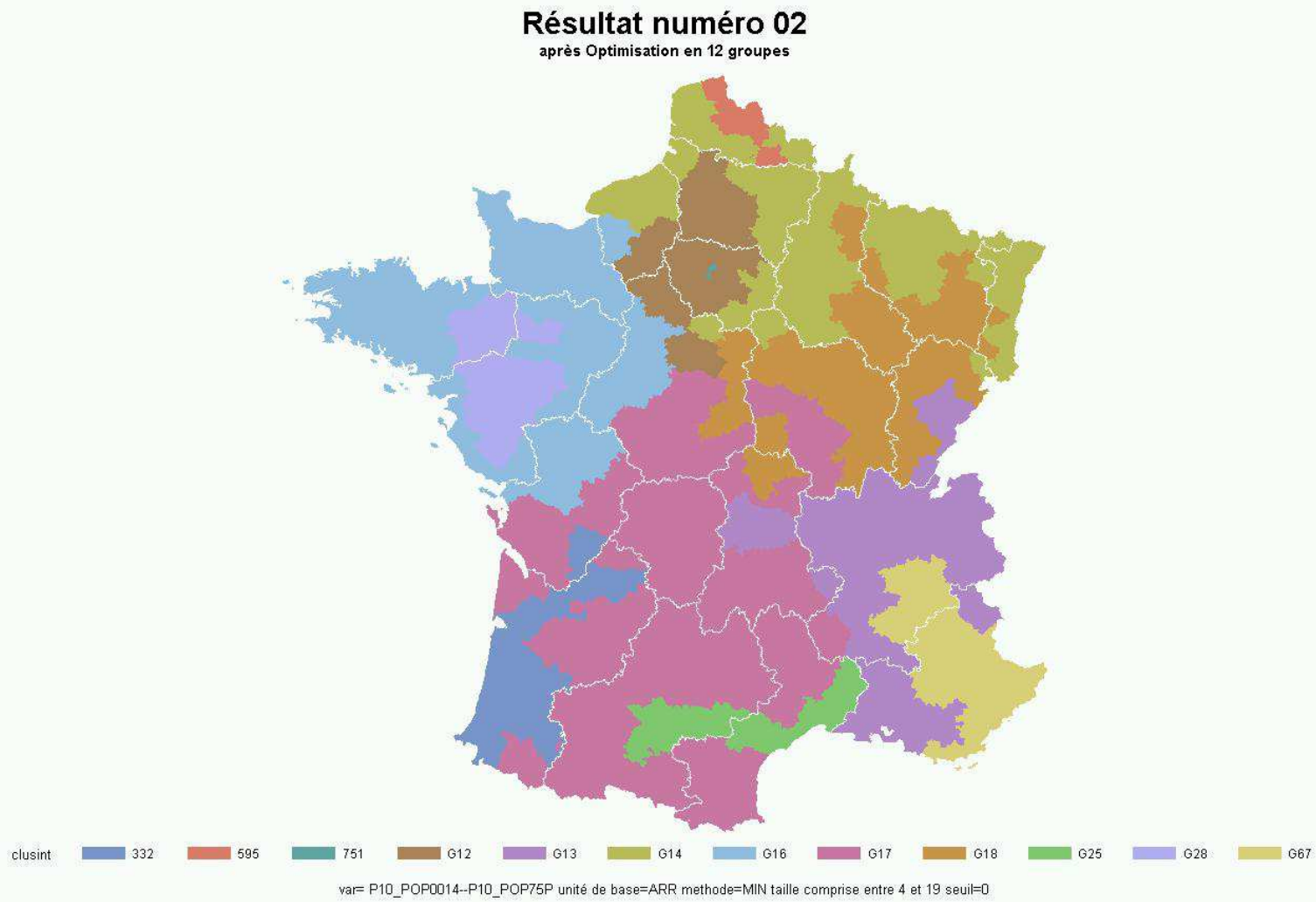
- Variable = Population en 6 tranches d'âge, agrégats de base=ARR et pas de seuil d'étendue géographique.

- Situation ***après*** la seconde phase.

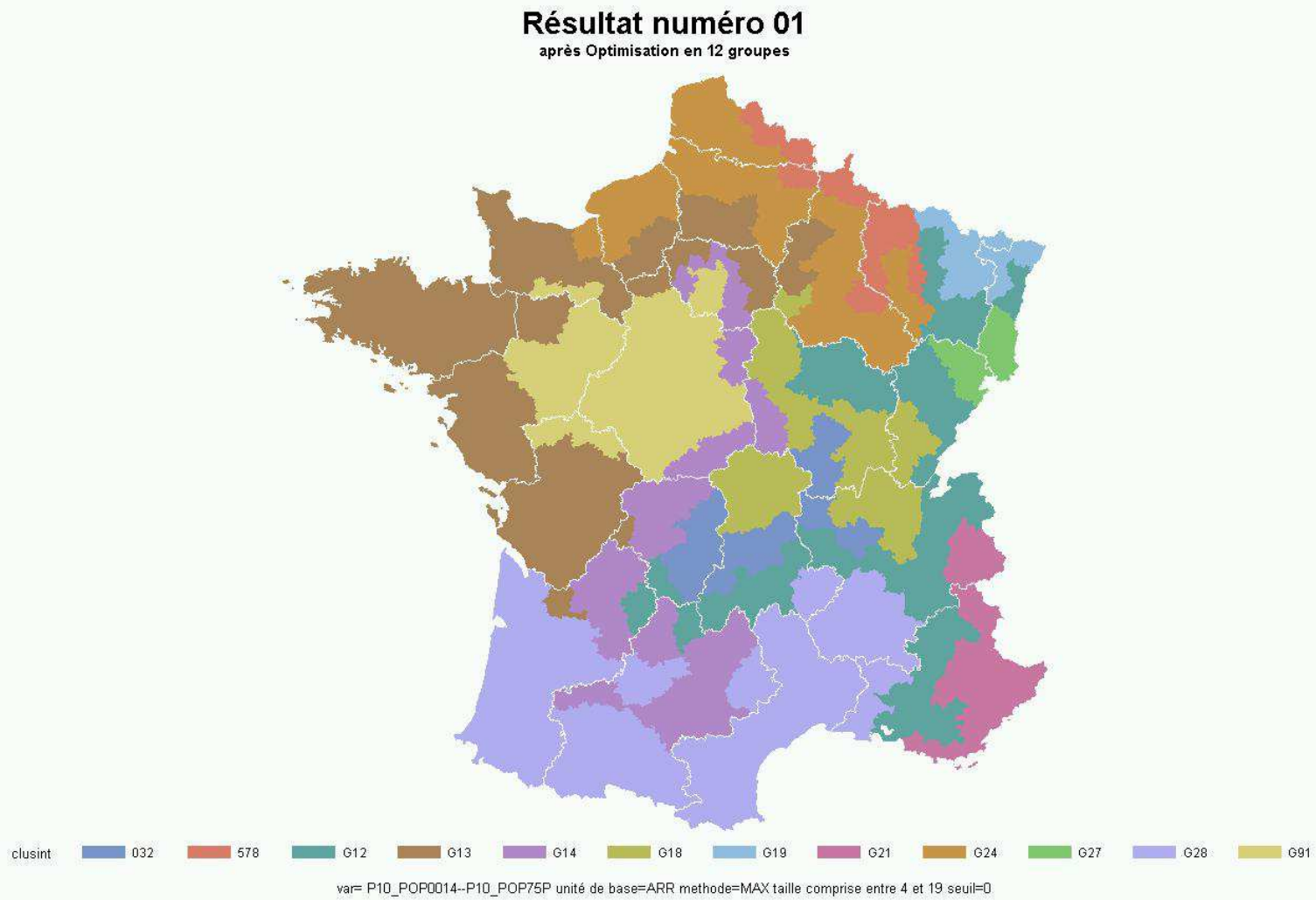
POP – ARR – 0

Indicateurs	MAX	MIN
Inertie	0.83	0.41
T. Min	1.88 (G19)	4.01 (332)
T. Max	18.76 (G13)	17.87 (G12)

POP – ARR – 0 – MIN


Résultat numéro 02
après Optimisation en 12 groupes
clusint
332
595
751
G12
G13
G14
G16
G17
G18
G25
G28
G67
var= P10_POP0014--P10_POP75P unité de base=ARR methode=MIN taille comprise entre 4 et 19 seuil=0

POP – ARR – 0 – MAX


Résultat numéro 01
après Optimisation en 12 groupes
clusint
032
578
G12
G13
G14
G18
G19
G21
G24
G27
G28
G91
var= P10_POP0014--P10_POP75P unité de base=ARR methode=MAX taille comprise entre 4 et 19 seuil=0

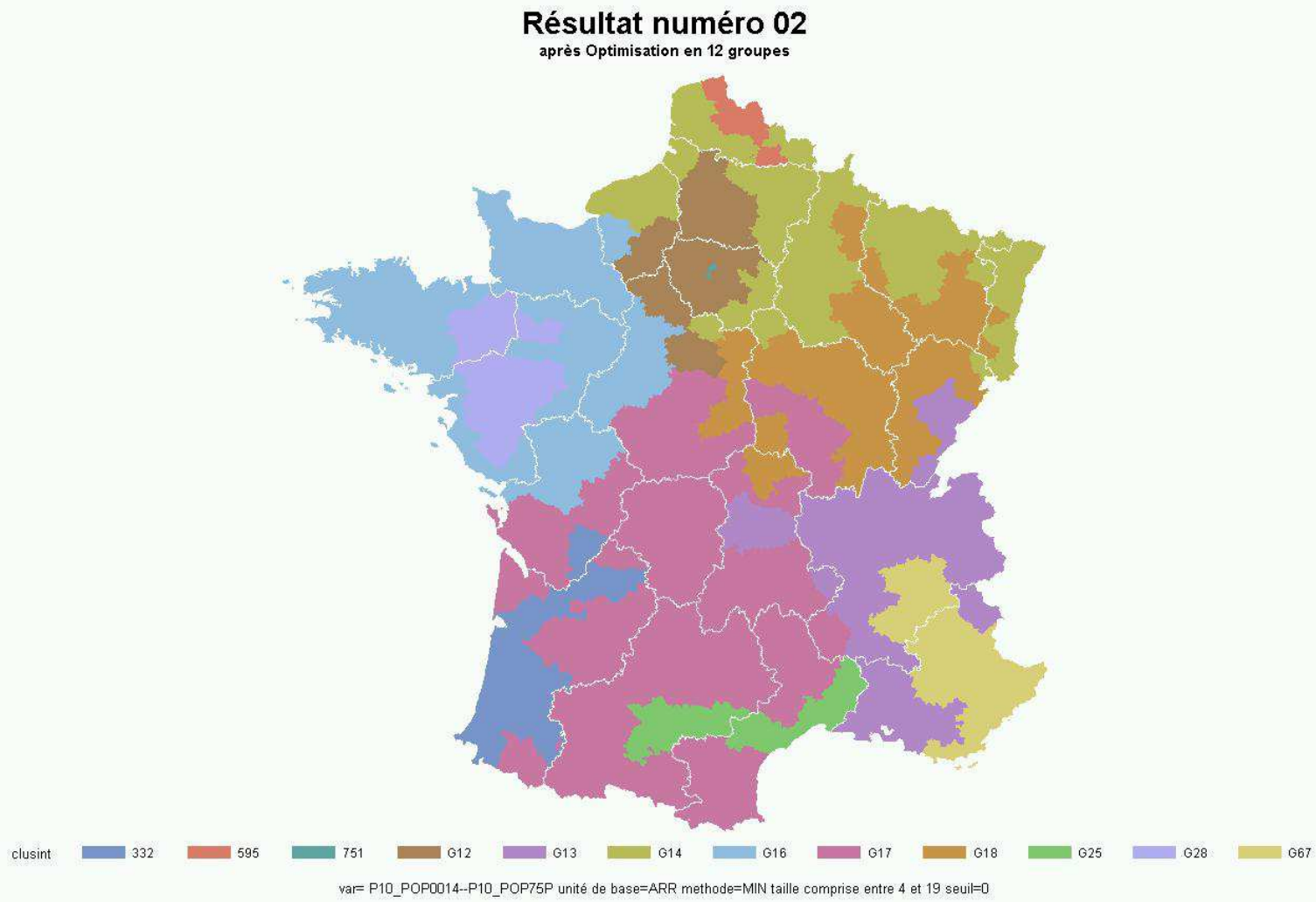
C/ Utilisation de plusieurs variables

- Variables utilisées :
  - Population en 6 tranches d'âge (POP)
  - Médiane des revenus des ménages (REV)
  - Population en 6 tranches d'âge et revenu (POP&REV)
- agrégats de base=ARR , MINimisation de l'inertie intra et pas de seuil d'étendue géographique.
- Situation ***après*** la seconde phase

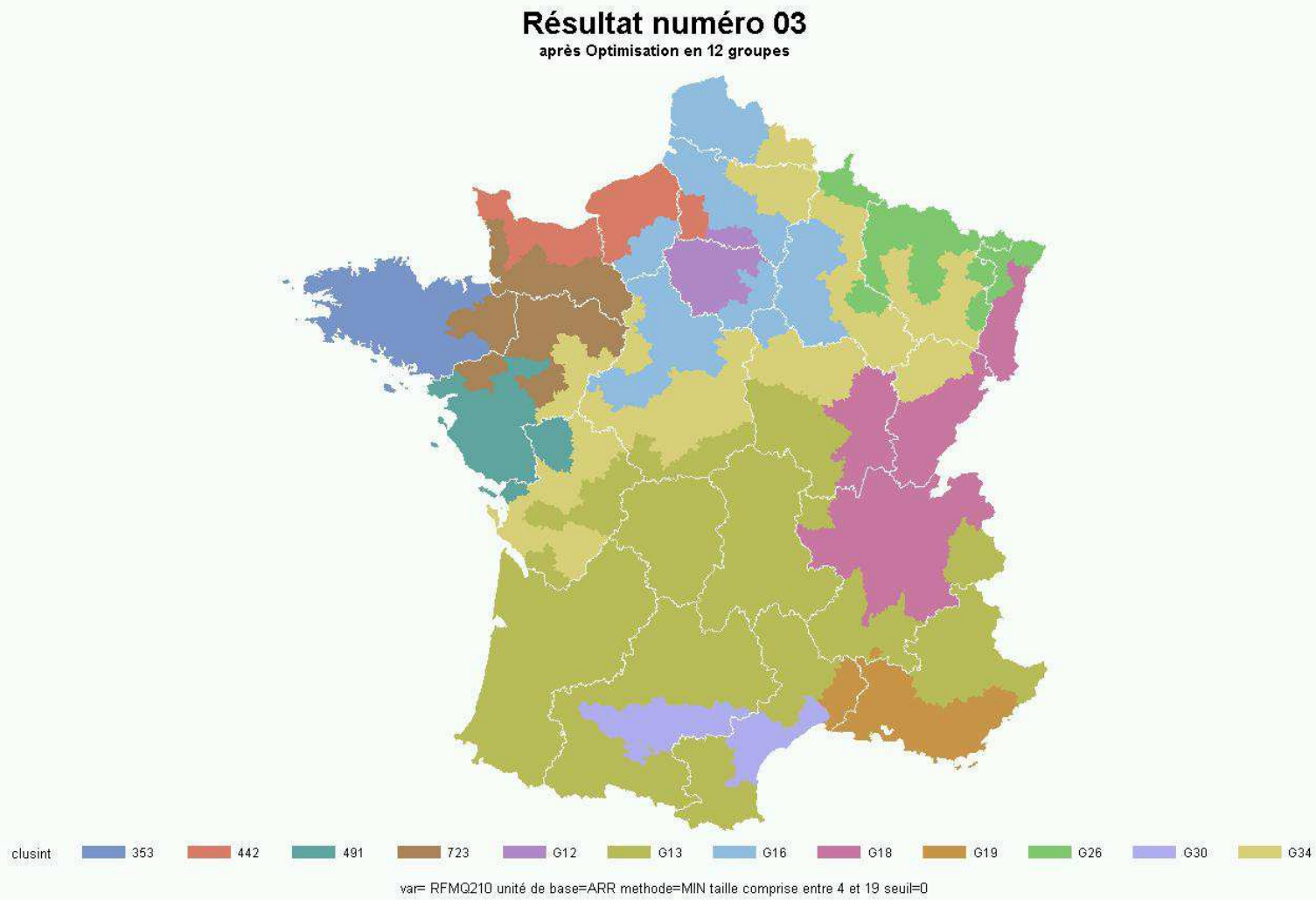
ARR – 0 – MIN

Indic.	POP	REV	POP&REV
In. Intra	0.41	0.56	0.41
T. Min	4.01% (332)	3.94% (353)	4.01% (G76)
T. Max	17.87% (G12)	18.81% (G12)	16.9% (G23)

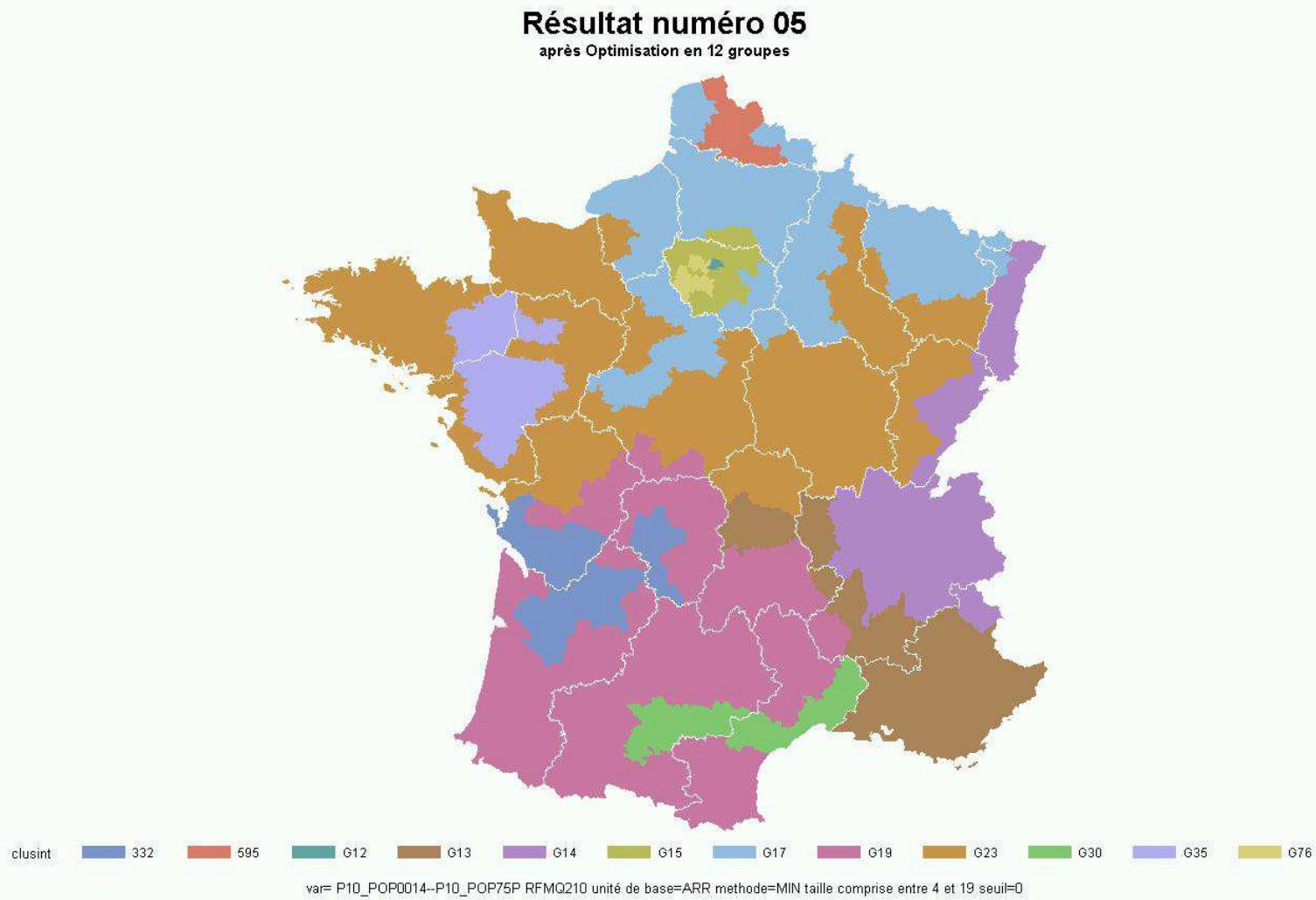
ARR – 0 – MIN – POP


Résultat numéro 02
après Optimisation en 12 groupes
clusint
332
595
751
G12
G13
G14
G16
G17
G18
G25
G28
G67
var= P10_POP0014--P10_POP75P unité de base=ARR methode=MIN taille comprise entre 4 et 19 seuil=0

ARR – 0 – MIN – REX


Résultat numéro 03
après Optimisation en 12 groupes
clusint
353
442
491
723
G12
G13
G16
G18
G19
G26
G30
G34
var= RFMQ210 unité de base=ARR methode=MIN taille comprise entre 4 et 19 seuil=0

ARR – 0 – MIN – POP & REV



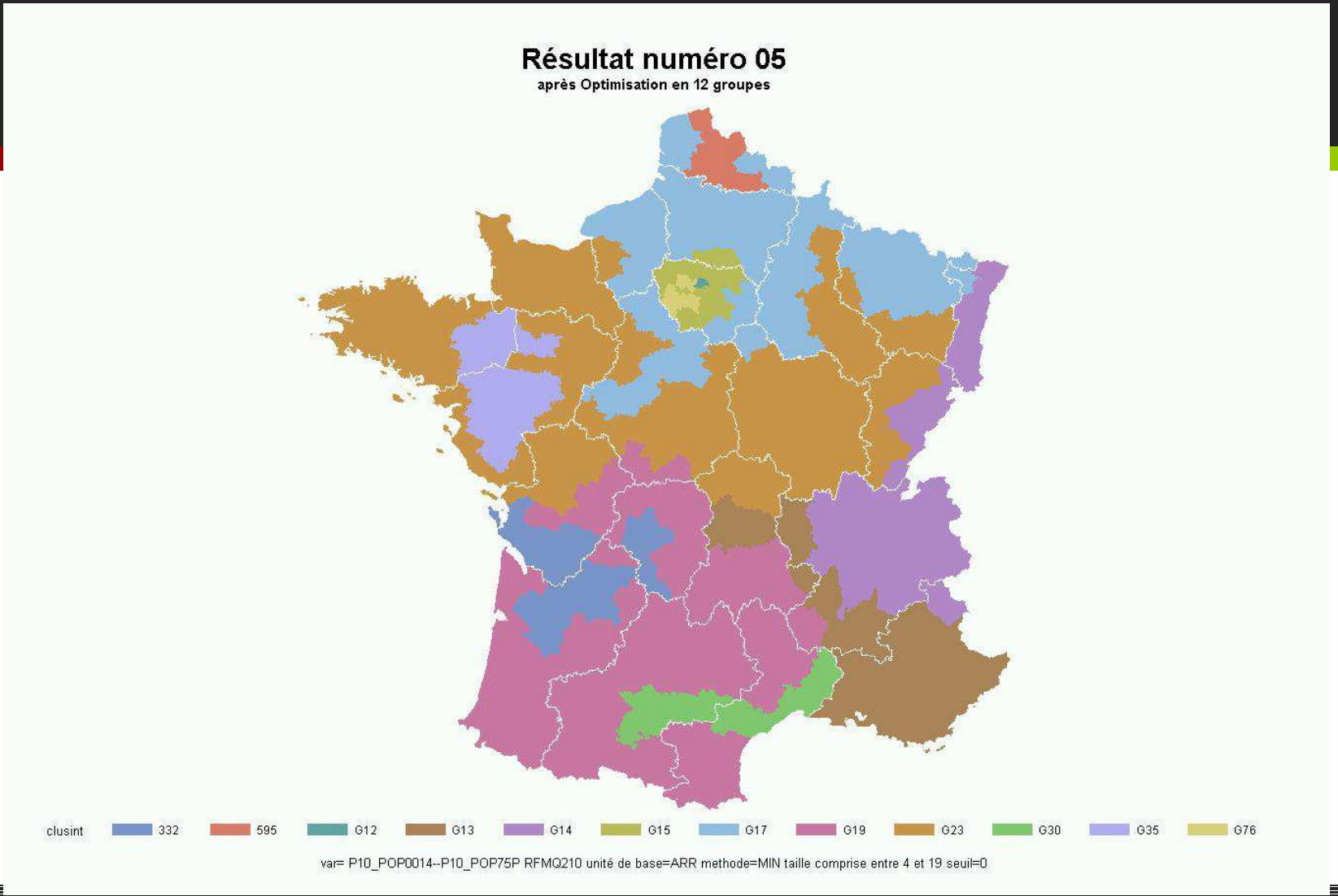
D/ Effet du seuil d'étendue

- Objectif : essayer de créer des classes « compactes ». On essaie de regrouper des unités ***seulement si elles s'inscrivent dans un rectangle dont la longueur de la diagonale est inférieure à un seuil donné (= seuil d'étendue géographique)***.
- Cette contrainte n'est prise en compte que lors de la première phase.
- Seuils testés : pas de seuil (0), 400 km et 500 km
- Variable = Population en 6 tranches d'âge et revenu, agrégats de base=ARR et minimisation de l'inertie intra.
- Situation ***après*** la seconde phase.

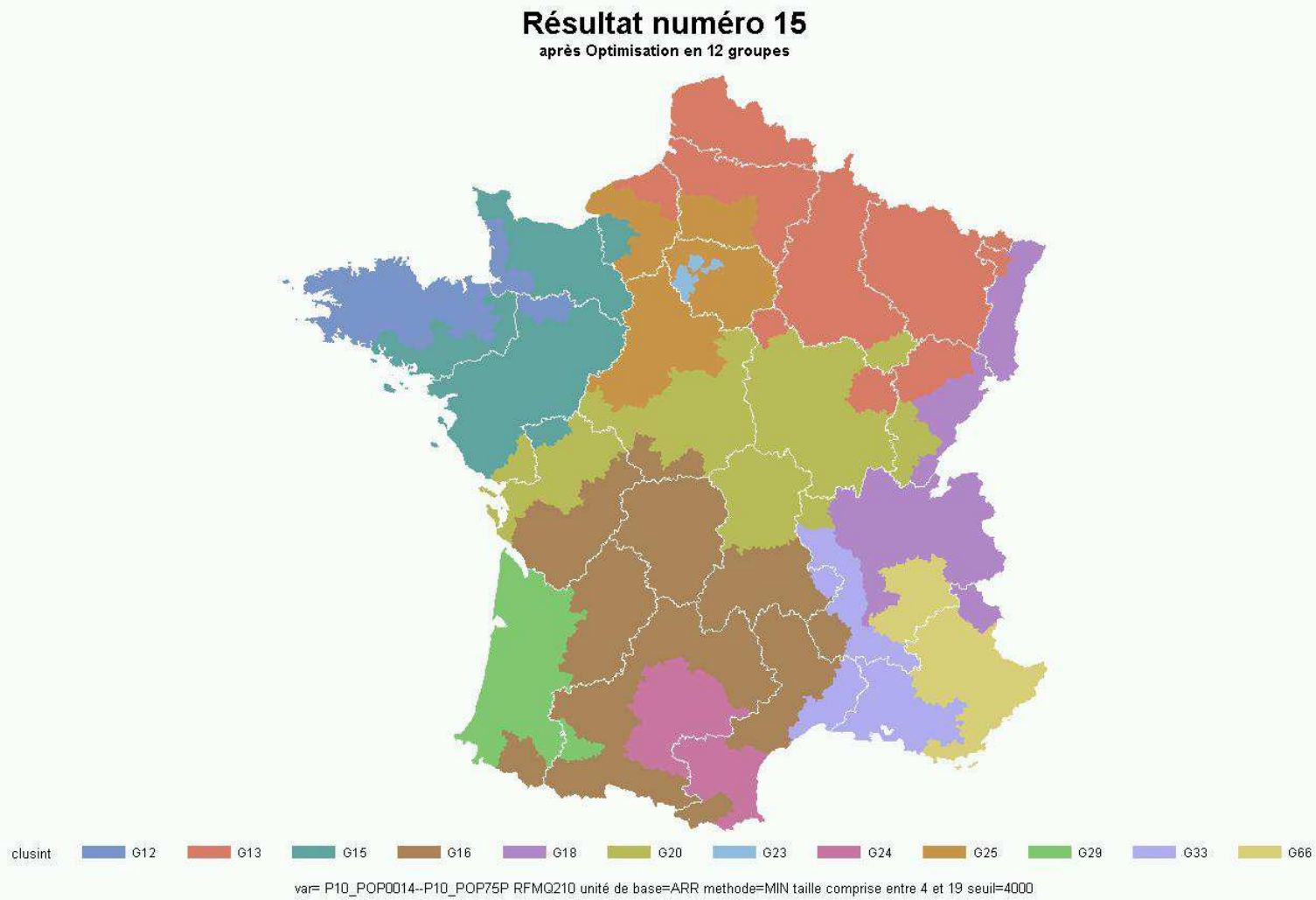
ARR – MIN – POP & REV

Indic.	0	400 km	500 km
In. Intra	0.41	0.55	0.56
T. Min	4.01% (G76)	4.07% (G24)	4.00% (G34)
T. Max	16.9% (G23)	15.74% (G13)	17.85% (G20)

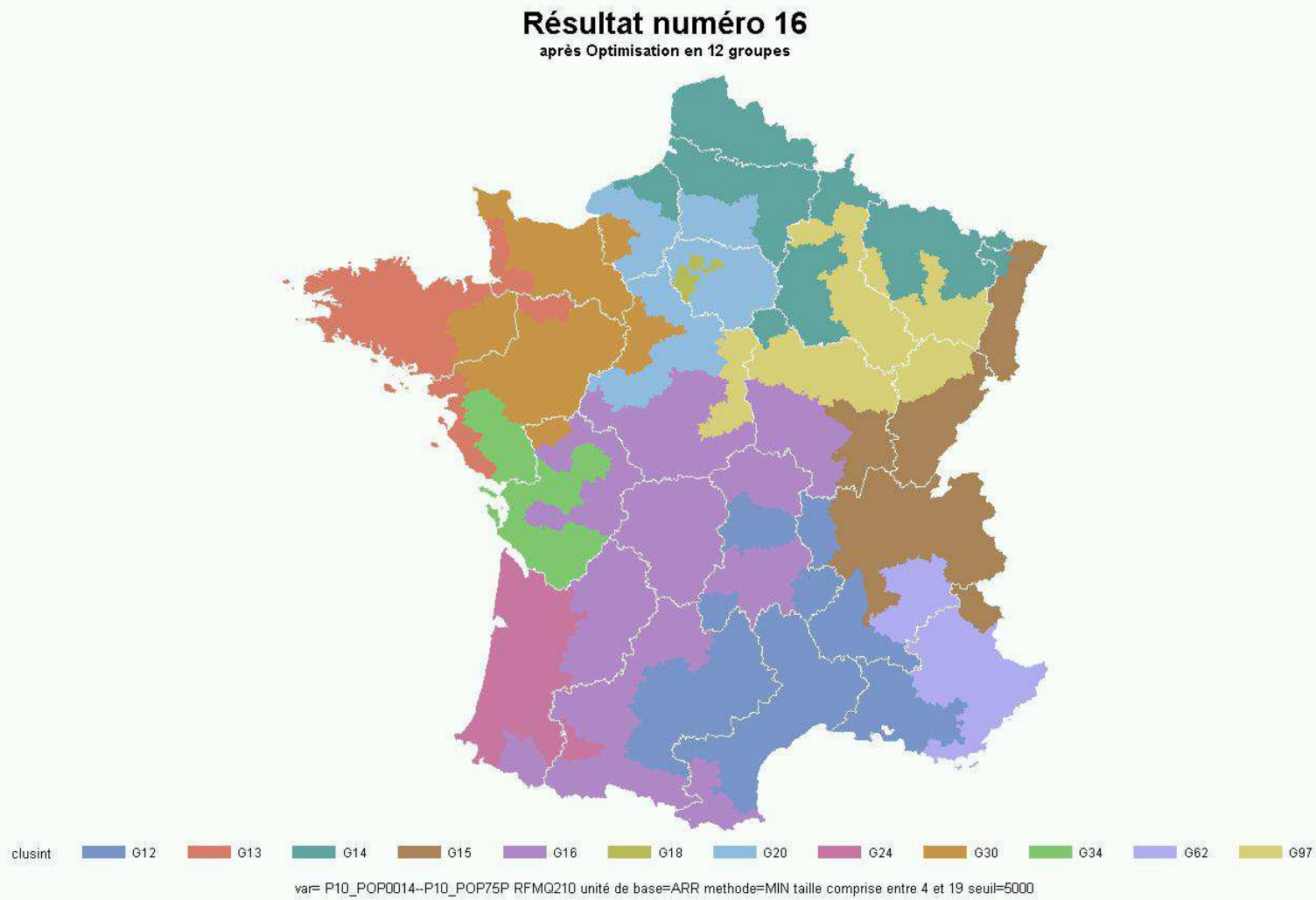
ARR – MIN – POP & REV – 0



ARR – MIN – POP & REV – 400 km



ARR – MIN – POP & REV – 500 km



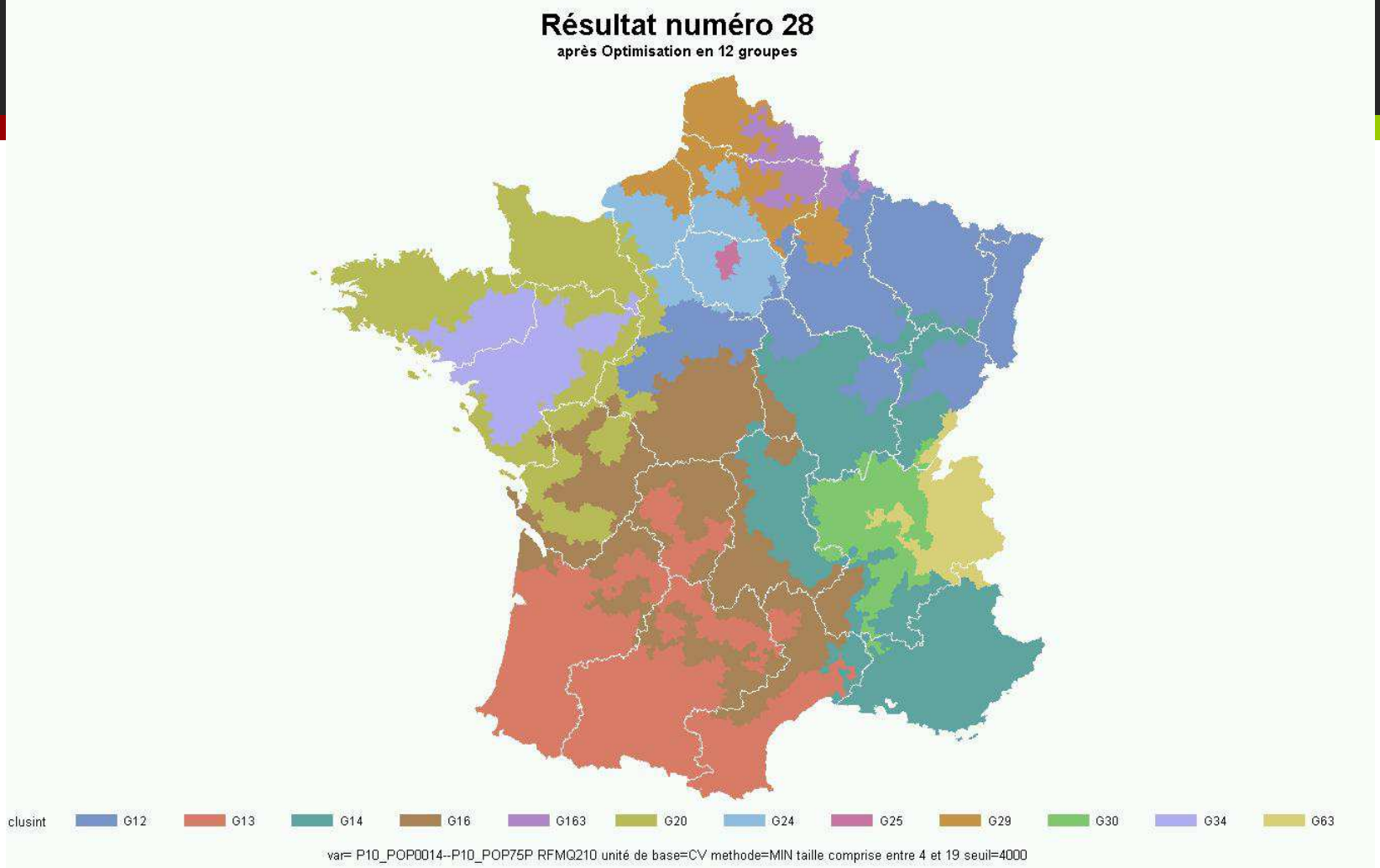
E/ Agrégats de base

- Différents résultats selon l'agrégat de base utilisé :
  - Cantons – villes
  - Arrondissement
  - Zone d'emploi
  - Département
  - Régions
- Minimisation de l'inertie intra, variables : Population en 6 tranches d'âge et revenu, seuil à 400 km
- **Attention : les inerties ne sont pas comparables.**
- Situation ***après*** la seconde phase.

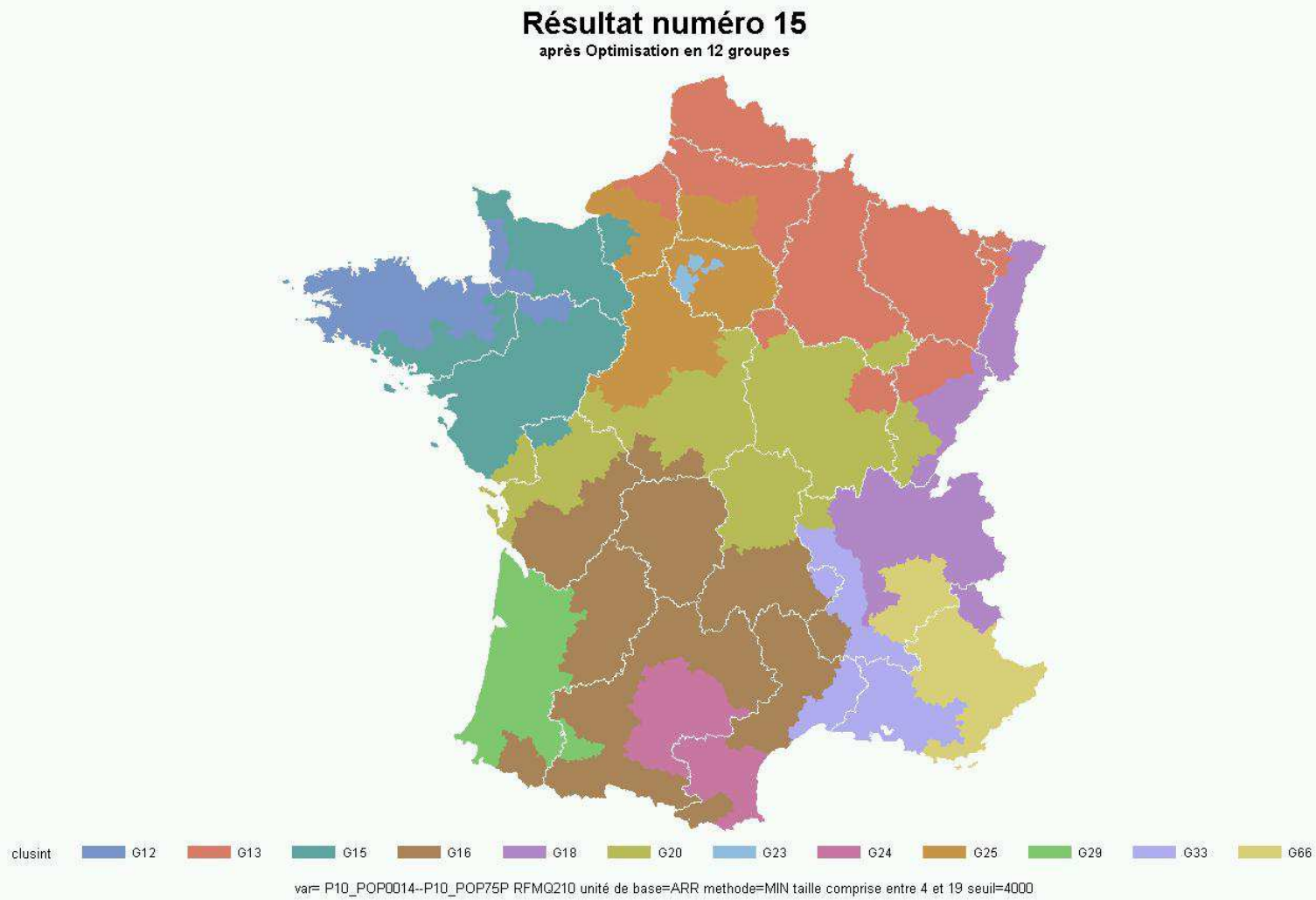
Agrégats de base

	In. Intra	T_Min	T_Max
Cantons	0.68	4.01	12.7
Zone emploi	0.35	4.10	14.2
Arr.	0.55	4.09	18.1
Dép.	0.34	3.53	18.9
Régions	0.02	4.22	18.9

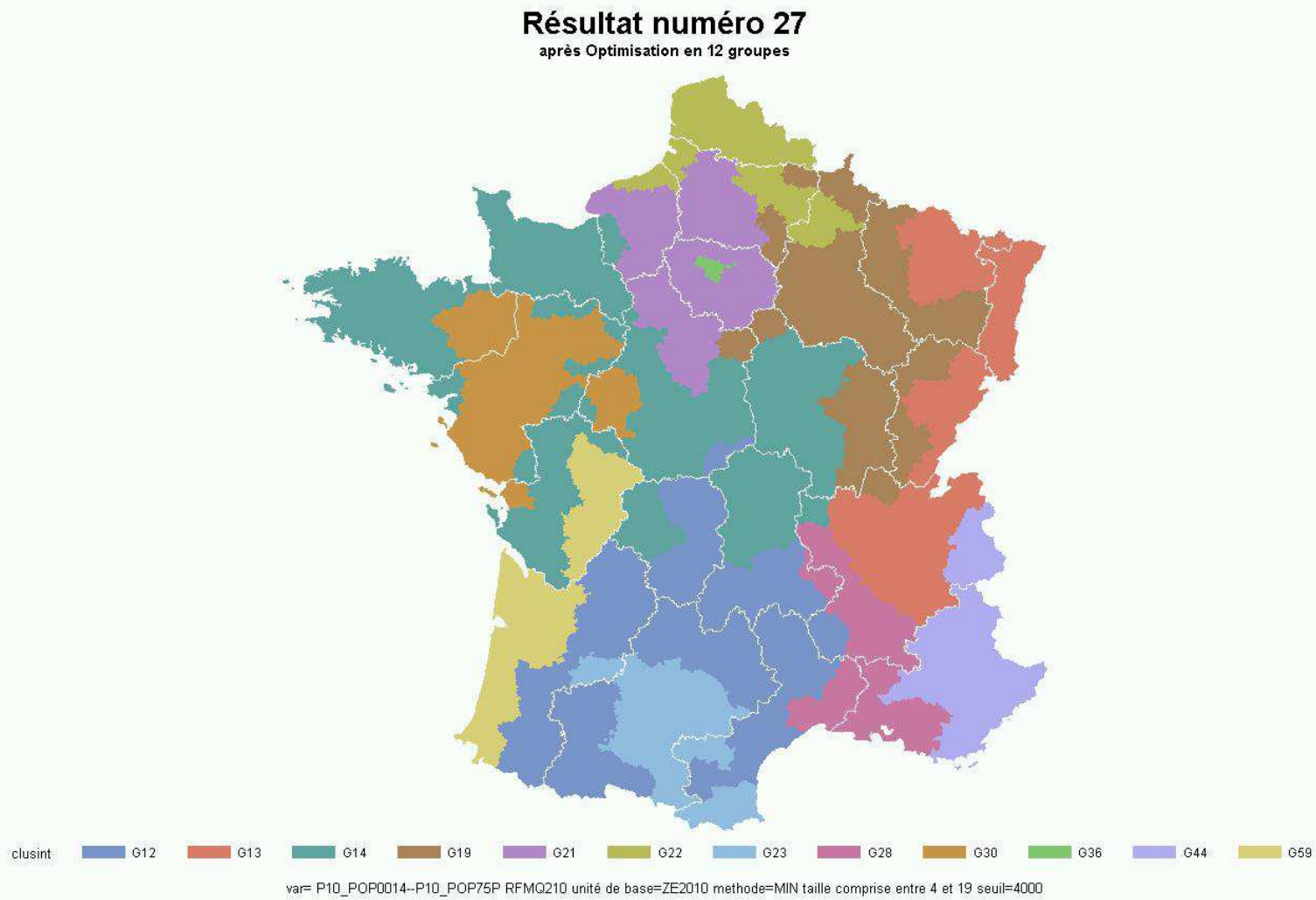
« Cantons » – MIN – POP & REV – 400 km



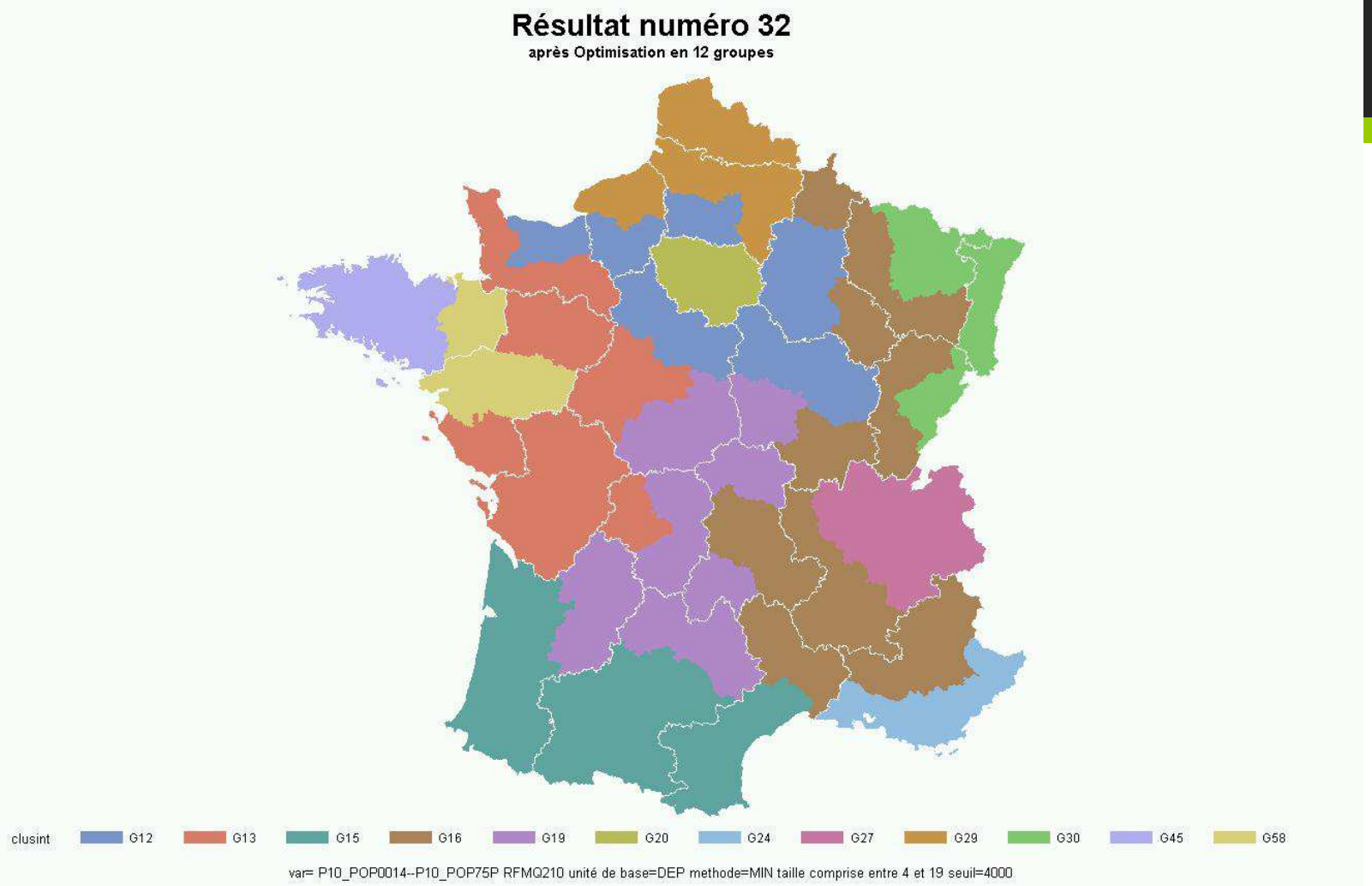
ARR –MIN – POP & REV – 400 km



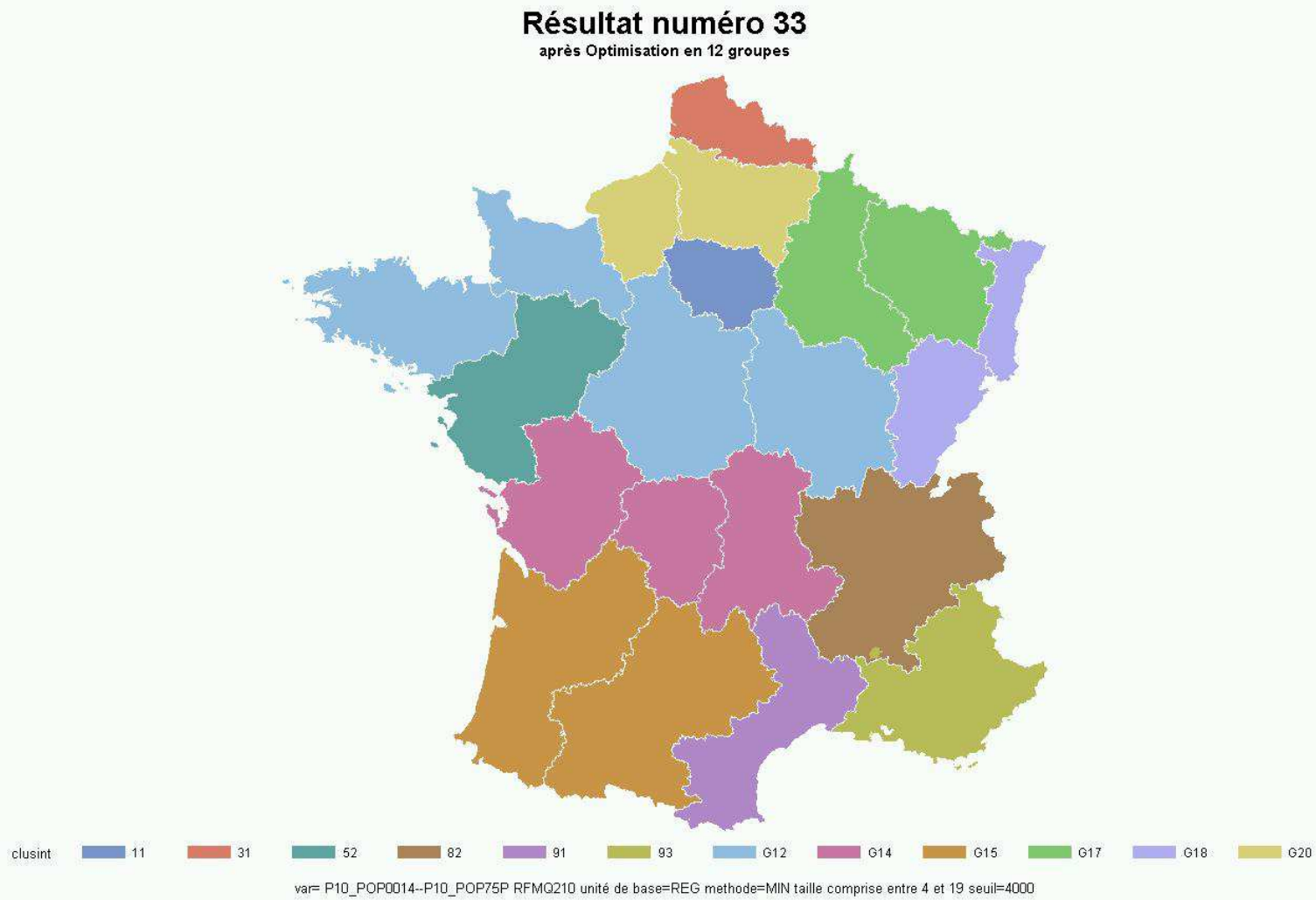
ZE – MIN – POP & REV – 400 km



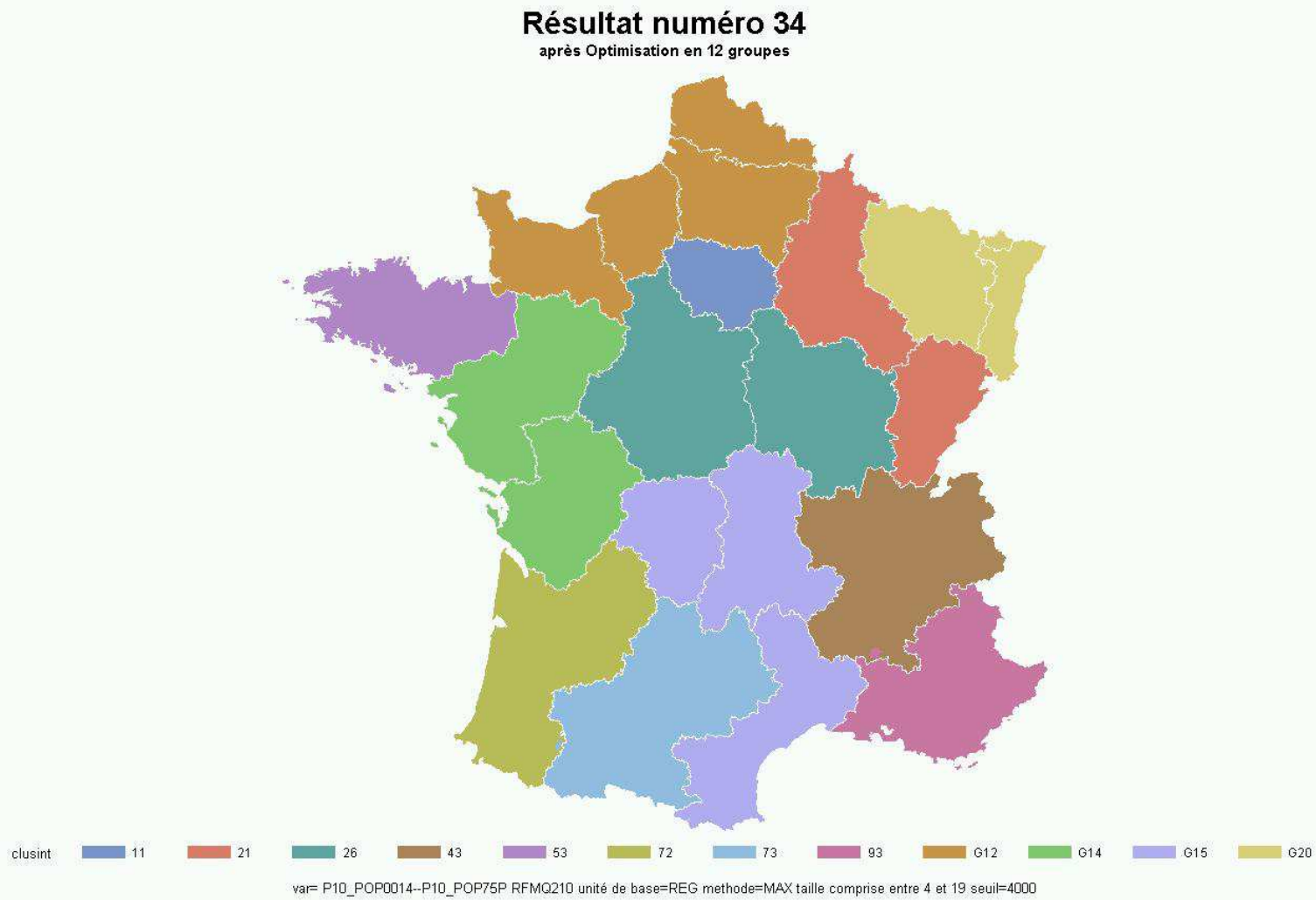
DEP – MIN – POP & REV – 400 km



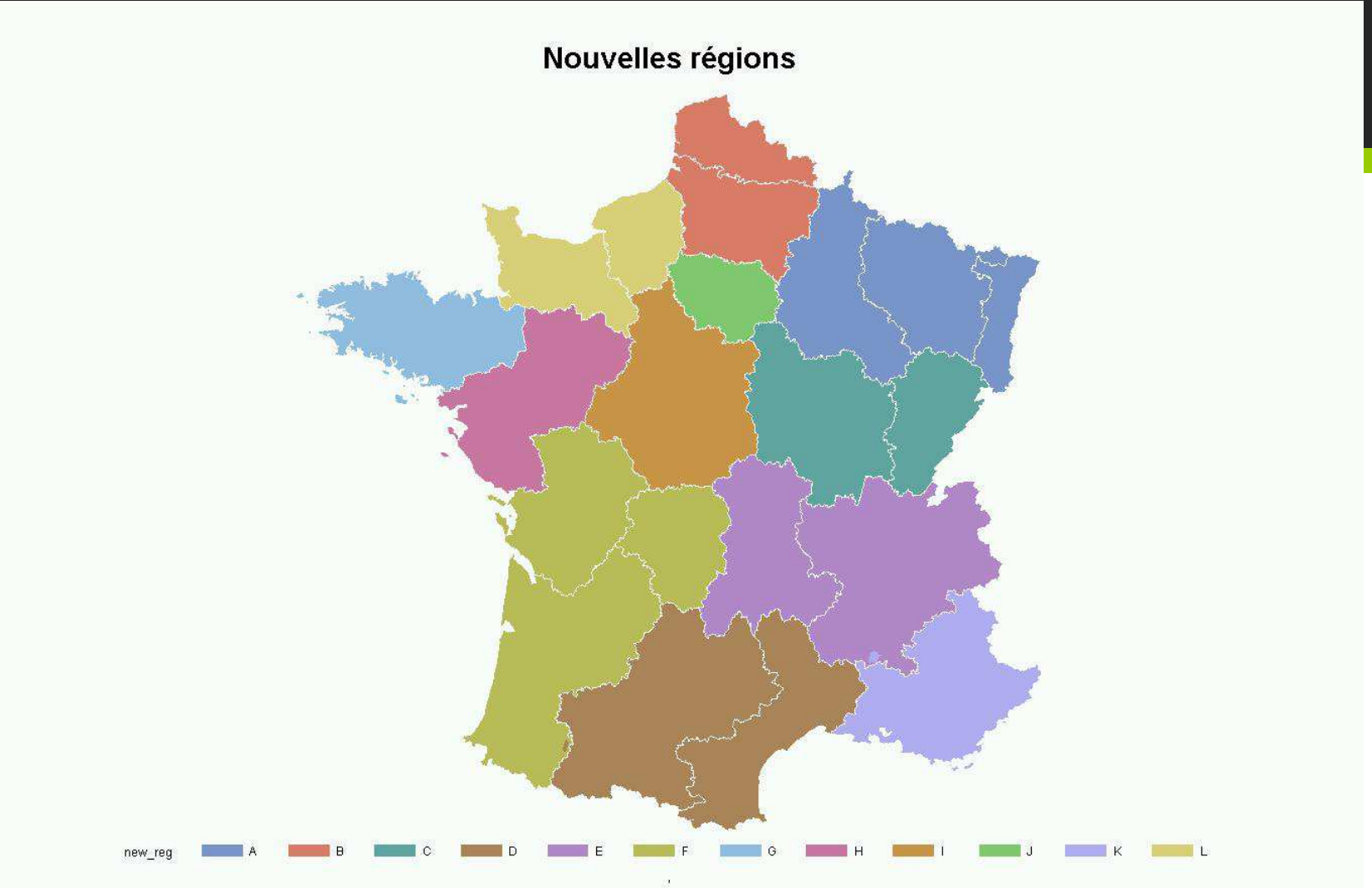
REG – MIN – POP & REV – 400 km



REG – MAX – POP & REV – 400 km



Nouvelles régions



Résultats pour les nouvelles régions

Résultat pour la variable « répartition de la population en tranches d'âge », limites de taille fixées à 4 et 19.

Agrégats de base	Inertie intra résultant de la méthode MIN	Inertie intra résultant des nouvelles régions	Inertie intra résultant de la méthode MAX
Région	0.0165772	0.1183861	0.2087198 (*)
Département	0.2923304	0.5055625	0.8271314 (*)
Arrondissements	0.4097203	0.6453773	0.8612547

Résultat pour la variable revenu médian (considérée comme une moyenne de revenu). limites de taille fixées à 4 et 19.

Agrégats de base	Inertie intra résultant de la méthode MIN	Inertie intra résultant des nouvelles régions	Inertie intra résultant de la méthode MAX
Région	0.039418	0.0795469	0.1139917
Département	0.1883157	0.5045428	0.8119782
Arrondissements	0.3850081(*)	0.7085313	0.8742114(*)

(*) : il existe un agrégat plus petit que la plus petite région créée par la loi.

7. Conclusion

- Ces simulations sont un exercice d'école, qui n'a testé que certaines hypothèses
- Plusieurs difficultés
  - Algorithmes complexes
  - Temps de calcul augmentant avec le nombre d'unités à traiter
  - Contraintes incompatibles
  - Solutions instables
- Voies de progrès
- ***Le politique a-t-il besoin de la statistique ? (et vice-versa…)***

=> Tout dépend de la formalisation des objectifs poursuivis !

Nota : Une macro SAS est disponible sur demande

Merci de votre attention

marc.christine@insee.fr

michel.isnard@insee.fr