Problèmes théoriques et pratiques de la construction de l'EMEX.

Marc CHRISTINE, INSEE (UMS)

Laurent WILMS, INSEE (RRP)

EMEX = Echantillon-Maître pour les Extensions régionales d'enquêtes nationales.

1. Objectifs de l'EMEX.

- Répondre à des *demandes régionales…*
- … pour lesquelles l'échantillon-maître (EM) est insuffisant.
- Construire une *offre cohérente* : standardiser et automatiser le tirage d'extensions.
- Possibilité *d'exploiter simultanément* les données issues des deux échantillons.

Première application :

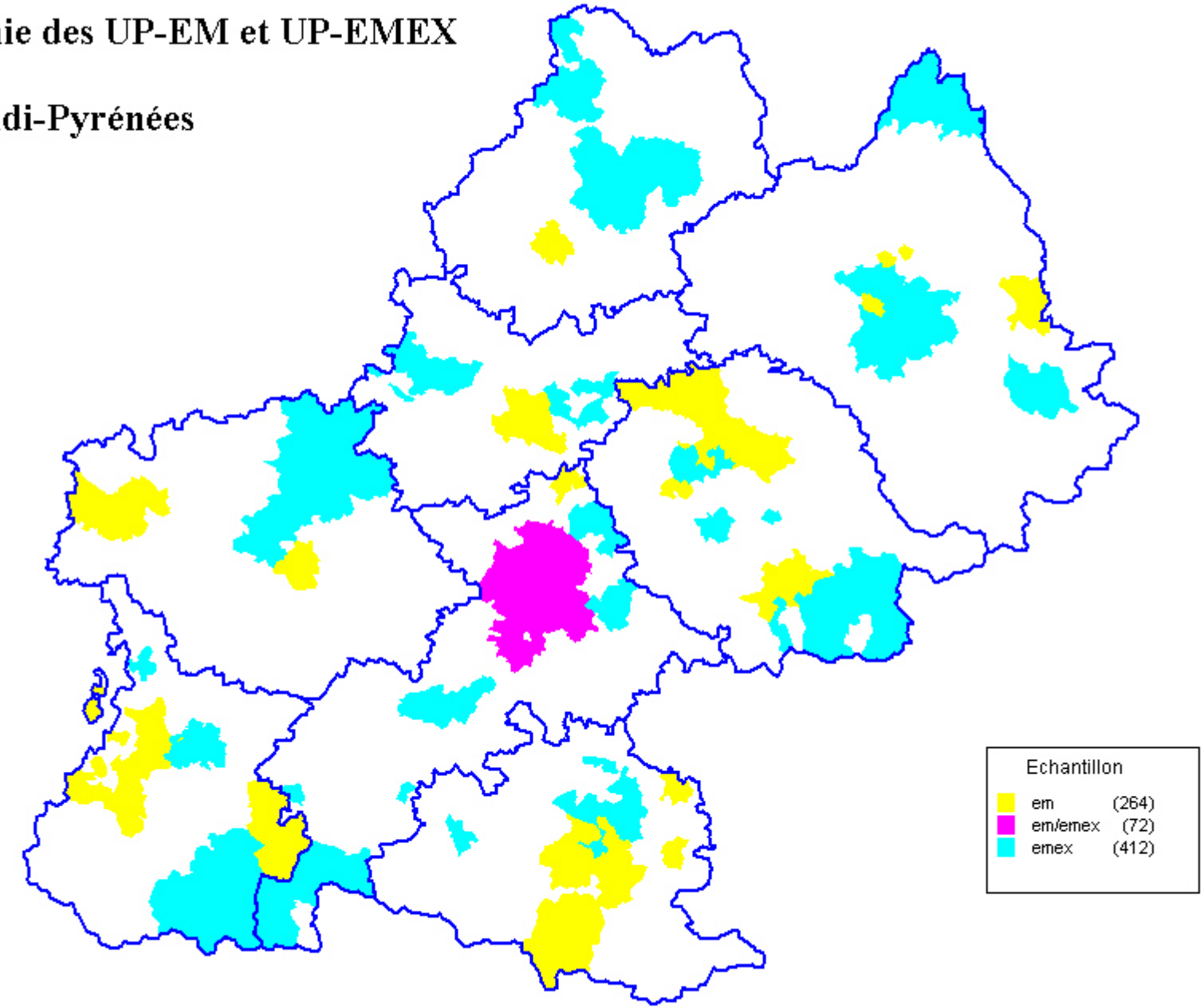
- L'enquête Santé 2002
  (5 régions concernées).

2. La problématique de l'EMEX.

- Il faut tirer un complément à l'EM (en dehors de celui-ci).
- Mais on ne peut le tirer « n'importe comment » :
  - Contraintes d'organisation et de coût
    => concentrer les zones d'enquête
    => tirage d'unités primaires
  - Assurer une « représentativité » régionale de l'ensemble EM + EMEX.

Cartographie des UP-EM et UP-EMEX

Midi-Pyrénées



- La difficulté principale est d'atteindre la représentativité régionale de l'ensemble EM + EMEX (que l'on va traduire par la notion d'*équilibrage*)…
- … alors que l'on ne peut jouer que sur le tirage de l'EMEX…
- … et que l'EM a déjà été tiré une fois pour toutes (avec équilibrage super-régional).

3. Nécessité d'une approche théorique.

- Le tirage de l'EMEX se fait *conditionnellement* au tirage de l'EM..
- … ce qui va conduire à la notion de probabilité d'inclusion conditionnelle
- L'ensemble EM + EMEX doit permettre de construire des estimateurs sans biais… Π_i^{2/S_1}
- … et doit vérifier des conditions d'équilibrage au niveau régional.

Une équation d'équilibrage s'écrira sous la forme :

$$\hat{T} = T$$

- où T est un total connu sur la population
- et $\hat{T}$ est un estimateur sans biais de ce même total sur l'échantillon.

On cherche à résoudre ce problème en partant d'une classe très générale d'estimateurs :

$$\hat{T} = \sum_{i \in S_1} a_i(S_1) Y_i + \sum_{j \in S_2} b_j(S_1) Y_j$$

- Les *coefficients aléatoires* sont obtenus en écrivant des conditions d'absence de biais pour l'estimation de ***tout*** total.

- On reconnaît comme cas particulier dans cette famille les estimateurs à *coefficients fixes* : HORVITZ-THOMSON.

$$\hat{T} = \alpha \sum_{i \in S_1} \frac{Y_i}{\Pi_i^1} + (1-\alpha) \sum_{i \in S_2} \frac{Y_i}{\Pi_i^2}$$

- Ou :

$$\hat{T} = \sum_{i \in S_1 \cup S_2} \frac{Y_i}{\Pi_i}$$

Le dilemme de l'équilibrage.

- L'équation $\hat{T} = T$

  va se réécrire :

$$\sum_{j \in S_2} b_j(S_1) Y_j = T - \sum_{i \in S_1} a_i(S_1) Y_i$$

- Lorsque l'on tire un échantillon S_2 selon une loi de tirage conditionnelle à la réalisation du 1er échantillon, l'équation précédente s'interprète comme une *relation d'équilibrage inadéquate*…

$$\sum_{j \in S_2} b_j(S_1) Y_j = T - \sum_{i \in S_1} a_i(S_1) Y_i$$

- ….portant sur un total qui, en général :
- *n'est ni le total des variables d'équilibrage sur la population dans laquelle on tire, …*
- *… ni l'espérance, vis-à-vis de la loi conditionnelle de tirage du 2ème échantillon, de l'estimateur d'un total construit à partir de ce 2ème échantillon.*

- En effet, pour être adaptée au tirage conditionnel, cette condition devrait s'écrire :

$$\sum_{i \in S_2} \frac{Z_i}{\Pi_i^{2/S_i}} = \sum_{i \in P - S_1} Z_i$$

- => La méthode du tirage équilibré (au moyen de CUBE) n'est en général pas utilisable dans ce cadre.

4. Contextes particuliers où la propriété d'équilibrage peut être atteinte.

4.1 Tirages successifs équilibrés et à probabilités égales.

- Quand le 1er échantillon tiré dans ***P*** est :
  - équilibré au niveau de ***P*** sur une variable Z
  - à probabilités égales
- Et quand le 2ème échantillon, tiré conditionnellement à S_1, vérifie les mêmes propriétés sur $\boldsymbol{P} \setminus S_1$
- Alors l'estimateur H-T relatif à $S_1 \cup S_2$ est bien équilibré sur Z au niveau de ***P***.

4.2 Estimateur de Horvitz-Thomson conditionnel à S_1

- L'estimateur suivant :

$$\hat{T}_1 = \sum_{i \in S_1} Z_i + \sum_{i \in S_2} \frac{Z_i}{\Pi_i^{2/S_1}}$$

- est sans biais et possède la propriété d'équilibrage sur Z au niveau de ***P***, si S_2 est sélectionné dans ***P*** \ S_1 selon un tirage équilibré sur Z.

4.3. Une voie possible : l'équilibrage « inverse »

- Lorsque la propriété d'équilibrage ne peut pas être atteinte sur l'estimateur de Horvitz-Thomson global (ce qui est en général le cas pour une loi de tirage quelconque),…

- …on peut chercher à construire une nouvelle loi de tirage qui conduise à un estimateur d'Horvitz-Thomson, cette fois-ci équilibré.

Formalisation du problème.

$$\underset{\tilde{\Pi}_i}{Min} \sum_{i \in P} d(\tilde{\Pi}_i, \Pi_i)$$

Sous la contrainte (C) :

$$\sum_{i \in s_1} \frac{x_i}{\tilde{\Pi}_i} + \sum_{i \in s_2} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i \in P} x_i \quad \forall s_2$$

où, dans (C), s_1 est fixé et s_2 est aléatoire.

Résolution du problème.

- L'équation (C) est équivalente à

$$\sum_{i \in S_2} \frac{z_i}{\widetilde{\Pi}_i^{2/S_1}} = \sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\widetilde{\Pi}_i} \quad \forall s_2 \quad \text{(C1):}$$

en posant : $$z_i = x_i \frac{\widetilde{\Pi}_i^{2/S_1}}{\widetilde{\Pi}_i} \quad \forall i \in P \setminus s1$$

- L'astuce est de convertir (C1) en une véritable équation d'équilibrage sur z :

$$\sum_{i \in S_2} \frac{z_i}{\widetilde{\prod}_i^{2/S_1}} = \sum_{i \in P \setminus S_1} z_i \qquad \forall S_2$$

… Car on sait tirer des S_2 dans $P \setminus S_1$ qui vérifient ce type d'équation !

- On va donc chercher les $\tilde{\Pi}_i$ assurant l'égalité

$$\sum_{i\in P} x_i - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i\in P-S_1} z_i$$

- ce qui conduit en définitive à résoudre :

$$\sum_{i\in P} x_i - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i\in P-S_1} x_i \frac{1-\frac{\Pi_i^1}{\tilde{\Pi}_i}}{1-\Pi_i^1}$$

Remarques sur cette méthode.

- On peut équilibrer sur plusieurs variables.
- Les équations de contrainte n'admettent pas nécessairement de solutions admissibles, c'est-à-dire telles que :

$$\Pi_i^1 \leq \tilde{\Pi}_i \leq 1 \quad \forall i \in P - S_1$$

Dans ce cas, il faut relâcher des contraintes.

- Il faut éventuellement disposer d'un critère de jugement sur l'éloignement de la loi de tirage solution par rapport à la loi initiale.

5. Tirage effectif des UP rurales et des UU de –20 000 hab.

- Dans quelques régions, on compare deux types de tirage d'UP pour l'EMEX :
  - (1) en application de la méthode de l'équilibrage inverse, un tirage régional de taille fixe, équilibré au niveau régional sur 5 variables ;
  - (2) un tirage régional systématique, avec tri des UP par département et probabilités finales d'inclusion de celles-ci proportionnelles au nombre de résidences principales.

Avantages-Inconvénients.

- Le tirage équilibré conduit à des estimateurs d'H-T plus précis pour des variables d'intérêt bien corrélées aux variables d'équilibrage

(résultats de simulations sur 17 variables).

Avantages-Inconvénients.

Région	Nombre de simulations	Variables d'équilibrage	EQM cube	EQM syst.	Biais relatif Cube	Biais relatif Systématique
11	100	1	0.22	0.24	0.06	0.12
21	100	5	0.09	0.56	0.03	0.48
23	100	3	0.09	0.16	0.02	0.11
24	100	5	0.06	0.18	0.02	0.14
25	100	3	0.04	0.22	0.008	0.14
82	100	5	0.08	0.43	0.03	0.27

Avantages-Inconvénients.

- Le tirage (2) n'est pas équilibré, mais les départements seront mieux représentés.
- Le tirage (2) assure une charge par UP constante.
- Le tirage (1), sur certaines régions, induit des charges par UP trop variables

  ($\tilde{\Pi}_i$ trop éloignés des Π_i)

- => On retient finalement le systématique pour le tirage effectif.

6. Tirage des districts.

- En strates de gestion 2, 3 et 4, les districts-EM avaient été tirés dans chaque UU avec des *probabilités égales* et un équilibrage sur 5 variables au niveau de chaque UU.
- Le tirage des districts-EMEX, conditionnellement aux districts-EM, est aussi à probabilités égales et équilibré sur les mêmes variables.

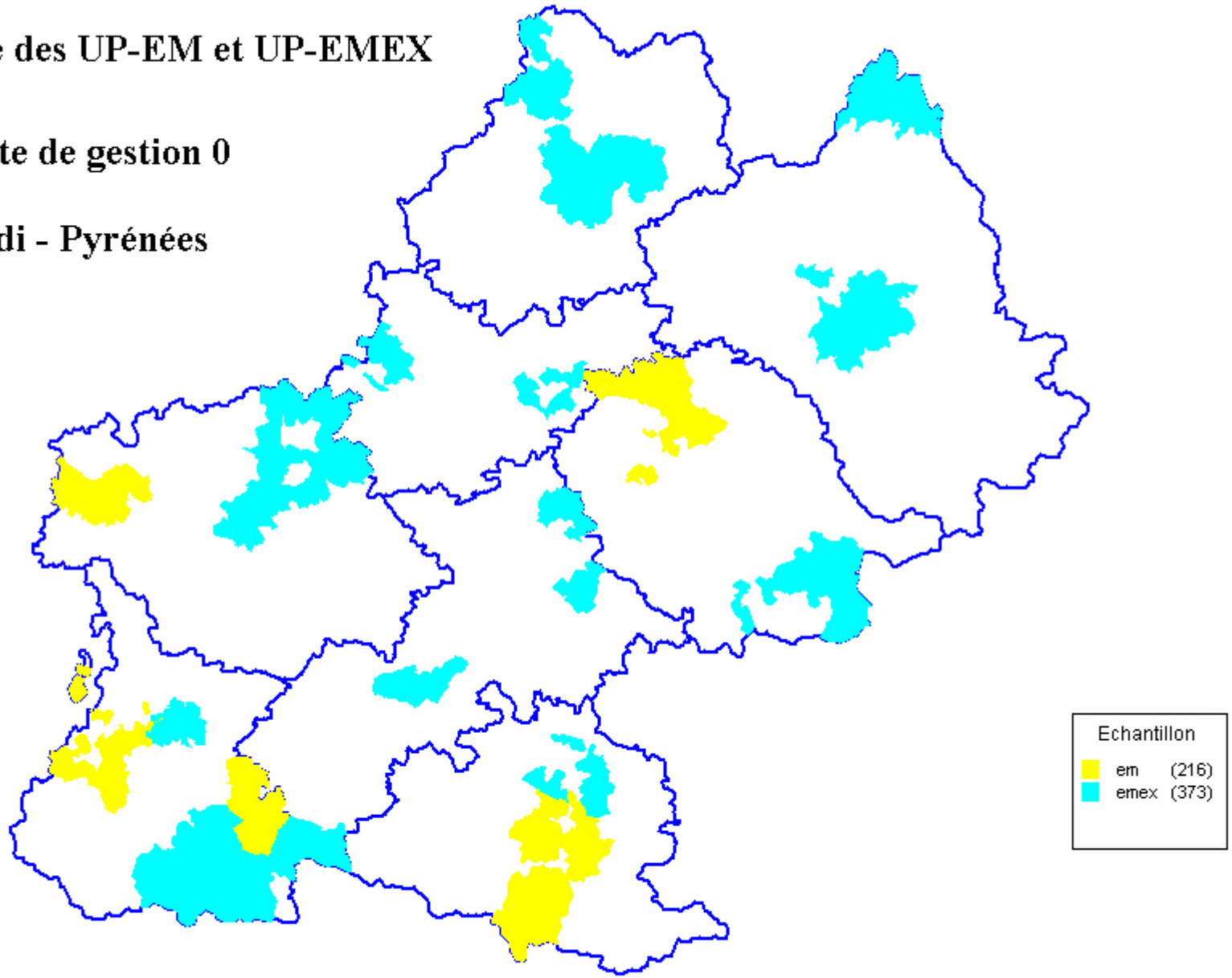
- Alors l'échantillon de l'ensemble des districts EM et EMEX est aussi équilibré au niveau de chaque UU, sur chacune des variables d'équilibrage.

Cartographie de l'EM et de l'EMEX par strate de gestion.

Cartographie des UP-EM et UP-EMEX

Strate de gestion 0

Midi - Pyrénées

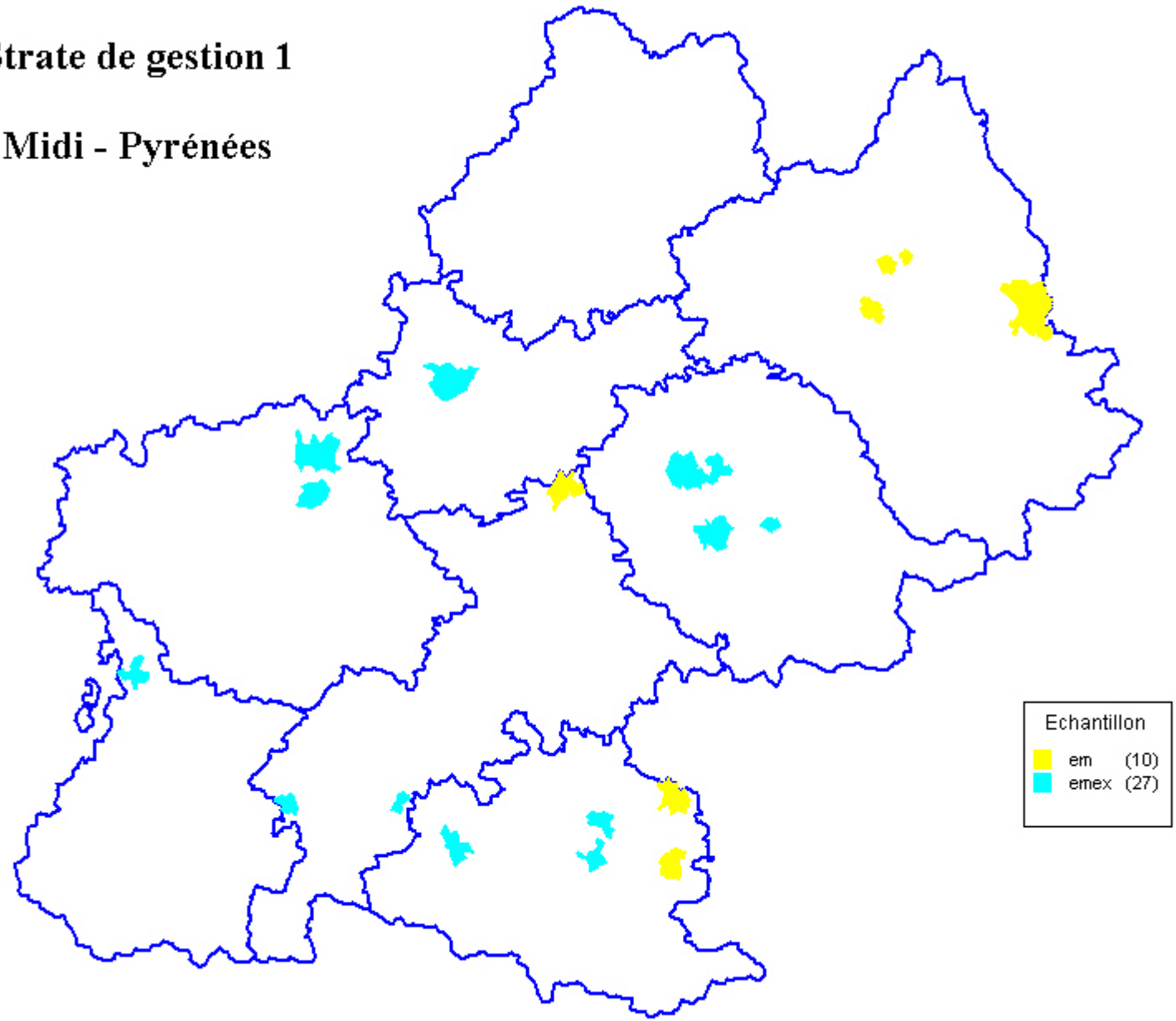


- Dimensionnement adapté à la taille moyenne des extensions…
- … visant à doubler le taux de sondage moyen d'une enquête standard.

Cartographie des UP-EM et UP-EMEX

Strate de gestion 1

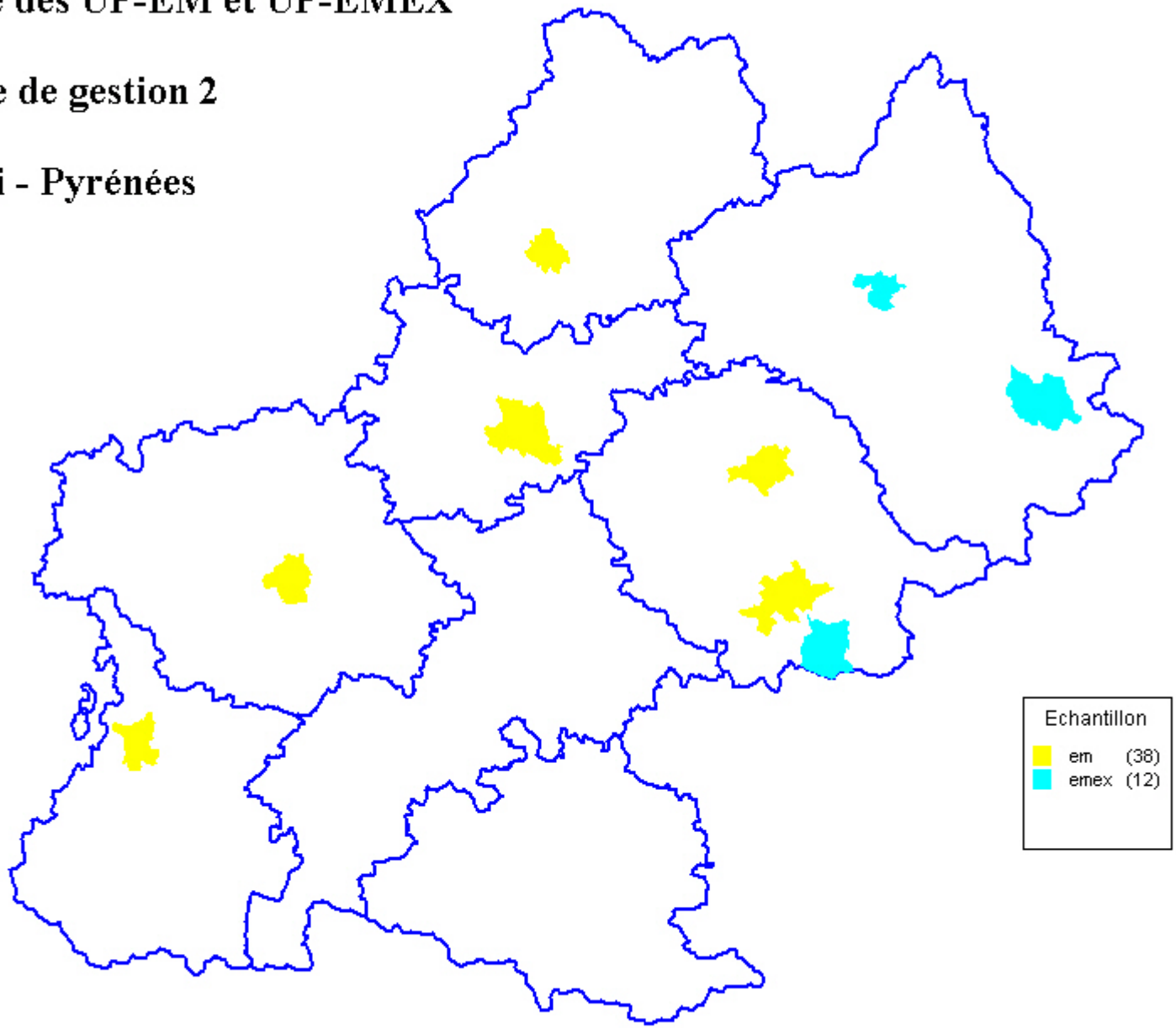
Midi - Pyrénées



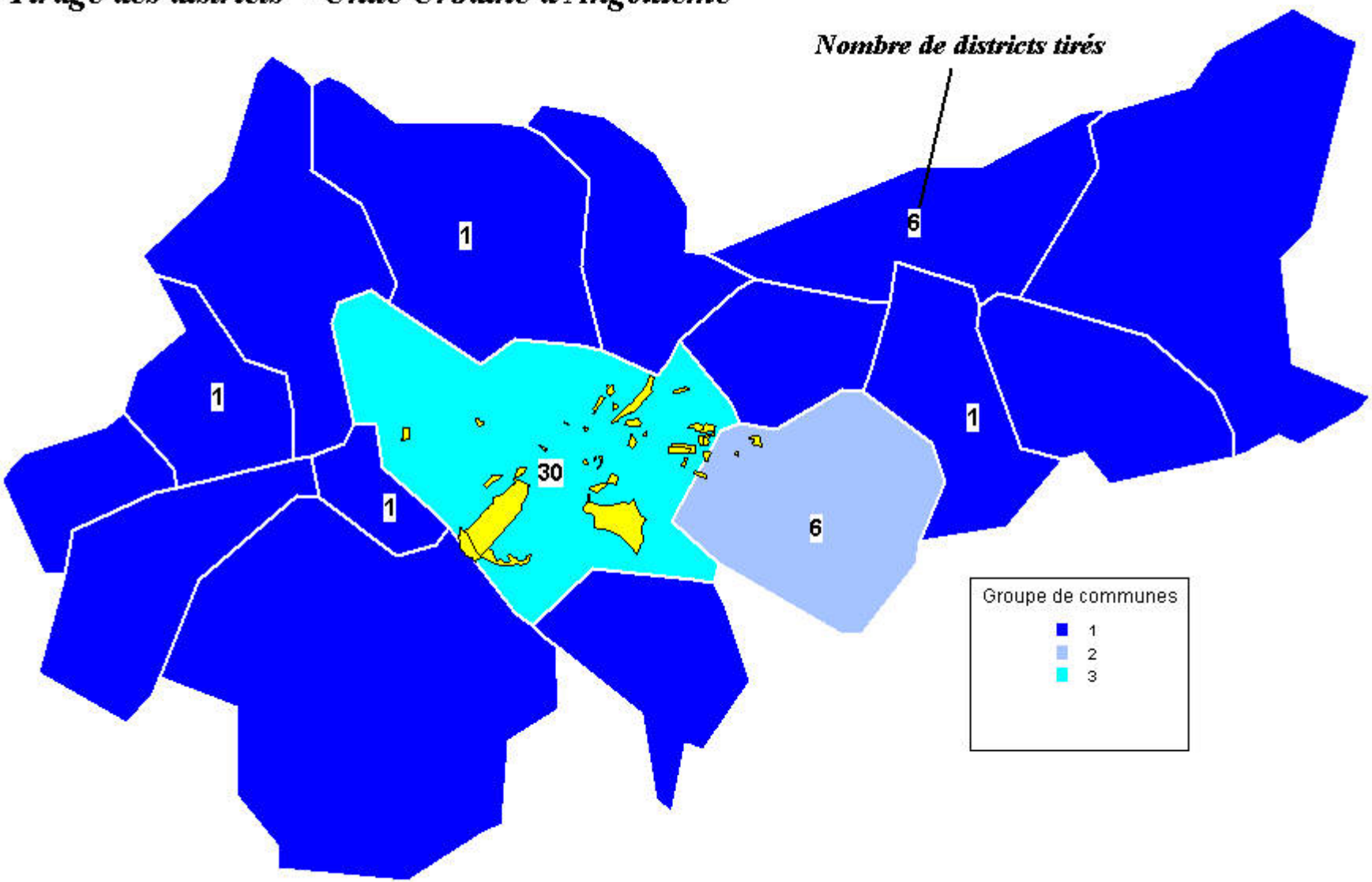
Cartographie des UP-EM et UP-EMEX

Strate de gestion 2

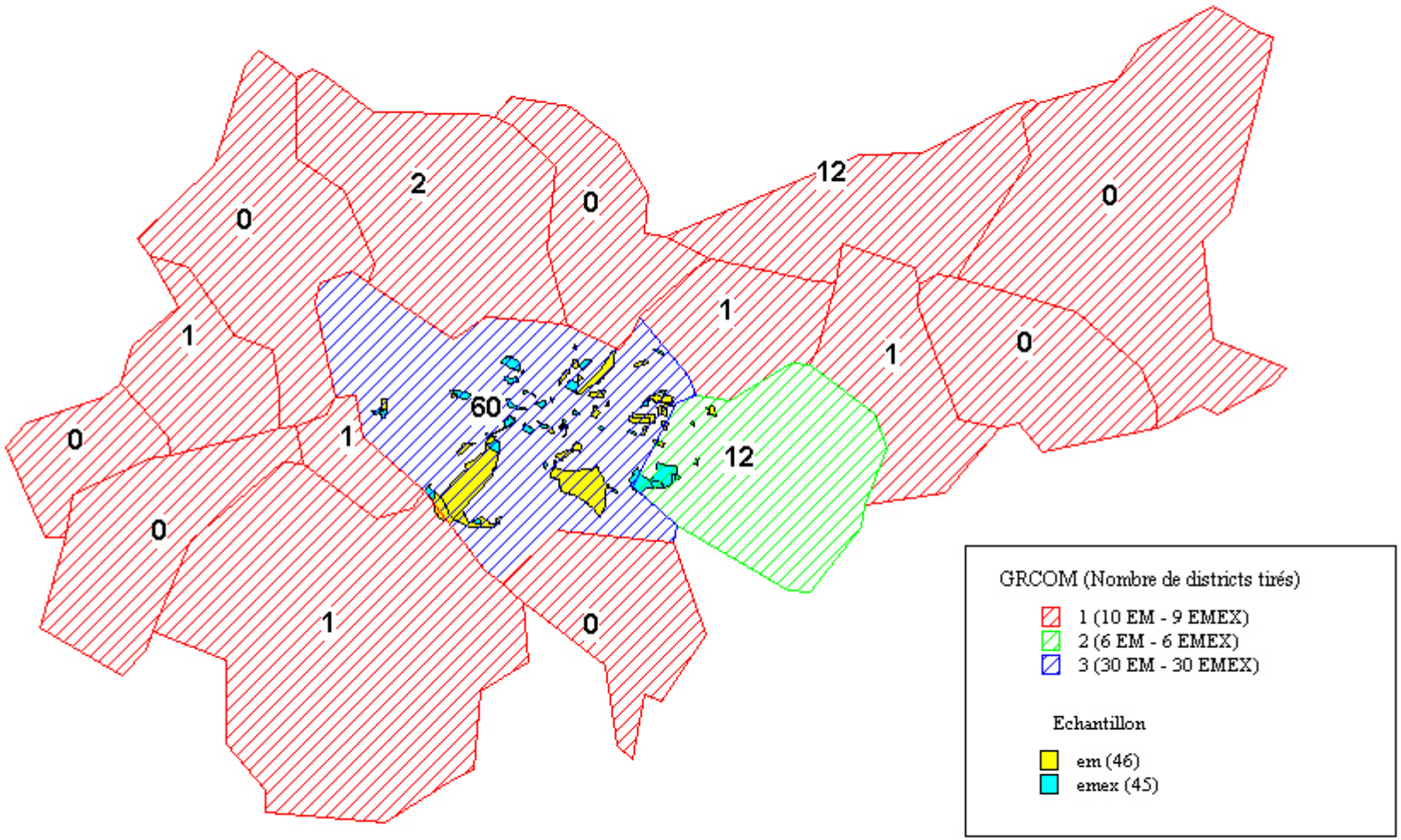
Midi - Pyrénées



Tirage des districts - Unité Urbaine d'Angoulème

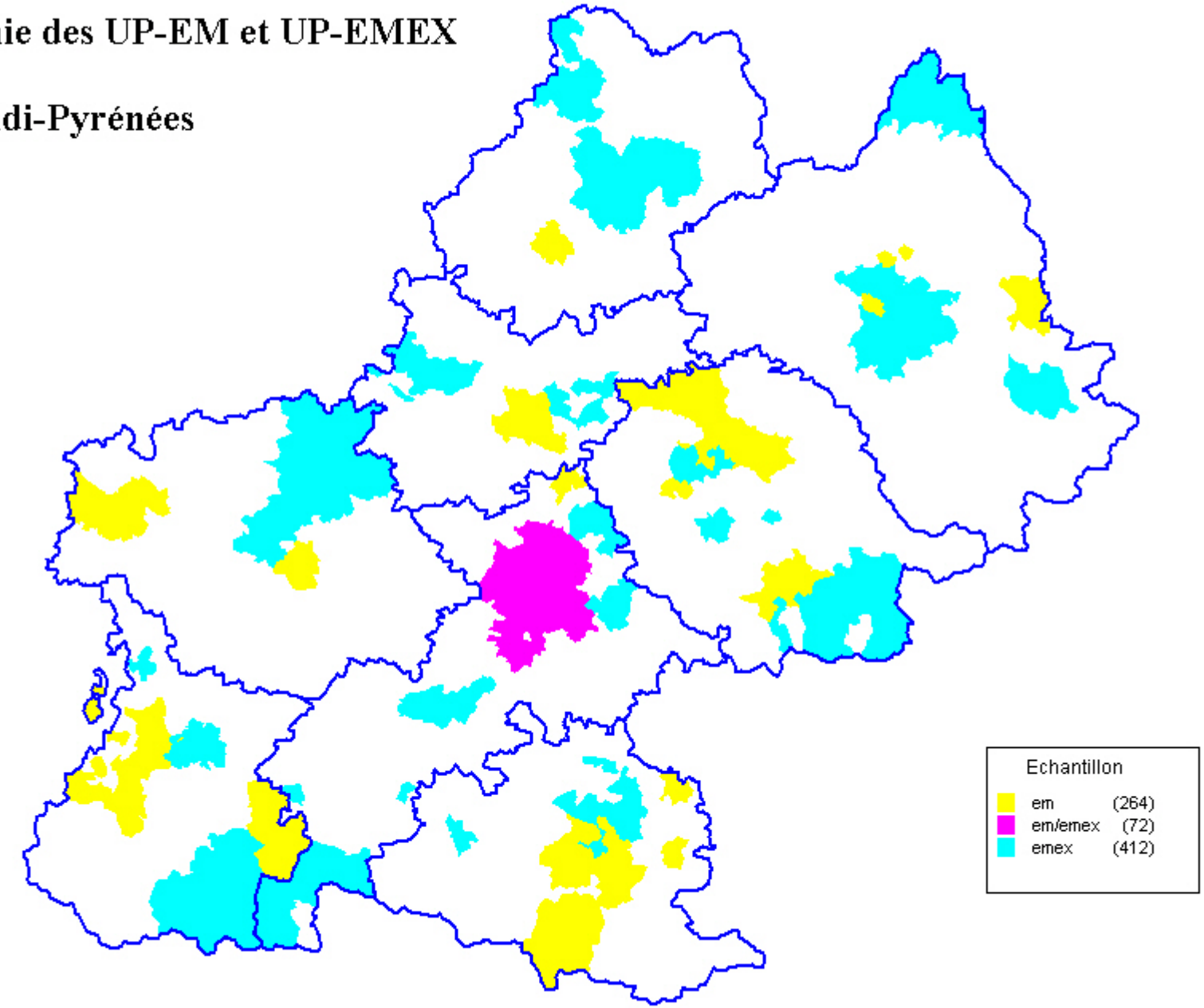


Répartition des districts EM et EMEX par commune
Unité urbaine d'Angoulême

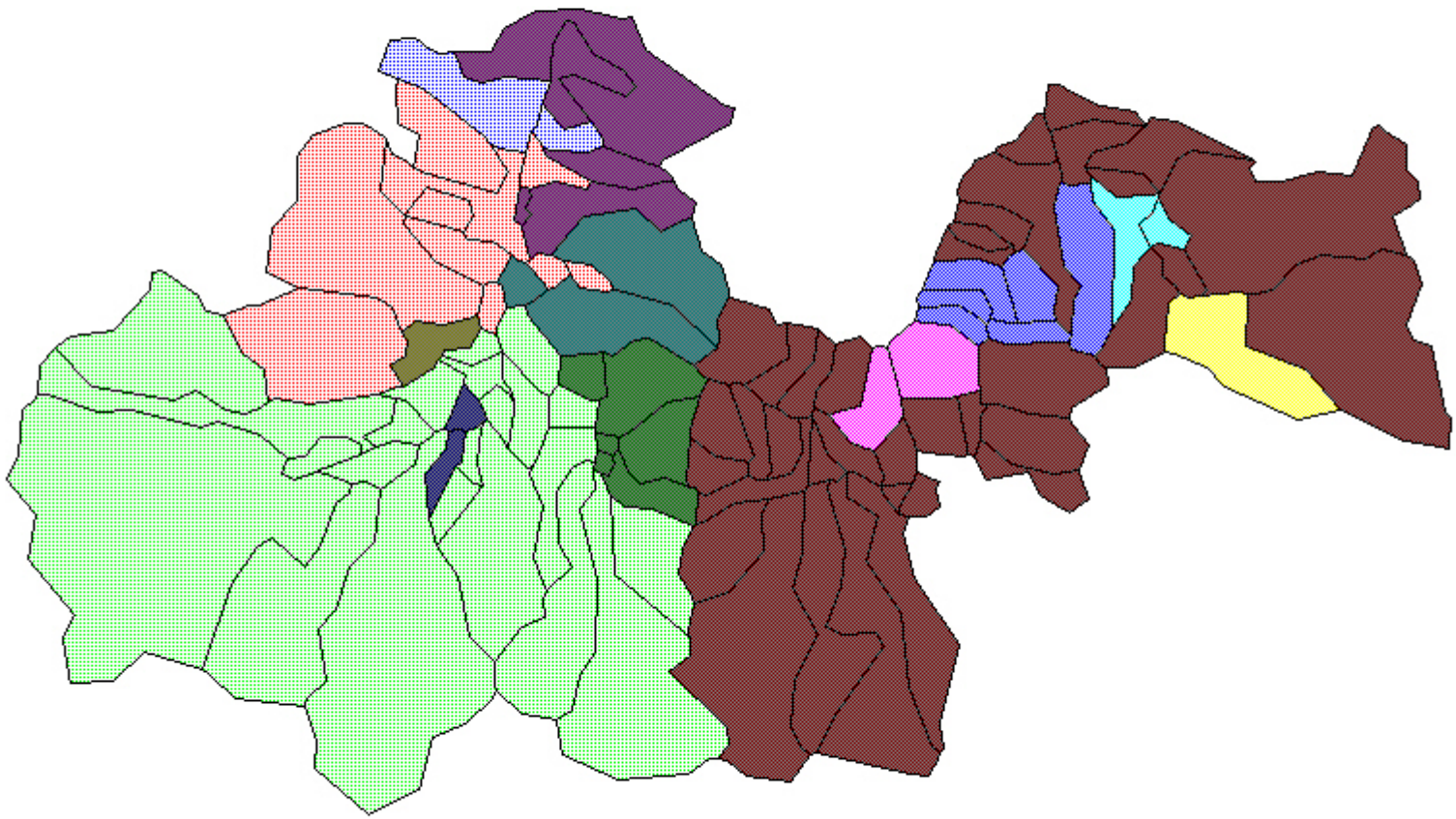


Cartographie des UP-EM et UP-EMEX

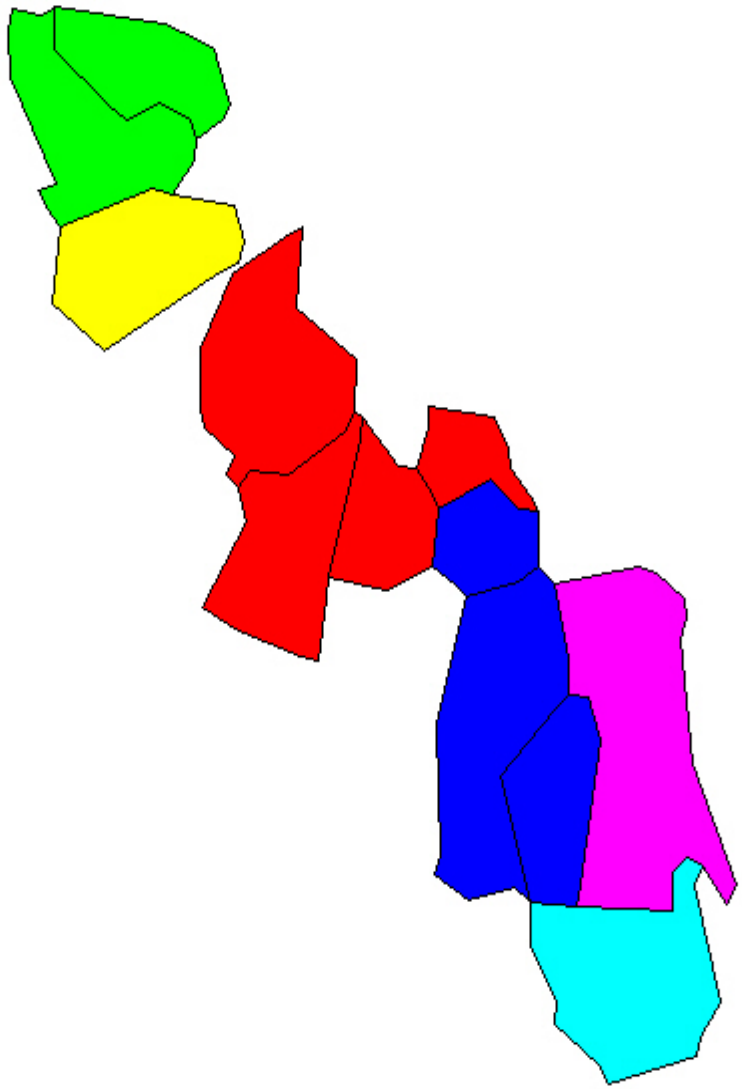
Midi-Pyrénées



Exemple de GRCOM créés dans une UP rurale de Midi-Pyrénés



Exemple de GRCOM créés dans une UP rurale de Champagne-Ardenne



7. Les traitements en aval.

- Correction de la non-réponse totale.
- Correction de la non-réponse partielle.
- Calage.
- Calcul de précision.

8. Conclusion.

- Facilités de gestion du réseau d'enquêteurs.
- Tirage intégré (chaîne informatique unique).

- Recherche de la satisfaction des « clients » :
  - Pondération unique.
  - Enrichissement mutuel des estimations nationales et régionales.
  - Homogénéisation des méthodes de traitement et comparabilité :
    - entre le national et le régional
    - entre deux régions à extension.

Bilan global à compléter ultérieurement :

- Cet outil ne s'applique pas aux enquêtes locales..
- .. et ne permet pas de gérer des zones d'enquête à la demande.

- => Qu'en pensent les régionaux ?

ANNEXE : traitement des logements neufs.

- Une base de logements neufs est constituée dans les nouvelles zones de l'EMEX.
- Elle est tirée dans la base SITADEL des permis *achevés*.
- Sa qualité peut être moindre que celle de la BSLN (pas de suivi terrain).

A SUPPRIMER

On montre que si ces 3 conditions sont remplies :

- (i) les probabilités finales d'inclusion sont solutions de l'équation :

$$\sum_{i\in P} x_i - \sum_{i\in S_1} \frac{x_i}{\tilde{\pi}_i} = \sum_{i\in P-S_1} x_i \frac{1-\frac{\pi_i^1}{\tilde{\pi}_i}}{1-\pi_i^1}$$

- (ii) le tirage, conditionnel à s1, de s2 dans P\s1 est compatible avec les probabilités finales d'inclusion solutions de (i)
- Le tirage conditionnel de s2 est équilibré sur la variable z définie par :

$$z_i = x_i \frac{\tilde{\pi}_i^{2/S_1}}{\tilde{\pi}_i} \quad \forall i \in P \setminus s1$$

Programme général à résoudre

- Si l'on souhaite équilibré simultanément sur J variables x^j, on résout :

$$\underset{\tilde{\Pi}_i}{Min} \sum_{i \in P} d(\tilde{\Pi}_i, \Pi_i)$$

Sous les J contraintes :

$$\sum_{i \in P} x_i^j - \sum_{i \in S_1} \frac{x_i^j}{\tilde{\Pi}_i} = \sum_{i \in P - S_1} x_i^j \frac{1 - \frac{\Pi_i^1}{\tilde{\Pi}_i}}{1 - \Pi_i^1} \qquad j = 1 \ldots J$$

- On en déduit plusieurs familles d'estimateurs, parmi lesquelles:

α_i

$$\hat{T} = \sum_{i \in S_1} \frac{\alpha_i Y_i}{\Pi_i^1} + \sum_{i \in S_2} \frac{(1 - \alpha_i) Y_i}{\Pi_i^{2/S_1} \left(1 - \Pi_i^1\right)}$$

- où α_i est un élément de [0,1], Π_i^1 la probabilité d'inclusion d'ordre 1 de l'unité i dans le 1er échantillon
- et Π_i^{2/S_1} la probabilité d'inclusion dans le 2ème échantillon (EMEX), *conditionnellement à la réalisation* s_1 *de l'échantillon* S_1.

Dans les deux cas :

- Les coefficients de pondération des observations dépendent des différentes probabilités d'inclusion.

- Il existe des relations entre ces probabilités :

$$\Pi_i = \Pi_i^1 + E[\Pi_i^{2/S_1} 1_{i \notin S_1}]$$

- qui se simplifient dans certains cas particuliers :
- $\Pi_i = \Pi_i^1 + \Pi_i^2$ dans le cas d'échantillons *disjoints* ;

$\Pi_i = (1 - \mu_i)\Pi_i^1 + \mu_i$ lorsque la probabilité conditionnelle prend une valeur μ_i indépendante des réalisations du 1er échantillon.

Les contraintes :

- Les probabilités d'inclusion des unités tirées dans l'EM sont fixées.
- Les probabilités conditionnelles lors du tirage de l'EMEX ou les probabilités finales se déduisent les unes des autres :
  - Soit on impose les probabilités finales.
  - Soit on introduit des conditions d'autopondération lors du tirage des unités secondaires au sein du tirage de l'EMEX…,
  - … ce qui implique des conditions sur les probabilités conditionnelles.

- Objectif idéal : on veut tirer, conditionnellement à s_1 un deuxième échantillon s_2 dans $\boldsymbol{P} \setminus s_1$ respectant :

1) des probabilités finales d'inclusion des unités, fixées a priori et notées π_i

2) la propriété d'équilibrage suivante :

$$\sum_{i \in s_1} \frac{x_i}{\Pi_i} + \sum_{i \in s_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i$$

Objectifs de l'équilibrage inverse.

- Construire une loi de tirage de s_2 conditionnelle à s_1 telle que:
  - Les probabilités finales d'inclusion sont « proches » de celles initialement fixées
  - L'estimateur d'Horvitz –Thomson vérifie la propriété d'équilibrage sur la variable x:

$$\sum_{i \in s_1} \frac{x_i}{\tilde{\pi}_i} + \sum_{i \in s_2} \frac{x_i}{\tilde{\pi}_i} = \sum_{i \in P} x_i$$

$$\widetilde{\Pi}_i^{2/S_1} = \frac{\widetilde{\Pi}_i - \Pi_i^1}{1 - \Pi_i^1}$$