ESTIMATION DE LA PRÉCISION EN PRÉSENCE DE DONNÉES IMPUTÉES PAR UN MODÈLE

David LEVY

INSEE, Direction régionale de Rhône-Alpes

1. Echantillonnage

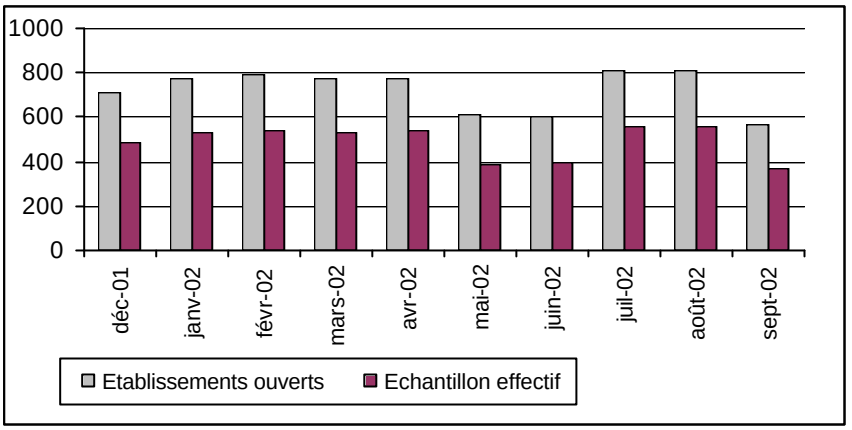
1.1 Contexte

Une enquête auprès des hébergements collectifs a été mise en place en Rhône-Alpes afin de suivre mensuellement l'activité de ce type de structure. Le champ, composé par exemple des villages de vacances, auberges de jeunesse, centres de vacances, vient compléter celui des enquêtes de fréquentation hôtelière. L'objectif de l'enquête est de fournir chaque mois des estimations départementales d'indicateurs d'activité tels que le taux d'occupation, le nombre de nuitées ou d'arrivées.

Il a été décidé, pour pouvoir produire des résultats départementaux et compte tenu du nombre relativement faible d'hébergements dans certains départements, de retenir exhaustivement tous les hébergements de certains départements. Ainsi, les départements de l'Ain (01), de l'Ardèche (07), de la Drôme (26), de la Loire (42) et du Rhône (69) sont enquêtés exhaustivement (302 établissements), alors que les départements de l'Isère (38), de la Savoie (73) et de la Haute-Savoie (74) sont échantillonnés (609 établissements).

On peut constater par ailleurs que les taux de fermeture - et donc la taille attendue chaque mois de l'échantillon d'hébergements ouverts - varient sensiblement d'un mois sur l'autre. Il est donc admis que la précision se « détériore » (de manière relative) certains mois, notamment en mai, juin et septembre. Au total, l'échantillon est constitué en moyenne annuelle de 570 établissements.

Nombre d'établissements ouverts et taille de l'échantillon par mois


1000
800
600
400
200
0
déc-01
janv-02
févr-02
mars-02
avr-02
mai-02
juin-02
juil-02
août-02
sept-02
Etablissements ouverts
Echantillon effectif

1.2 Stratification

Afin d'améliorer la précision, nous avons procédé à une stratification préalable de la partie échantillonnée de la base de sondage. Pour déterminer les variables de stratification pertinentes, des classifications hiérarchiques ont été effectuées à partir des données des enquêtes passées sur les hébergements collectifs[1]. Ces classifications ont donné lieu à des classes d'hébergements sur la base des variables d'intérêt liées à la capacité : occupation des équipements, nuitées passées, type de public et nombre de séjours. Ces classes ont été caractérisées ensuite par les variables disponibles dans la base de sondage : type de centre, offre d'équipements d'hébergement…

Les estimateurs qui seront utilisés étant du type «ratio » sur la capacité d'hébergement (variable auxiliaire), les variables utilisées ont été en réalité, non pas les informations brutes collectées, mais les résidus issus des régressions linéaires (sans terme constant) des variables d'intérêt sur la variable « capacité d'hébergement ». En effet, le paramètre d'intérêt est le taux d'occupation noté R, rapport entre le nombre d'équipements occupés et le nombre d'équipements offerts.

$$R = \frac{\sum_i OCCi}{\sum_i OFFi} = \frac{\overline{OCC}}{\overline{OFF}}$$

avec OCCi le nombre d'équipements occupés du centre i et OFFi le nombre d'équipements offerts du centre i

La variance de l'estimateur correspondant peut s'écrire, après linéarisation (taille d'échantillon théoriquement « suffisamment grande »), de la manière suivante (voir Cochran [1]) :

$V(R) \approx \frac{1-f}{n.\overline{OFF}^2} \frac{\sum_i (OCCi - R.OFFi)^2}{N-1}$ où f est le taux de sondage

L'estimation de la variance dépend ainsi de la variable $Zi = (OCCi - R.OFFi)$.

On a donc pu constater in fine que les informations significativement discriminantes sont :

- la capacité d'accueil ;
- le département ;
- le mode de gestion, sous 2 modalités : lucratif/non lucratif.

L'exercice a porté sur les données des mois de janvier, avril et août 1999 afin de s'assurer de la stabilité des résultats obtenus. On a finalement constitué 30 strates (capacités en 5 modalités / département en 3 modalités / mode de gestion en 2 modalités).

La capacité d'hébergement est la première variable discriminante. Les limites des strates constituées sur la base de cette variable ont été définies par la méthode « optimale » de Dalenius (voir Cochran [1]) . A ce stade, l'allocation par strate s'approche théoriquement de l'allocation optimale. Parallèlement, le calcul des dispersions « intra-strate » de la variable « résidu issu de la régression du

1 Les enquêtes existent mensuellement depuis 1993. On dispose donc de séries de données sur un champ certes plus réduit que le champ de l'enquête rénovée.

nombre d'équipements occupés sur le nombre d'équipements offerts » montre que l'origine d'une part importante de la variance intra provient de la strate regroupant les hébergements disposant du plus grand nombre d'équipements. Aussi, a-t-il été décidé, par mesure de précaution, de sonder exhaustivement cette strate.

De même, compte tenu du très faible nombre d'hébergements gérés sur le mode lucratif, nous avons prévu de tirer exhaustivement tous les individus de ces strates. L'annexe 1 présente les effectifs par strate et la dispersion intra-strate.

Finalement, sur les 12 strates donnant lieu à un véritable échantillonnage, on procède à une allocation proportionnelle.

1.3 Tirage et renouvellement de l'échantillon

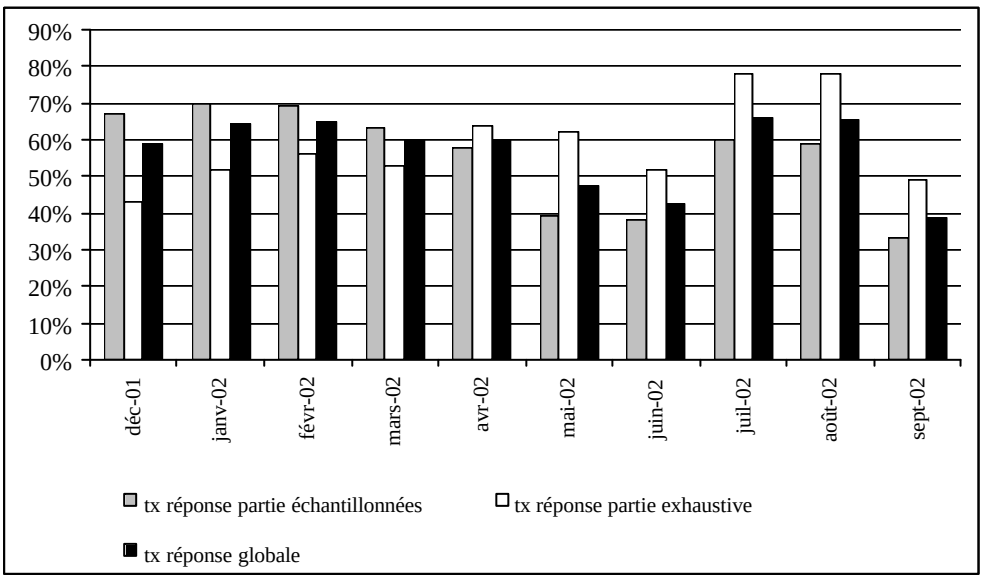
Il s'agit d'un sondage aléatoire simple, strate par strate. L'échantillon est tiré pour une année : il n'y a aucune mise à jour en cours d'année. On prévoit, d'une année sur l'autre, de renouveler par moitié l'échantillon des strates non exhaustives, afin de limiter les charges d'enquêtes.

En revanche, dans les strates et départements exhaustifs il s'agira d'un véritable panel : les hébergements concernés seront donc enquêtés chaque mois tant que l'enquête durera.

2. Correction de la non-réponse : choix de la méthode d'imputation

Les taux de réponse sont assez variables d'un mois sur l'autre, pouvant descendre jusqu'à près de 50%. Il convient donc de corriger cette non-réponse, source de biais non négligeable.

Graphique 1 :Taux de réponse à l'enquête par mois



Le taux de réponse s'entend ici comme le taux de questionnaires renvoyés par les enquêtés. Certains d'entre-eux sont incomplets, ce qui dégrade le taux de réponse de certaines questions.

Une étude des données existantes a permis de mettre en lumière une relation assez forte entre les variables d'occupation des établissements et les variables d'offre d'équipements, dites variables auxiliaires (voir annexe 2).

Le modèle suivant, sous-jacent à l'imputation, repose sur l'hypothèse d'une relation linéaire entre la variable d'intérêt et la variable auxiliaire :

$$y_i = a + bx_i + e_i$$
$$\text{avec : } \mathrm{E}(e) = 0$$
$$\mathrm{V}(e) = s^2 .$$

Les caractéristiques de fréquentation d'un hébergement collectif y_i sont la réalisation de variables aléatoires expliquées par une variable x_i connue (l'équipement offert). L'analyse des corrélations et l'observation de la forme des nuages de points tendent à confirmer cette hypothèse de linéarité.

On s'intéresse à un total $Y = \sum_N y_i$, représentant la fréquentation d'un hébergement. L'estimateur s'écrit :

$$\hat{Y} = \sum_N y_i = \sum_r y_i + \sum_{N-r} \hat{y}_i$$

où N représente l'ensemble de la population, de taille N, et r l'ensemble des répondants, de taille r. L'estimateur s'écrit :

$$\hat{Y} = \sum_r y_i + \sum_{N-r} (\hat{a} + \hat{b}x_i)$$

où $\hat{a}$ et $\hat{b}$ sont les meilleurs estimateurs linéaires sans biais de a et b, au sens des moindres carrés ordinaires .

Ainsi on obtient :

$$\begin{aligned}\hat{Y} &= r\bar{y}_r + \left(\sum_N x_i - \sum_r x_i \right) \hat{b} + (N - r)\hat{a} \\ &= r\bar{y}_r + (X - r\bar{x}_r)\hat{b} + (N - r)(\bar{y}_r - \hat{b}\bar{x}_r) \\ &= N\left[\hat{b}(\bar{X} - \bar{x}_r) + \bar{y}_r\right]\end{aligned}$$

On retrouve l'estimateur par la régression de la moyenne construit sur l'échantillon des répondants.

Le plan de sondage étant stratifié, on a le choix entre l'estimateur par la régression combiné et celui séparé. Le premier consiste à ajuster un modèle sur l'ensemble des répondants, le second à ajuster un modèle sur les répondants mais strate par strate. La comparaison des modèles dans chaque strate conduit à choisir l'estimateur séparé. En annexe 2 sont présentés quelques nuages de points et des régressions par strate.

L'estimateur retenu est donc :

$$\hat{Y} = N \sum_h W_h \left(\bar{y}_h + \hat{b}_h \left(\bar{X}_h - \bar{x}_h \right) \right),$$

où W_h représente le poids de la strate h dans la population. Cet estimateur stratifié est également utilisé pour la partie exhaustive où les strates sont constituées de chacun des départements.

3. Calcul de précision

Pour écrire la variance de l'estimateur, il convient de séparer la partie échantillonnée de la base de la partie exhaustive pour laquelle des hypothèses supplémentaires sont nécessaires.

3.1 Partie échantillonnée

On est dans le cas de l'estimateur par la régression séparé avec imputation des non-répondants par un modèle de régression.

La variance de l'estimateur par la régression séparé s'estime sur l'échantillon, d'après Sarndäl [2] par :

$$\hat{V}\left(\hat{Y}\right) = \sum_h N_h^2 \frac{1-f_h}{n_h} S_{y_h}^2 (1-\hat{r}_h^2)$$

où f_h est le taux de sondage dans la strate h, n_h la taille de l'échantillon de la strate h, $S_{y_h}^2 = \frac{1}{n_h - 1}\sum_{k\in h}(y_k - \bar{y}_h)^2$ et $\hat{r}_h^2$ l'estimation du coefficient de corrélation linéaire dans la strate h.

Afin de prendre en compte le modèle d'imputation permettant de corriger la non-réponse, la variance totale est décomposée d'après Särndal [3] en deux termes : variance due à l'échantillonnage et variance due à l'imputation.

En notant :

U, la population de taille N, s, l'échantillon de taille n

r la taille de l'ensemble des répondants et o la taille de l'ensemble des non-répondants

On observe y_k, la variable d'intérêt, sur r et on impute les valeurs manquantes par un modèle. L'ensemble des données est ainsi constitué des valeurs notées $y_{\bullet k}$ définies par :

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k = x_k \hat{B} & \text{si } k \in o \end{cases}$$

où x_k est une variable auxiliaire connue sur N.

L'estimateur de la variable d'intérêt Y_U s'écrit $\hat{Y}_{\bullet s} = \sum_s w_k y_{\bullet k}$, compte tenu des valeurs manquantes.

En supposant que $\hat{Y}_{\bullet s}$ est non biaisé, la variance totale s'écrit :

$$V_{totale} = E_{imp} E_p E_q \left[\left(\hat{Y}_{\bullet s} - Y_U \right)^2 \right],$$

où "imp" signifie selon le modèle d'imputation, p selon le plan de sondage et q selon le mécanisme de non-réponse.

En notant que $\left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - Y_U\right)^2 = \left[\left(\hat{\mathrm{Y}}_{\mathrm{s}} - Y_U\right) + \left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - \hat{\mathrm{Y}}_{\mathrm{s}}\right)\right]^2$, on a :

$$\mathrm{V}_{\text{totale}} = \mathrm{E}_{\text{imp}}\mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}}\left(\hat{\mathrm{Y}}_{\mathrm{s}} - Y_U\right)^2 + \mathrm{E}_{\text{imp}}\mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}}\left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - \hat{\mathrm{Y}}_{\mathrm{s}}\right)^2 + 2\mathrm{E}_{\text{imp}}\mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}}\left(\hat{\mathrm{Y}}_{\mathrm{s}} - Y_U\right)\left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - \hat{\mathrm{Y}}_{\mathrm{s}}\right),$$

soit

$$\mathrm{V}_{\text{totale}} = \mathrm{E}_{\text{imp}}\mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}}\left(\hat{\mathrm{Y}}_{\mathrm{s}} - Y_U\right)^2 + \mathrm{E}_{\text{imp}}\mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}}\left[\left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - \hat{\mathrm{Y}}_{\mathrm{s}}\right)^2 + 2\left(\hat{\mathrm{Y}}_{\mathrm{s}} - Y_U\right)\left(\hat{\mathrm{Y}}_{\bullet \mathrm{s}} - \hat{\mathrm{Y}}_{\mathrm{s}}\right)\right].$$

Ainsi :

$$\mathrm{V}_{\text{totale}} = \mathrm{E}_{\text{imp}} V_s + \mathrm{E}_{\mathrm{p}}\mathrm{E}_{\mathrm{q}} V_{imp} \ ;$$

où V_s désigne la variance de $\hat{\mathrm{Y}}_{\mathrm{s}}$ selon le plan de sondage. Il s'agit en réalité de la variance dans le cas de réponse complète. La composante V_{imp} est la variance de l'erreur d'imputation, conditionnelle à p et q. Elle se compose de deux termes, le deuxième étant négligeable devant le premier.

La variance totale se décompose en deux termes, la variance d'échantillonnage et la variance due au modèle d'imputation :

$$\boxed{V_{tot} = V_s + V_{imp}}.$$

Une formulation et une procédure de calcul des estimation $\hat{V}_s$ et $\hat{V}_{imp}$ dans le cas général d'imputation par un modèle sont donnés par Deville et Särndal [4].

3.2 Partie exhaustive

On considère que l'ensemble des répondants forme un échantillon aléatoire simple. On fait donc des hypothèses sur la probabilité de réponse, notée p. Par analogie avec le cas précédent, cela revient à considérer un échantillon dans lequel il n'y a pas de répondant. L'estimateur utilisé reste celui de la régression.

Soit N la taille de la population et r le nombre de répondants. On suppose que r suit une loi de Bernoulli : $r \to b(N, p)$. La variance de l'estimateur par la régression est alors conditionnée à r et s'écrit :

$$V(\hat{Y}_{reg}) = V_r\left[E(\hat{Y}_{reg} / r)\right] + E_r\left[V(\hat{Y}_{reg} / r)\right] .$$

De plus, on suppose l'estimateur par la régression sans biais :

$$E(\hat{Y}_{reg} / r) = Y ,$$

donc

$$V_r\left[E(\hat{Y}_{reg} / r)\right] = 0 ,$$

ainsi,

$$V(\hat{Y}_{reg}) \approx E_r\left[V(\hat{Y}_{reg} / r)\right] .$$

Dans le cas d'un sondage aléatoire simple et lorsque la taille de l'échantillon est suffisamment grande, la variance de l'estimateur par la régression peut s'écrire d'après Ardilly [5] :

$$V(\hat{Y}_{reg}) = \frac{1-f}{n} S_u^2$$

où S_u^2 est la variance calculée sur les résidus de la régression.

Ainsi, lorsque r est suffisamment grand, on a :

$$V(\hat{Y}_{reg}) = E\left[\left(1-\frac{r}{n}\right)\frac{S_u^2}{r}\right],$$

soit

$$\boxed{V(\hat{Y}_{reg}) = \left(E\left(\frac{1}{r}\right) - \frac{1}{N}\right) S_u^2}.$$

Le calcul de $E\left(\frac{1}{r}\right)$ se fait par développement limité qui donne l'approximation suivante :

$$E\left(\frac{1}{r}\right) \approx \frac{1}{E(r)}\left(1+\frac{V(r)}{(E(r))^2}\right).$$

Or *r* suit la loi $b(\mathrm{N},\mathrm{p})$, donc $E(r) = Np$ et $V(r) = (1-p)Np$,

ainsi

$$E\left(\frac{1}{r}\right) \approx \frac{1}{Np}\left(1+\frac{1-p}{Np}\right).$$

On estime *p* par $\hat{p} = \frac{r}{N}$ dans chaque groupe homogène constitué ici des départements.

Finalement, on a :

$$\boxed{V(\hat{Y}_{reg}) = \left(\frac{1}{Np}\left(1+\frac{1-p}{Np}\right) - \frac{1}{N}\right) S_u^2}.$$

4. Résultats

On cherche à estimer le total des nuitées par département pour un mois donné. Le modèle d'imputation est :

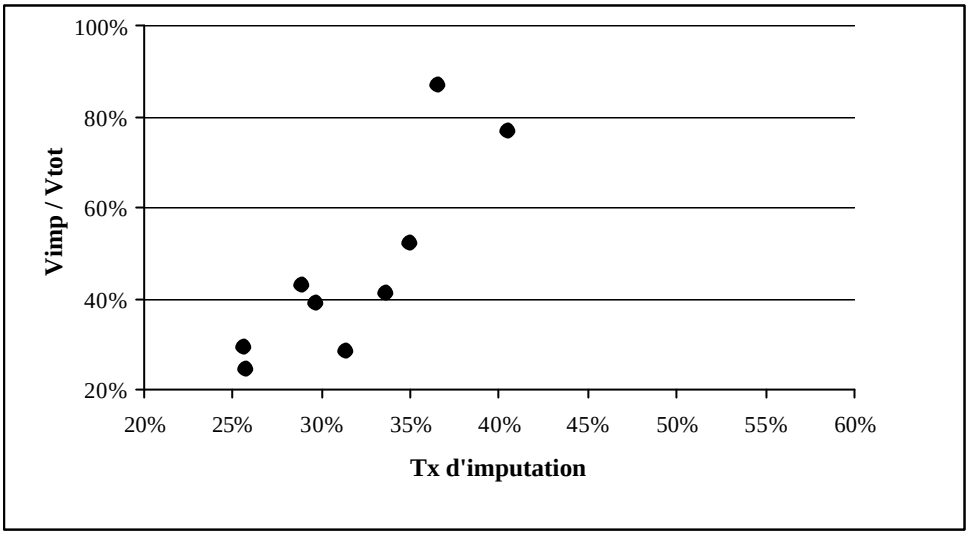
NUITEES = a + b.CAPA + e

où CAPA est la capacité totale en chambres de l'établissement. Les exemples portent sur la saison hiver 2001-2002 et la saison été 2002.

4.1 Part de la variance due à l'imputation dans la variance totale

Le graphique ci-dessous présente la part de la variance due à l'imputation dans la variance totale en fonction du taux d'imputation.

Graphique 2 : Part de la variance due à l'imputation sur la variance totale en fonction du taux d'imputation



Chaque point représente un mois d'enquête, de décembre 2001 à septembre 2002. La part de la variance due à l'imputation augmente de manière quasi-linéaire en fonction du taux d'imputation.

4.2 Comparaison de l'estimateur simple et de celui par la régression

On compare l'estimation d'un total issue de l'estimateur simple stratifié et sa variance à celui de la régression séparée avec sa variance (y compris la prise en compte du modèle d'imputation). Dans le cas de l'estimateur simple, la non-réponse est corrigée par repondération des répondants. Les estimations sont calculées sur la saison hiver 2001-2002 et été 2002.

Tableau 1 : L'estimateur simple et l'estimateur par la régression

	Estimateur simple		Estimateur par la régression		
	Estimation(*)	Précision relative	Estimation(*)	Précision relative	avec prise en compte de l'imputation
déc-01	478	19%	440	15%	20%
janv-02	1 690	11%	1 610	8%	12%
févr-02	1 957	11%	1 880	9%	10%
mars-02	1 377	14%	1 326	8%	11%
avr-02	969	21%	888	9%	16%
mai-02	549	31%	552	3%	25%
juin-02	551	33%	389	11%	23%
juil-02	1 632	14%	1 456	9%	15%
août-02	1 980	16%	1 738	9%	15%
sept-02	396	48%	237	9%	41%

(*) En milliers de nuitées

La valeur estimée de l'estimateur simple est toujours supérieure à celle de l'estimateur par la régression. Il est intéressant de noter que la variance de ce dernier est toujours la plus faible. Cela s'explique par la relation entre les deux variances (voir Ardilly [5])

$$V(\hat{Y}_{reg}) = (1 - r^2)V(\hat{Y}),$$

ce qui entraîne :

$$\boxed{0 \leq V(\hat{Y}_{reg}) \leq V(\hat{Y})}.$$

Nous pouvons constater que la variance de l'estimateur par la régression avec prise en compte du modèle d'imputation est très souvent inférieure à celle de l'estimateur simple.

Bibliographie

[1] Cochran, W.G., « Sampling Techniques », Wiley, New-York, 1977.

[2] Särndal, C.E., Swenson, B., and Wretman, J., « Model Assisted Survey Sampling », Springer, New York, 1992.

[3] Särndal, C.E., « Méthodes pour estimer la précision des estimations lorsqu'il y a eu imputation », Recueil du Symposium 90 de Statistique Canada, octobre 1990.

[4] Deville, J.C., et Särndal, C.E., «Estimation de la variance en présence de données imputées par modèle », Document de travail de l'INSEE n° 2F9102, 1991.

[5] Ardilly, P., « Les techniques de sondages », éditions Technip, 1994.

ANNEXE 1 : MESURE DE DISPERSION

Strates					Statistiques *	
N°	Capacité	DEP	MODEGEST(*)	Effectif base 98	MOYENNE	Variance "intra"
1	1	38	1	25	- 63,7	19 491,0
2	1	38	2	0		
3	1	73	1	19	- 73,3	15 941,6
4	1	73	2	4	3,7	1,3
5	1	74	1	15	- 5,3	15 321,5
6	1	74	2	2	- 62,3	28 284,5
7	2	38	1	15	28,1	66 610,4
8	2	38	2	2	12,9	4 588,7
9	2	73	1	24	- 38,3	45 450,0
10	2	73	2	1	- 323,5	
11	2	74	1	24	- 23,3	58 259,5
12	2	74	2	4	- 67,8	35 770,2
13	3	38	1	7	276,6	89 162,0
14	3	38	2	0		
15	3	73	1	24	- 75,4	141 406,1
16	3	73	2	1	381,2	
17	3	74	1	20	- 135,4	174 139,3
18	3	74	2	1	121,0	
19	4	38	1	7	- 503,9	251 001,0
20	4	38	2	1	- 826,0	
21	4	73	1	18	- 174,5	192 545,4
22	4	73	2	1	- 886,2	
23	4	74	1	14	103,0	185 003,2
24	4	74	2	0		
25	5	38	1	2	156,6	130 313,8
26	5	38	2	1	3 335,9	
27	5	73	1	18	- 54,4	393 969,6
28	5	73	2	5	1 718,9	821 452,2
29	5	74	1	7	10,8	1 507 247,3
30	5	74	2	1	1 991,1	
Ensemble					**- 1,9**	**4 175 958,5**

(*) MODEGEST vaut 2 pour un établissement à but lucratif, 2 pour les autres.
La variable utilisée est constituée du résidu de régressions linéaires (sans terme constant) des variables d'intérêt sur la variable « capacité d'hébergement ».

ANNEXE 2 : Nuages de points par strate

On présente pour certaines strates des nuages de points formés des variables "nombre de chambres offertes" et "nombre de chambres occupées" dans quelques strates.