IMPUTATION MASSIVE DANS L'ENQUÊTE MENSUELLE HÔTELIÈRE EN FRANCE

Pascal Ardilly

Insee, Département des méthodes statistiques

Enquête mensuelle de fréquentation touristique :
12 000 hôtels tirés parmi 18 000 (environ).

Nombre de chambres occupées, de nuitées, d'arrivées, part de clientèle d'affaire … ?

Echantillonnage annuel complexe, composé de deux tirages stratifiés successifs avec s.a.s. dans chaque strate. L'échantillon est rotatif avec gestion du recouvrement annuel.

Une grande originalité : **aucune pondération** !

⇓

Les données des hôtels *non échantillonnés* et celles des *non-répondants* sont donc **imputées**.

En option, l'échantillon à t exclut tous les échantillonnés à $t-1$ jugés « mauvais répondants », c'est-à-dire :

- nombre de mois d'ouverture >= 6

ET

- % de mois avec réponse parmi les mois d'ouverture $\geq$ *seuil au choix*.

L'imputation s'appuie sur des **modèles de comportement**.

L'intérêt va à des estimations de moyennes à tous niveaux, mais en particulier sur des petits domaines (partenaires locaux) $\Rightarrow$ approche *déterministe*.

On connaît le nombre de chambres offertes.

C'est le point de départ d'une modélisation 'en cascade' portant (presque toujours) sur des *taux* :

Nombre de chambres occupées / Nombre de chambres offertes

$\Downarrow$

Nombre de nuitées / Nombre de chambres occupées

$\Downarrow$

Nombre de nuitées étrangères / Nombre de nuitées

Principales difficultés liées à cette approche

i) Importance des effets locaux non captables par les informations de la base de sondage (par exemple la qualité de l'accueil, ou l'environnement immédiat).

Les variables explicatives sont limitées :

- localisation de l'hôtel ($\Leftrightarrow$ espace touristique)
- catégorie d'hôtel (7 modalités),
- type d'hôtel (4 modalités),
- nombre de chambres offertes (3 tranches).

iii) Sélection des variables explicatives :

- Quelles modalités (regroupement ou non de modalités fines) ?
- Interactions ou pas ?
- Traitement des variables non-significatives ?

En amont : possibilité de partitionnement géographique à géométrie variable (regroupement automatisé des espaces touristiques les plus fins) $\rightarrow$ guidé par un nombre minimum de répondants (paramétré).

iii) *In fine*, situation de compromis à trouver, la masse de données étant considérable :

- grand nombre de modèles concurrents,
- 8 variables d'intérêt,
- 13 (voire 22) régions,
- 12 mois de traitement - chaque année !

Comment prendre en compte toute cette diversité pour choisir le modèle ?

Dans tous les cas, avantage des modèles « parcimonieux » :

corrélation ≠ causalité

causalité ⇔prédiction

Toutes les variables sont initialement des **variables de comptage**.
La **fréquence des zéros** va structurer le type de modélisation.

Variables exceptionnellement nulles

Nombre de chambres occupées, de nuitées, d'arrivées…

1/ Modèles linéaires à effets fixes (PROC GLM)

$$Y_i = X_i \cdot B + \varepsilon_i$$

où $E\varepsilon_i = 0$ et $V\varepsilon_i = \sigma^2$.

X_i = *systématiquement* zone géographique, catégorie d'hôtel, type d'hôtel, tranche de taille.

En sus, éventuellement :

- ajout d'effets croisés (impliquant 2 variables);
- possibilité de retenir les régresseurs de *p-value* < *seuil* au choix (0.10 ?). On enchaine deux étapes de sélections successives utilisant le critère *p-value.*

X_i est constitué par des *indicatrices* (fort gain de souplesse, versus le 'CLASS' de SAS).

On applique aussi la PROC GLMSELECT (sélection optimale de régresseurs - méthode *stepwise*).

2/ Modèles linéaires avec effets aléatoires indépendants (PROC MIXED)

$$Y_i = X_i \cdot B + v_g + \varepsilon_i$$

où g *zone géographique*, avec

$$E(v_g) = 0 \qquad V(v_g) = \sigma_v^2 \qquad Cov(v_g, v_h) = 0$$

$\rightarrow$ économie de paramètres (évite tous les régresseurs géographiques).

Il peut y avoir un échec ($\hat{\sigma}_v^2 = 0$). On relâche alors les contraintes techniques de convergence de l'algorithme de Newton (processus EMV).

3/ Modèles linéaires avec effets aléatoires corrélés : distance $d_{g,h}$ entre les zones g et h, puis

$$Cov(v_g, v_h) = \lambda \cdot \exp(-\mu . d_{g,h}) .$$

Nota : le paramètre μ ne participe pas à l'EMV et doit être fixé *a priori* 'à la main' (limitation de SAS ?).

Soit au final 26 modèles différents !

Au passage, on édite tous les hôtels qui ont des résidus *fréquemment* atypiques.

Appréciation de la pertinence de l'ajustement :

- Utilisation des tests (*p-value*) ;
- Techniques habituelles (distribution des résidus, R^2, AIC, BIC, validation croisée, etc.) ;
- Corrélation entre les taux imputés issues de tous les modèles ;
- CV des différents taux imputés, hôtel par hôtel.

Critère '*universel*' de sélection de modèle : on forme $(Y_i - \hat{Y}_i)^2$ sur chaque répondant, puis on somme.
Mais c'est sensible au nombre de régresseurs.

Questions techniques embarrassantes sur le choix des régresseurs ‘fixes’ :

- seuil de taille d’échantillon répondant par modalité explicative (regroupement de modalités si besoin) ?
- interactions sélectionnées si elles concernent un nombre d’hôtels répondant supérieur à un certain seuil (par exemple entre 5 et 10).
- méthode de sélection automatique : faut-il se débarrasser des variables non significatives ? Seuil de *p-value* (retenu : 0.10) ?

Prise en compte du *passé* si possible : ajouter aux régresseurs la variable d’intérêt du même mois de l’an passé (t-1).

Un sérieux problème : il faut que l’hôtel soit répondant (et donc échantillonné) le même mois de l’année t-1.

Réponse à la variable d'intérêt		Niveau d'implication de l'hôtel
par le passé	le mois d'imputation	
OUI	OUI	Ajustement du modèle *avec* régresseur « passé »
NON	OUI	Ajustement du modèle *sans* régresseur « passé »
OUI	NON	Imputation à partir du modèle avec régresseurs passé
NON	NON	Imputation à partir du modèle sans régresseur passé

Les régresseurs passés s'avèrent très efficaces mais hélas peu d'hôtels sont concernés (cas de la 3ème ligne).

Variables pouvant être nulles avec un nombre limité de zéros

Typiquement : nombre de nuitées étrangères (NUIET)

1/ Loi de Poisson ou loi binomiale-négative (PROC GENMOD)

$Y_i \rightarrow P(\lambda)$ avec
$Log\,(\lambda) = X_i \cdot B$

$Y_i \rightarrow BN(n, p)$
$Log\,(EY_i) = X_i \cdot B$

$$\text{M V} \quad \Rightarrow \quad \hat{Y}_i = \exp\,(X_i.\hat{B})$$

2/ Loi de Poisson avec effet aléatoire géographique (PROC GLIMMIX)

$Y_i \rightarrow P(\lambda)$ avec
$Log(\lambda) = X_i \cdot B + v_g \quad \Rightarrow \quad \hat{Y}_i = \exp(X_i.\hat{B} + \hat{v}_g)$

et v_g : effet aléatoire géographique (seul modèle à effets aléatoires)

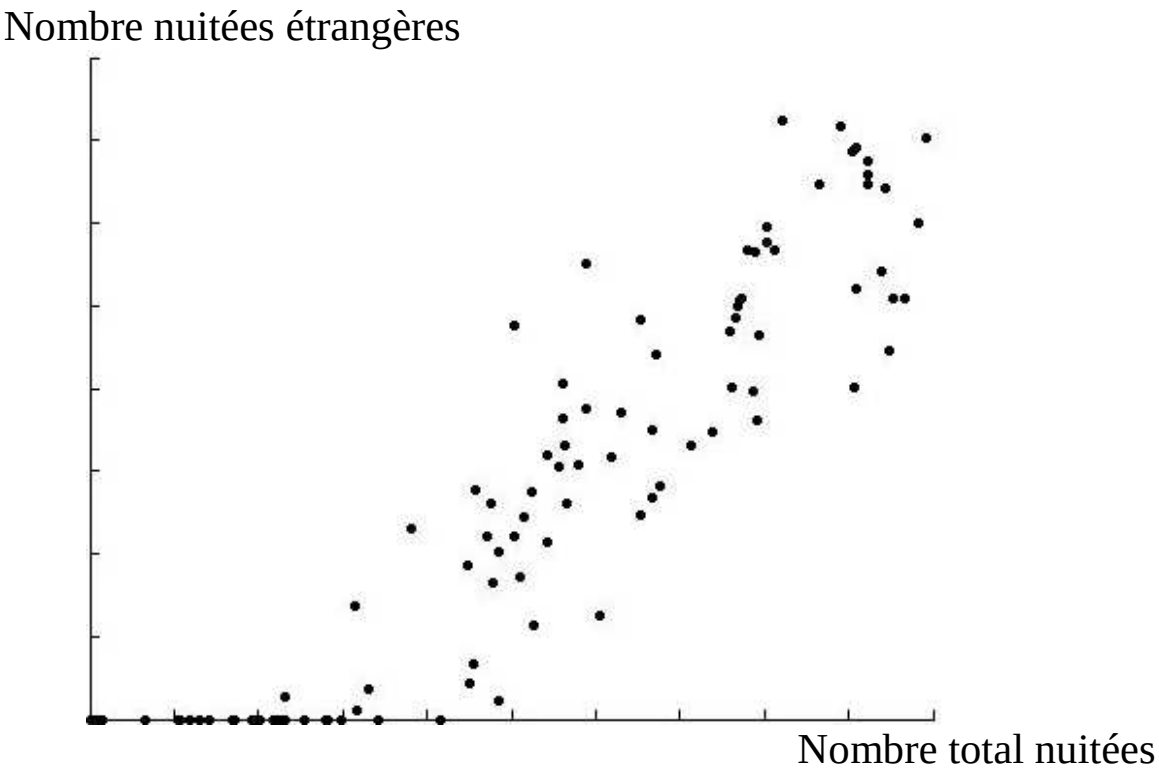
3/ Modèles Tobit (PROC LIFEREG)

Variable collectée $Y_i \rightarrow$ Variable sous-jacente Y_i^*

$Y_i^* = X_i \cdot B + \varepsilon_i$ où $E\varepsilon_i = 0$ et $V\varepsilon_i = \sigma^2$

$Y_i = 0$ si $Y_i^* \leq 0$

$Y_i = Y_i^*$ si $Y_i^* > 0$



Imputation : $\hat{Y}_i = EY_i$

4/ Modèles dits *Zero-inflated*

(PROC GENMOD + option ZEROMODEL)

Probabilité ω : $Y = 0$
Probabilité $1-\omega$: $Y \rightarrow P(\lambda)$

Logit $(\omega) = (NUITOT, categ, zone).B_1$
Log $(\lambda) = X_i \cdot B_2$

NUITOT peut être collectée ou imputée.

Imputation : $\hat{Y}_i = (1-\hat{\omega}) \cdot \hat{\lambda}$

18 modèles concurrents au final

Les indicateurs d'ajustement « habituels » sont compliqués (déviance, AIC, BIC,…) et ne se comparent qu'en cas de modèles *emboités*.

Comparateur universel de qualité : de nouveau utilisation des erreurs $(Y_i - \hat{Y}_i)^2$ pour les *répondants*.

Les variables de comptage avec valeurs nulles fréquentes

Typiquement : nombre de nuitées étrangères *par pays*.

- ventilation de NUIET à assurer selon 47 pays !
- des données collectées de (très) mauvaise qualité …

i : zone géographique

j : condensé de la catégorie et du type d'hôtel

k : pays de résidence

Une nuitée étrangère quelconque a une probabilité $\pi_{i,j,k}$ de satisfaire aux critères (i, j, k) : schéma *multinomial.*

On regroupe les nationalités « rares » dans une seule modalité : limiter le nombre de nationalités *principales* (entre 10 et 20 ?) $\rightarrow$ évite les problèmes numériques.

1/ Modèle logit généralisé

PROC LOGISTIC (/ Link = GLOGIT)

Les régresseurs sont des *indicatrices* construites spécifiquement à partir des modalités de i et de j.

$$Log \frac{\pi_{i,j,k}}{\pi_{i,j,K}} = a^k + b_i^k + c_j^k + \gamma_{i,j}^k$$

Interactions $\gamma_{i,j}^k$ en option (filtre selon le nombre de répondants dans la case (i, j)) + passage éventuel d'un test de significativité.

$$\Rightarrow \qquad \pi^*_{k|i,j} = \frac{\hat{\pi}_{i,j,k}}{\sum_{k=1}^{K} \hat{\pi}_{i,j,k}}$$

Nota : la procédure la plus 'naturelle' est PROC CATMOD $\Rightarrow$ traitement SAS plus rapide mais moins de souplesse dans le choix des régresseurs (pas de sélection).
Absence possible de prédiction $\hat{\pi}_{i,j,k}$ si aucun répondant n'est présent dans (i,j).

2/ Modèle Log linéaire

Modèle portant sur la structure des probabilités, ne distinguant pas les variables explicatives de la variable expliquée.

PROC CATMOD

i) Version très contraignante (modèle d'indépendance) :

$$Log\pi_{i,j,k} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z$$

ii) Version moyennement contraignante :

$$Log\pi_{i,j,k} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{i,j}^{xy} + \lambda_{i,k}^{xz} + \lambda_{j,k}^{yz}$$

iii) Version saturée (sans valeur ajoutée !) :

$$Log\pi_{i,j,k} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{i,j}^{xy} + \lambda_{i,k}^{xz} + \lambda_{j,k}^{yz} + \lambda_{i,j,k}^{xyz}$$

$$\Rightarrow \hat{\pi}_{i,j,k} = \frac{n_{i,j,k}}{n}$$

Estimation des λ puis des $\pi_{i,j,k}$ par maximum de vraisemblance. Grand nombre de paramètres à estimer (plusieurs centaines parfois).

Il n'y a (quasiment) pas de valeurs imputées à zéro - ce qui peut surprendre l'utilisateur final des données.

Passage aux nuitées étrangères par pays :

$\forall h \in (i, j)$:

$$\pi^{*}_{k|i,j} = \frac{\hat{\pi}_{i,j,k}}{\sum_{k=1}^{K} \hat{\pi}_{i,j,k}} \quad \Rightarrow \quad NUIET_{h,k} = \pi^{*}_{k|i,j}.NUIET_{h}$$

et ensuite calcul des erreurs de prédiction des répondants.

1er sentiment : les variables retenues sont peu explicatives (décevant). Poids des erreurs de mesure ?

Confirmé par la comparaison entre les hôtels des structures empiriques des nuitées par nationalité, à $(région, i, j)$ fixé : très gros CV (de 100% à 200%) !

Pourquoi pas les méthodes de prédiction de la partie précédente ? → Gain de programmation considérable + réticence *intuitive* à cause de la fréquence des zéros !

Un point gênant (en théorie) : cette approche fournit *au passage* les $\hat{\pi}_{i,j}$, qui permettraient aussi d'imputer le nombre total d'étrangers dans (i, j).
Ce peut être en contradiction avec la phase d'imputation de NUIET retenue (logique « descendante » et logique « montante » pourraient donc différer fortement).

Conclusion

L'univers des choix de modèle n'est (en général) pas très familier au statisticien d'enquêtes (qui devrait peut-être s'y plonger davantage …).

- L'embarras du choix dans les modèles ;

- Des erreurs 'emboitées' : inquiétude légitime sur la qualité obtenue en bout de chaine ;

- En phase de production, le processus idéal de validation semble hors de portée ;

- Validation pratique ?
  - critère automatisé basé sur l'erreur globale de prédiction pour les répondants ;
  - apprécier la vraisemblance des estimations sur domaines (par avis d'expert).

⇓

Phase de test : on s'apprête à donner un avantage aux modèles 'classiques' avec des interactions.