L'ENQUÊTE MENSUELLE HÔTELIÈRE EN FRANCE : ÉCHANTILLONNAGE CONTRAINT PAR DOMAINE ET IMPUTATION MASSIVE

Pascal Ardilly

Insee, Département des méthodes statistiques, 165 Bd Garibaldi 69003 Lyon, France
pascal.ardilly@insee.fr

Résumé. L'enquête mensuelle hôtelière de l'Insee utilise largement la méthode du tirage stratifié avec sondage aléatoire simple dans chaque strate. Néanmoins, au sein des régions, il existe des zones géographiques distinctes des strates sur lesquelles les échantillons doivent respecter des tailles prédéfinies. Il s'agit donc de contraintes relatives à des domaines. Les zones géographiques en question forment un zonage, c'est-à-dire une partition de la région. On doit pouvoir faire face à un système de contraintes portant sur deux zonages distincts, plus ou moins indépendants. La solution du problème se traduit par la résolution d'une minimisation de fonction quadratique sous des contraintes d'égalité et d'inégalités, que le logiciel R prend en charge. Par ailleurs, s'agissant d'une enquête qui n'affecte pas de poids de sondage aux hôtels répondants, l'estimation procède par imputation massive des hôtels non échantillonnés ou non-répondants. Cette approche par modélisation conduit à tester différents modèles de prédiction pour chaque variable d'intérêt.

Mots-clés. Estimation par domaine, stratification, optimisation sous contraintes, imputation par modèle.

1 Introduction

Chaque mois, l'Insee procède à une enquête de fréquentation touristique auprès des hôtels (classés ou non) situés sur l'ensemble du territoire français. Le questionnaire collecte des informations relatives au nombre de chambres occupées, au nombre de nuitées, au nombre d'arrivées, à la part de clientèle d'affaire. Pour la plupart des statistiques produites, on distingue les clients selon leurs pays d'origine. Il s'agit d'une enquête 'entreprises' à échantillon aléatoire tiré une fois par an dans une base de sondage rassemblant l'ensemble des établissements constituant le parc hôtelier de la France (environ 18 000 hôtels). L'estimation est originale car elle n'utilise aucune pondération : les données relatives aux hôtels non échantillonnés et aux hôtels échantillonnés mais non-répondants sont entièrement imputées. Il s'agit donc d'une approche par modélisation massive. En 2015, l'Insee a souhaité rénover l'ensemble de la procédure d'échantillonnage et d'estimation. Ce document met d'abord l'accent sur une difficulté particulière rencontrée lors de la phase de mise au point de l'échantillonnage, puis aborde le contexte de la modélisation.

2 Un échantillonnage contraint par domaine

Tous les échantillonnages effectués sont des tirages stratifiés avec sondage aléatoire simple dans chaque strate. Les strates croisent les régions administratives, au nombre de 22 en 2015, un découpage géographique pérenne de la région appelé 'espace touristique', la catégorie

d'hôtel (il y en a sept, distinguant le niveau de confort), le type d'hôtel (deux types sont distingués, opposant les hôtels de chaine aux hôtels indépendants) et la tranche de taille en nombre de chambres (quatre modalités). Contrairement au cas des enquêtes sur les ménages, on a à faire à des méthodes d'échantillonnage simples qui *a priori* sont bien maitrisées et ne posent aucun problème sérieux. Néanmoins, il a fallu faire face à certaines difficultés techniques dont la principale provient d'une exigence de taille d'échantillon par domaine. En effet, les territoires des régions sont découpés en zones géographiques qui forment une partition discriminant les comportements touristiques. Ces découpages viennent en sus de l'espace touristique et peuvent être sans relation avec lui. On parle alors de zonage de diffusion au sein du territoire régional - ou de zonage tout court. Pour un zonage donné (certaines régions distinguent près de dix zonages différents), les partenaires de l'Insee en région attendent des statistiques produites dans chaque zone, ce qui impose de contrôler les tailles d'échantillon par zone et de s'éloigner ainsi sensiblement de l'allocation 'en moyenne ' proportionnelle. Dans certaines zones, qui peuvent comprendre très peu d'hôtels, on souhaite un fort taux de sondage, allant même jusqu'à pratiquer une enquête exhaustive. Si ces zones de diffusion étaient des strates de tirage - comme l'est l'espace touristique - il n'y aurait pas d'obstacle car on contrôlerait la taille d'échantillon, mais en la circonstance il s'agit seulement de domaines parce que les stratifications en place relèvent de critères jugés prioritaires et produisent déjà des échantillons de petite voire très petite taille par strate. Considérant la précision des estimations dans ce contexte, les exigences de production au niveau national fixent des allocations de taille d'échantillon par strate qui sont *a priori* en contradiction avec les objectifs régionaux. Pour ajouter à la difficulté, certaines régions veulent respecter des tailles d'échantillon par zone relativement à deux zonages différents, conçus indépendamment l'un de l'autre.

Pour faire face à cette double exigence de respect d'objectifs portant sur les tailles des échantillons national et régionaux, l'idée centrale sur laquelle repose la méthodologie d'échantillonnage rénové consiste à procéder en deux temps. On effectue deux tirages successifs d'hôtels dans la base de sondage disponible : un tirage dit "principal" qui est une opération relativement optimisée et qui justifie son existence par l'objectif national, et dans la foulée un tirage dit "complémentaire" qui est sensiblement moins optimisé mais qui s'attache plus à la satisfaction des besoins régionaux. Le second est conçu comme un complément du premier pour en quelque sorte compenser ce que le premier n'a pas pu faire pour la satisfaction des besoins locaux. Il en résulte une perte d'optimisation de l'échantillon final considéré dans son ensemble mais elle se justifie par le second objectif qui est d'initiative "locale". Toute la difficulté consiste à fixer un partage d'influence entre ces deux objectifs, qui peuvent très bien être contradictoires[1]. Cet arbitrage passe essentiellement par le choix de la taille de l'échantillon principal : plus cette taille principale est grande, plus il sera difficile de satisfaire les contraintes locales mais plus l'échantillon final sera "efficace" pour produire une estimation nationale.

On détaille maintenant le processus, qui se déroule à l'intérieur de chaque région. Les plus gros hôtels de la région sont tout d'abord retenus avec une probabilité égale à 1, ainsi que certains hôtels répondants à des critères *ad hoc* définis en extension (par exemple des hôtels au comportement jugé atypique). On définit ensuite un ensemble d'hôtels dits 'exclus', que l'on ne veut jamais interroger (il y en a peu en nombre, cette exclusion n'étant pas correcte sur le plan théorique). Par ailleurs, bien que cette option ne soit pas satisfaisante sur le plan

[1] Un exemple caricatural : si dans une région donnée aucune exigence n'est formulée au niveau département, l'objectif local consistera à ne tirer aucun hôtel dans ce département. En tout cas, en pratique, moins il y en aura et plus la région sera satisfaite. Ce n'est évidemment pas la façon de voir du niveau national ...

théorique, on donne la possibilité au sondeur d'exclure de tout échantillonnage une liste d'hôtels déclarés 'mauvais répondants'. Le caractère de 'mauvais répondant' relève du comportement passé des hôtels déjà échantillonnés. On peut discuter de la pertinence de cette option très empirique, mais il faut reconnaitre qu'il n'est pas utile de sélectionner des hôtels dont on sait pertinemment par avance qu'ils ne répondront jamais à aucun questionnaire, et mieux vaut alors reporter les moyens de suivi et de relance sur une population potentiellement coopérative. Le tirage principal est paramétré par un nombre total d'hôtels à tirer dans la région, lequel reste 'au choix' du sondeur mais nécessairement inférieur au nombre total d'hôtels à enquêter dans la région. Par exemple l'échantillon final concerne 1000 hôtels, dont 700 seront tirés par l'échantillon principal. Une allocation est ensuite calculée par strate : selon les circonstances, soit il s'agit d'une allocation proportionnelle standard (la clé de ventilation peut être le nombre total d'hôtels dans la strate mais ce peut être aussi le nombre total de chambres offertes - le critère est laissé au choix, en option), soit on utilise une allocation optimisée de type Neyman en exploitant la dispersion des taux d'occupation des mois passés quand c'est possible. Les strates régionales sont en théorie un croisement de l'espace touristique, de la catégorie d'hôtel, du type d'hôtel et de la tranche de taille : cela conduit à des découpages d'une finesse excessive et impose donc en phase opérationnelle des regroupements. De fait, les véritables strates de tirage sont en réalité des agrégations 'à façon' des strates théoriques. Pour compliquer le paysage, ces regroupements évoluent d'une année sur l'autre parce qu'on a cherché à en minimiser le nombre... Quoi qu'il en soit, à la fin de l'étape du tirage principal, on dispose d'un échantillon d'hôtels qui relève d'une certaine optimisation. Aucun processus n'est parfait mais en la circonstance il semble que la plupart des pistes d'amélioration de l'échantillonnage principal aient été exploitées.

Venons-en maintenant au tirage complémentaire, qui est le plus délicat. Dans le cadre de l'échantillonnage rénové, il a été proposé à toutes les régions de définir deux zonages (mais pas davantage) pour lesquels les partenaires du tourisme ont montré un intérêt tout particulier à obtenir des estimations zone par zone. On les appelle 'zonages prioritaires'. Certaines régions peuvent se contenter d'un seul zonage prioritaire. Concrètement, chaque zone est définie par un ensemble de communes. Les zones d'un zonage prioritaire constituent toujours une partition parfaite de la région. On considère alors l**'intersection** de ces deux zonages prioritaires, laquelle conduit à une partition fine de la région formant un découpage que l'on pourrait convenir d'appeler « découpage atomique ».

Le schéma ci-dessous donne l'exemple d'une région découpée selon un zonage horizontal et un zonage vertical (imaginons une logique respectivement est-ouest et sud-nord). Chaque case (i,j) est une intersection élémentaire de ces deux zonages (case hachurée) et l'ensemble de ces cases constitue le découpage atomique de la région.

		ZONAGE 2			
		1	…	j	J
ZONAGE 1	1				
	…				
	i			(i,j)	
	…				
	I				

Le nombre total d'hôtels en case (i,j) est $N_{i,j}$ (il peut être nul dans certaines cases). Par exemple, une dimension peut distinguer des pays touristiques, et l'autre un ensemble d'agglomérations que l'on veut surreprésenter, le complément de ces agglomérations dans la

région formant la dernière modalité du second zonage (il n'y a donc pas mécaniquement autant de dimensions qu'il y a de territoires de diffusion).

Cette approche offre une grande souplesse, car les deux zonages prioritaires peuvent être complètement indépendants. Ils peuvent être aussi de natures très différentes. Par exemple si on dispose d'un zonage de niveau 'pays' en cinq zones (zonage prioritaire N°1) et que l'on veut par ailleurs distinguer à titre de second zonage prioritaire une (seule) commune qui se trouve être par exemple dans la zone-pays N°3 du premier zonage prioritaire, la matrice de taille (5,2) prend la forme suivante :

PAYS	Commune distinguée	Complément dans la région
1	0	$N_{1,2}$
2	0	$N_{2,2}$
3	$N_{3,1}$	$N_{3,2}$
4	0	$N_{4,2}$
5	0	$N_{5,2}$

Si on distingue une agglomération N°1 dans le pays 1 et une agglomération N°2 à cheval sur les pays 4 et 5 cela donne

PAYS	Agglo 1	Agglo 2	Complément dans la région
1	$N_{1,1}$	0	$N_{1,3}$
2	0	0	$N_{2,3}$
3	0	0	$N_{3,3}$
4	0	$N_{4,2}$	$N_{4,3}$
5	0	$N_{5,2}$	$N_{5,3}$

Une zone de « complément » est donc souvent indispensable pour réaliser une partition parfaite du territoire régional. Le tirage complémentaire doit par ailleurs tenir compte d'une spécificité : au-delà des critères d'exhaustivité nationaux, on offre aux régions la possibilité de désigner des hôtels à enquêter de manière certaine, donc avec une probabilité de sélection égale à 1. On les nomme hôtels « exhaustifs régionaux ».

L'Insee et ses partenaires en région se sont entendus pour produire un échantillon final formé par D_i^1 hôtels en zone i du premier zonage et D_j^2 hôtels en zone j du second zonage. Ces effectifs doivent (évidemment) être inférieurs aux tailles des populations associées présentes dans la base de sondage. Si ce n'est pas le cas, la procédure informatique le détecte et diminue autoritairement les effectifs demandés D_i^1 et D_j^2.

La stratégie de tirage complémentaire va ensuite rechercher des allocations $n_{i,j}$ associées aux cases (i, j). Pour ancrer ces valeurs, il faut définir une allocation par case traduisant une situation hautement souhaitable, que l'on appellera allocation-cible. Le choix de la valeur cible de l'allocation par « case » de la matrice s'est porté sur l'allocation proportionnelle au

nombre total d'hôtels de la case, que l'on notera $\tilde{N}_{i,j}$. Soit $\tilde{n}_{i,j}$ cette valeur-cible dans la case (i,j). Si on note n la taille d'échantillon régionale globale de l'échantillon complémentaire (égale à la taille de l'échantillon global moins la taille de l'échantillon principal), on a

$$\tilde{n}_{i,j} = n \cdot \frac{\tilde{N}_{i,j}}{\sum_{i,j} \tilde{N}_{i,j}}$$

L'option de ventilation proportionnellement au nombre total de chambres, qui était possible dans le tirage principal, ne l'est plus ici. Primo, avec une allocation basée sur le nombre total d'hôtels, on augmente sensiblement les chances de pouvoir trouver une allocation par case qui respecte les tailles d'échantillon demandées par zone de chaque zonage. Secundo, on sera plus en cohérence avec la stratégie d'échantillonnage stratifié qui sera appliqué par la suite au sein des cases (cf. infra). On précise que le choix a été fait de ne pas soustraire du calcul des $\tilde{N}_{i,j}$ les hôtels exclus et de ne pas soustraire non plus les hôtels « mauvais répondants », quoi qu'il arrive. De même, les hôtels tirés dans l'échantillon principal sont bien inclus dans ce dénombrement. Cette stratégie a paru la meilleure parce que les exclus comme les « mauvais répondants » font partie du champ de l'enquête, et c'est bien entendu la même chose avec les hôtels de l'échantillon principal. En revanche, les hôtels exhaustifs régionaux sont exclus de $\tilde{N}_{i,j}$ parce qu'ils constituent une population spécifique à part qui n'a pas vocation à être échantillonnée.

Le tirage principal s'est déroulé bien évidemment sans tenir compte du découpage atomique. On **constate** qu'il a produit certaines allocations par case (simple comptage *a posteriori*). A ces allocations, on ajoute les effectifs d'hôtels considérés comme exhaustifs au niveau régional. Cela conduit à des allocations $n^0_{i,j}$ pour chaque case (i,j) du découpage atomique. On en déduit les effectifs déjà tirés par zone de chaque zonage : $n^0_i = \sum_j n^0_{i,j}$ pour le premier zonage et $n^0_j = \sum_i n^0_{i,j}$ pour le second zonage. Si on soustrait ces effectifs aux demandes des partenaires exprimées par zone, respectivement D^1_i et D^2_j, on obtient les tailles d'échantillon à respecter au niveau des marges de la matrice pour définir le tirage complémentaire :

$$\bar{n}^{(1)}_i = D^1_i - n^0_i \quad \text{et} \quad \bar{n}^{(2)}_j = D^2_j - n^0_j$$

Ces allocations marginales doivent évidemment être positives (au pire nulles). Plus la taille de l'échantillon principal est grande, plus le risque d'obtenir une valeur négative est grand : le tirage complémentaire devient impossible lorsque la taille principale dépasse un certain seuil, seuil qu'il convient de trouver empiriquement. Quoi qu'il arrive, on ne sous-échantillonne **jamais** dans l'échantillon principal : si celui-ci, s'avère « trop important » en taille en ce sens où il excède certaines demandes D^1_i ou D^2_j après prise en compte des exhaustifs régionaux, il conserve sa taille sans être modifié (ou alors on recommence toute la procédure !). Dans ce cas, le processus augmente en conséquence les demandes D^1_i et/ou D^2_j afin qu'elles coïncident avec les valeurs constatées n^0_i et/ou n^0_j. Cette phase d'adaptation va au-delà de la simple augmentation de taille dans une unique zone, car de fait elle augmente la taille

d'échantillon global régional et impose donc de corriger l'ensemble des tailles requises dans les zones des zonages non concernés par ce phénomène de saturation, cela afin de garantir la cohérence numérique. Ce processus n'est *in fine* acceptable que si l'augmentation de la taille de l'échantillon régional qui en résulte reste raisonnable : au-delà d'un certain seuil d'augmentation, il faut plutôt chercher à baisser la taille de l'échantillon principal (ou renégocier avec les partenaires la ventilation de l'échantillon régional entre les zones !).

Il reste un aspect à traiter car il faut que le complément d'échantillon restant à tirer dans chaque case (i, j) au titre de l'échantillonnage complémentaire soit en taille inférieur ou égal à la taille de la base de sondage encore disponible. En effet le tirage complémentaire a lieu dans la base de sondage expurgée des hôtels exhaustifs régionaux, des hôtels exclus, des « mauvais répondants » (si cette option est choisie) mais surtout et essentiellement expurgée de tous les hôtels constituant l'échantillon principal. On dispose donc en case (i, j) d'une réserve de taille $N_{i,j}$ sensiblement plus petite que la base de sondage expurgée des seuls hôtels 'exhaustifs' et dont on ne contrôle pas bien la structure : en effet, les hôtels exhaustifs régionaux peuvent être concentrés dans certaines cases, de même que les mauvais répondants. La partie non exhaustive de l'échantillon principal est pour sa part répartie de manière probablement plus favorable si le tirage principal a utilisé l'option d'allocation proportionnelle au nombre d'hôtels mais en revanche de manière probablement plus anarchique s'il a utilisé l'option de tirage proportionnel, au nombre total de chambres. Plus accessoirement, l'optimisation de type Neyman peut aussi contribuer à déformer un peu les structures. On récapitule les variables du problème :

- $\tilde{n}_{i,j}$ désigne l'effectif-cible dans le croisement des zones i et j
- $N_{i,j}$ désigne la réserve disponible dans le croisement des zones i et j
- $\bar{n}_i^{(1)}$ est l'effectif à tirer au titre du complément dans la zone i du premier zonage
- $\bar{n}_j^{(2)}$ est l'effectif à tirer au titre du complément dans la zone j du second zonage

En pratique, on constate que de nombreux $N_{i,j}$ sont nuls. Le programme à résoudre consiste à trouver de nouvelles allocations par case $n_{i,j}$ aussi proches que possible des cibles $\tilde{n}_{i,j}$ mais respectant les tailles d'échantillon demandées par zone et compatibles avec les réserves disponibles dans chaque case, tout en restant positives, autrement dit :

$$\text{Minimiser } \sum_{i,j} \left(n_{i,j} - \tilde{n}_{i,j}\right)^2$$

Sous les contraintes :

$$\forall i \quad \sum_j n_{i,j} = \bar{n}_i^{(1)}$$

$$\forall j \quad \sum_i n_{i,j} = \bar{n}_j^{(2)}$$

$$\forall (i, j) \quad n_{i,j} \leq N_{i,j}$$

$$\forall (i, j) \quad 0 \leq n_{i,j}$$

La situation s'avère délicate à chaque fois qu'il existe un domaine (une zone d'un zonage) pour lequel la taille-cible diffère de la taille que produirait naturellement un tirage à

probabilités égales, et ce d'autant plus intensément que la différence entre les deux tailles est importante.

La programmation de cette optimisation sous contraintes a été assurée par le package *limSolve* de R. Le reste du programme est développé en SAS, ce qui nécessite un échange entre les univers SAS et R. Il n'y a pas de solution si et seulement si le système de contraintes est irréalisable (si le système de contraintes n'est pas vide, on montre qu'il y a nécessairement une unique solution au programme d'optimisation[2]). Cette situation survient parfois car la structure de la réserve disponible $N_{i,j}$ peut être très déséquilibrée, s'il y a dans certaines cases une accumulation d'hôtels déjà tirés (exhaustifs ou non), exclus et/ou de 'mauvais répondants' due à un mauvais hasard ou à des configurations défavorables. Il semble que le risque existe d'autant plus qu'il y a davantage de croisements vides entre les zones ; ce sera le cas si les zonages ont une certaine tendance à s'emboiter. Le contexte de deux zonages indépendants sera donc plus favorable. S'il y a un emboitement complet des deux zonages (un des zonages inclus dans l'autre), c'est qu'en réalité on peut entièrement décrire la question avec un seul zonage.

Dans ce dernier cas, ou plus simplement s'il n'y a qu'un seul zonage prioritaire défini à l'origine, le problème n'a qu'une dimension. La case élémentaire est alors indexée par h, les allocations cibles sont les $\tilde{n}_h$, les réserves disponibles sont les N_h. La différence essentielle avec le cas du croisement de deux zonages porte sur la contrainte d'égalité qui impose que la somme des allocations complémentaires finales n_h soit égale à la taille de l'échantillon complémentaire global régional, soit $\bar{n}$. De ce fait, l'optimisation aura pour objectif de réallouer l'allocation initiale entre les zones en respectant la taille globale $\bar{n}$: si la réserve N_h est insuffisante dans une zone, alors le programme augmentera en contrepartie la taille d'échantillon n_h dans une autre. C'est un peu dommage évidemment pour la qualité de la diffusion dans cette zone dont la réserve s'avère insuffisante et qui sera donc dégradée par cette perte, mais l'effectif régional $\bar{n}$ en revanche ne sera pas modifié (de fait, d'autres zones bénéficieront du report). Le programme devient

$$\text{Minimiser } \sum_h (n_h - \tilde{n}_h)^2$$

Sous les contraintes :

$$\sum_h n_h = \bar{n}$$

$$\forall h \quad n_h \leq N_h$$

$$\forall h \quad 0 \leq n_{ih}$$

Sur le plan informatique, on réutilise le code SAS conçu pour le cas de deux zonages mais dans la table des zonages on décrit un des zonages comme étant constitué d'une unique zone qui se trouve être la région toute entière. En revanche, le programme R est spécifique.

[2] Minimisation d'une fonction convexe sur un ensemble compact (fermé + borné) non vide de R^p : alors la fonction est bornée inférieurement est atteint sa borne minimale, c'est-à-dire qu'il y a toujours une unique solution.

Revenons au cas de deux zonages. Dans le cas d'absence de solution, il faut en premier lieu s'assurer de la cohérence des demandes par zone. Souvent, cela vient du fait que deux zones appartenant aux deux zonages respectifs sont en réalité une seule et même zone et que les tailles d'échantillon final réclamées respectivement dans les deux zones ne sont pas égales. Plus généralement, le phénomène peut se produire si une zone du premier zonage coïncide avec un ensemble de zones du second zonage et que la taille de la première diffère de la somme des tailles des secondes. Si on ne parvient pas, après examen 'à vue' de la structure des zones et des tailles d'échantillon réclamées, à détecter la source de l'incohérence, c'est qu'elle est plus subtile et probablement indétectable sans une procédure complexe. Il n'y a alors guère d'autre solution que d'agir sur les leviers déjà signalés : taille de l'échantillon principal évidemment (à réduire), répartition demandée de l'échantillon par zone (demandes peut-être excessives, donc irréalisables, dans certaines zones), traitement des « mauvais répondants » peut-être trop contraignant, changement du type d'allocation de l'échantillon principal, éventuellement (bien que cela soit marginal probablement) modification du paramétrage permettant de sélectionner les strates à allocation optimale.

En cas d'échec de l'optimisation, le package d'optimisation *limsolve* ne bloque pas mais produit des allocations fantaisistes, en général en partie négatives. On produit une sortie très utile pour diagnostiquer l'origine des blocages qui peuvent être constatés à l'issue de l'optimisation. Il s'agit d'une simple matrice portant en ligne les zones d'un des zonages et en colonne les zones de l'autre zonage. A l'intersection de la ligne i et de la colonne j on trouve la taille de la base de sondage encore disponible, c'est-à-dire le $N_{i,j}$. L'examen de la matrice permet - sauf peut-être si elle est d'une taille considérable - de repérer « à l'œil » la raison pour laquelle le domaine des contraintes n'a pas de solution. Cela est d'autant plus simple que la matrice est remplie de zéros, ce qui est le cas lorsque les zones des deux zonages respectifs ne se recoupent que de manière très occasionnelle (à l'expérience il s'avère qu'il y a souvent des inclusions entre les deux zonages).

On termine par quelques éléments sur le tirage proprement dit de l'échantillon complémentaire. On dispose désormais d'une allocation (réalisable) à assurer au niveau de chaque case de la matrice. Lorsqu'on se place dans une case, c'est-à-dire une sous-population d'hôtels bien particulière définie par un critère géographique (un ensemble de communes), le tirage des hôtels a lieu par stratification avec sondage aléatoire simple dans chaque strate. La stratification utilisée est une stratification simplifiée croisant un groupement d'espace touristique, la catégorie et le type d'hôtel. L'allocation choisie est systématiquement l'allocation proportionnelle au nombre total d'hôtels dans la strate. Les strates concernées sont en général en petit nombre (quelques strates par case). On passe ensuite une procédure d'arrondi des allocations. Comme à ce stade on traite souvent de petites ou très petites tailles par strate, on trouvera fréquemment des tailles arrondies à zéro, que l'on laisse en l'état : la situation est beaucoup moins préoccupante que lors de l'échantillonnage principal, s'agissant d'un complément *a priori* restreint et qui de toute façon laisse une probabilité de sélection globale strictement positive à chaque hôtel de la base (du fait de l'existence du tirage principal).

3 Le processus d'imputation massive

L'enquête mensuelle auprès des hôtels s'appuie sur un échantillon tiré de manière aléatoire, mais le système d'estimation est fondé sur une prédiction des valeurs des variables d'intérêt pour tout hôtel qui n'a pas fourni de réponse aux questions posées. Cela impose de formuler

des hypothèses de comportement, donc d'adopter une approche d'estimation « par modèle ». L'utilisation de modèles pose des difficultés particulières :

i) Il y a toujours un risque de mauvaise spécification du modèle car un modèle reste une hypothèse simplificatrice de la réalité. L'estimation finale est donc biaisée.

ii) On est confronté sans cesse à des situations de compromis, dans un contexte de masse considérable de données numériques : grand nombre de modèles concurrents, plusieurs variables d'intérêt, plusieurs régions, douze mois de traitement. Ce contexte très diversifié dans le temps et dans l'espace complexifie la stratégie à adopter pour les choix de modèle.

iii) Les phénomènes à mesurer relèvent de concepts simples mais dépendent d'effets multiples. Des phénomènes locaux sont en jeu et des caractéristiques propres aux hôtels doivent tenir un rôle important dans la fréquentation (par exemple la qualité de l'accueil). Les variables disponibles pour expliquer les phénomènes à mesurer sont limitées : catégorie d'hôtel, type d'hôtel, taille et bien entendu localisation (au travers de l'espace touristique).

iv) Fixer la liste des variables explicatives n'est pas évident : la théorie de la régression dit que plus on implique de variables explicatives, meilleur est l'ajustement. Néanmoins, la prédiction perd une partie de sa crédibilité si on augmente trop le nombre des régresseurs. C'est pourquoi on cherche en général des modèles « parcimonieux ». Se pose aussi la question du traitement qu'il faut réserver aux variables jugées non significatives.

Les modèles linéaires ne comprenant que des effets fixes du type $Y_i = X_i \cdot B + \varepsilon_i$ où $E\varepsilon_i = 0$ et $V\varepsilon_i = \sigma^2$ expliquent une variable quantitative - par exemple le taux d'occupation des chambres ou le nombre total de clients reçus dans le mois. On impute pour chaque hôtel l'espérance mathématique (estimée) de la variable d'intérêt soit $\hat{Y}_i = X_{i.} \cdot \hat{B}$. Les modèles linéaires avec effets aléatoires les plus simples expliquant une variable quantitative s'écrivent $Y_i = X_i \cdot B + v_g + \varepsilon_i$ où g désigne la zone géographique dans laquelle se situe l'hôtel i avec $E(v_g) = 0$ et $V(v_g) = \sigma_v^2$. Les effets aléatoires v_g traduisent une spécificité de la zone g, donc un effet géographique propre à chaque zone. Ils sont le plus souvent deux à deux indépendants. Avec ce modèle, il n'y a pas d'effet géographique en espérance mathématique mais en revanche on a $cov(Y_i, Y_j) = \sigma_v^2$ dès que i et j appartiennent à la même zone géographique. Si on veut introduire une corrélation entre ces effets locaux, on utilise un distancier pour calculer la distance $d_{g,h}$ entre les zones g et h puis on postule une covariance de la forme $cov(v_g, v_h) = \lambda \cdot exp(-d_{g,h})$. La stratégie de minimisation de l'erreur quadratique moyenne de prédiction d'un total consiste à prédire l'effet aléatoire v_g et à l'ajouter à l'estimation des effets fixes, soit $\hat{Y}_i = X_{i.} \cdot \hat{B} + \hat{v}_g$. Certaines variables d'intérêt qui n'ont pas une nature continue ne se traitent pas au moyen de modèles linéaires mais en utilisant des modèles linéaires généralisés, voire des modèles linéaires mixtes généralisés. C'est le cas du dénombrement des nuitées étrangères par exemple. Ainsi Y_i va suivre une loi paramétrée plus ou moins complexe (loi de Poisson, loi binomiale-négative, loi binomiale,…) $f(EY_i) = X_i \cdot B + v_g$ où f est une fonction non-linéaire (un logarithme, une fonction l.ogit, …) et v_g un effet aléatoire géographique. La prédiction optimale se construit selon $\hat{Y}_i = f^{-1}(X_{i.} \cdot \hat{B} + \hat{v}_g)$. Les modèles à effets aléatoires peuvent néanmoins conduire à un

échec. Certains modèles prédisent des effectifs pouvant prendre assez souvent des valeurs nulles. Par exemple c'est le cas du nombre de nuitées étrangères. On utilise alors des modèles spécifiques, comme les modèles Tobit ou les modèles dits « Zero-inflated ».

L'impact de la géographie (espace touristique) peut être appréhendé au travers d'effets fixes, auquel cas on trouve dans certaines régions des zones qui ne comprennent qu'un (très) petit nombre d'hôtels répondants. Afin de ne pas trop dégrader la précision des coefficients de régression associés, lorsque le nombre de répondants n'est pas suffisant, on a mis en place un système de regroupement des espaces touristiques. Il est effectué de manière automatique par le programme en utilisant un distancier. Certains modèles testés comprennent des interactions formées en croisant deux variables parmi les suivantes : zonage, catégorie d'hôtel, type d'hôtel, taille (soit 6 configurations de croisement). Pour éviter de prendre en compte des interactions construites sur un trop petit nombre d'hôtels répondants, on ne sélectionne les interactions que si elles concernent un nombre d'hôtels répondant supérieur à un certain seuil que l'on peut choisir (par exemple entre 5 et 10). Cela permet de limiter le nombre d'interactions intervenant dans le modèle. La plupart des modèles s'appuient sur des régresseurs prédéterminés, mais on a aussi testé l'application de méthodes de sélection optimale de régresseurs de type *stepwise*. Quel que soit le modèle considéré, on peut choisir de ne sélectionner que les variables explicatives dites « significatives » (en utilisant la *p-value* associée au test de nullité du coefficient de régression associé).

Lorsque c'est possible, on incorpore dans l'information explicative la variable d'intérêt représentant la fréquentation d'un mois passé, pour tirer profit des corrélations temporelles de l'activité. Le 'passé' est représenté par le même mois que le mois d'imputation considéré au titre de l'année précédente. Cette variable 'passé' s'avère à fort pouvoir explicatif mais elle est hélas limitée dans son application. En effet, la prédiction avec un tel modèle n'est possible que si et seulement si l'hôtel non-répondant (échantillonné ou non) a fourni l'an passé une réponse à la variable d'intérêt. Cela ne concerne qu'un effectif minoritaire d'hôtels parmi les hôtels non-répondants le mois d'imputation. En résumé, on trouve quatre situations :

Réponse à la variable d'intérêt		Niveau d'implication de l'hôtel
Par le passé	Au titre du mois d'imputation	
OUI	OUI	Ajustement du modèle avec régresseur « passé » (si non atypique)
NON	OUI	Ajustement du modèle sans régresseur « passé » (si non atypique)
OUI	NON	Imputation à partir du modèle avec régresseurs passé
NON	NON	Imputation à partir du modèle sans régresseur passé

In fine, la variable imputée est, soit le prédicteur issu du modèle **avec** passé si l'hôtel a répondu par le passé, soit le prédicteur issu du modèle **sans** passé si l'hôtel n'a pas répondu par le passé. Pour chaque variable expliquée traitée isolément, on produit différents modèles concurrents et on fournit, pour chaque hôtel de la région, un ensemble de ratios estimés qui permettent d'en déduire immédiatement la variable d'intérêt imputée. Différents indicateurs sont produits pour juger de la pertinence des modèles concurrents. On calcule en particulier les carrés des résidus $(Y_i - \hat{Y}_i)^2$ sur les hôtels répondants, puis on les somme. Ce type d'indicateur synthétique est néanmoins sensible au nombre de régresseurs impliqués.

Bibliographie

[1] Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press.
[2] Cameron, A. et Trivedi, P. (2013), *Regression Analysis of Count Data*, Cambridge University Press.
[3] Davison, A. (2009), *Statistical Models*, Cambridge University Press.
[4] Mc Culloch, C. , Searl, S. et Neuhaus, J. (2008), *Generalized , Linear, and Mixed Models*, Wiley.