Estimation du trafic de courrier distribué en France métropolitaine par sondage indirect

Pauline Lardin-Puech[1]

[1] *La Poste, 2 boulevard Newton, Champs-sur-Marne 77453 Marne-La-Vallée Cedex 2*
pauline.puech@laposte.fr

Résumé. Depuis une vingtaine d'années, La Poste a mis en place un plan de sondage à deux degrés afin d'estimer le trafic de courrier distribué en France métropolitaine pendant chaque trimestre. La Poste a beaucoup investi dans ses systèmes d'information et ses machines de tri depuis quelques années, afin de pouvoir adapter le plan de tri en fonction du trafic. Suite à ces investissements, les enquêteurs ont de moins en moins de temps pour réaliser l'étude par sondage et notre base de sondage n'est plus en adéquation avec la réalité sur le terrain. La méthode actuelle devient donc de plus en plus difficile à mettre en place et nous devons envisager une nouvelle méthodologie d'observation. Dans ce résumé, nous allons exposer le plan de sondage qui va être déployé début 2015. Celui-ci est basé sur un plan de sondage indirect : nous utiliserons un échantillon d'adresses pour capter des organisations locales sur lesquelles un échantillon d'objets sera observé.
Mots-clés. Sondage indirect, MGPP, transitivité, application, estimateur non biaisé.

1 Présentation du contexte de l'étude

Chaque jour, La Poste distribue des millions d'objets de correspondance en France. Afin d'améliorer la connaissance de son réseau postal et ses offres tarifaires, La Poste a décidé de mettre en place, il y a une vingtaine d'années, une étude par sondage qui permet d'estimer le trafic de courrier distribué en France métropolitaine (hors Corse), ainsi que sa structure (prix, poids, dimension, catégorie d'objets, etc.). Cette étude est notamment utilisée pour estimer le volume de courrier distribué sur une période T, pour estimer les taux d'évolution (nombre d'objets d'une caractéristique traités sur une période T de l'année N, comparé au nombre d'objets de même caractéristique sur la même période T de l'année $N-1$), et pour mesurer la qualité de service (délai d'acheminement du courrier).

Actuellement, cette étude est basée sur un plan de sondage à deux degrés. Au début de chaque trimestre, nous sélectionnons un échantillon de tournées de facteur. Ensuite, on affecte aléatoirement une date d'observation à chaque tournée de manière à obtenir un équilibrage temporel. Ainsi, les observations sont réparties de manière équitable en fonction des jours de la semaine, des mois et des positions de la date dans le mois. Cela permet à la fois de lisser les observations et les charges correspondantes mais aussi de prendre

en compte dans le plan les phénomènes temporels sous-jacents (saisonnalités très fortes à la fois en volume de trafic, mais aussi en type de courrier distribué). Enfin, le jour de l'observation, des enquêteurs se déplacent dans le bureau, afin d'observer un échantillon d'objets dans la tournée sélectionnée.

Ce plan de sondage repose donc sur une base de tournées de facteur, qui est considérée comme fixe pendant tout le trimestre. Depuis quelques années, La Poste a beaucoup investi dans ses systèmes d'information et ses machines de tri afin de pouvoir adapter son plan de distribution aux fluctuations journalières du volume de courrier. Suite à ces investissements, les enquêteurs ont de moins en moins de temps pour réaliser l'étude et notre base de sondage n'est plus en adéquation avec la réalité du terrain. Le plan de sondage actuel devient donc de plus en difficile à mettre en place et statistiquement non maîtrisable. Un nouveau plan de sondage, indépendant de l'organisation des bureaux, doit être mis en place. Nous avons décidé d'utiliser un plan de sondage indirect : à partir d'un échantillon d'adresses, nous allons capter les objets distribués par une organisation locale un jour donné. Au début des années 2000, des premiers travaux basés sur le sondage indirect des ilots INSEE avaient déjà été envisagés par Dessertaine & Fluteaux (2004). Ils ont été abandonnés du fait de la grande difficulté, sur le terrain, d'élaborer la matrice de liens entre la population des ilots d'un côté et des tournées de facteurs de l'autre. La Poste s'est entre temps dotée d'une base d'adresses de grande qualité, mise à jour en « temps réel » car utilisée pour la mécanisation du tri. La mise en œuvre des liens entre cette population et la population d'étude devient, de fait, beaucoup plus facile.

Dans la suite de ce document, nous allons présenter le nouveau plan de sondage que nous avons initialement envisagé, ainsi que les difficultés rencontrées lors des premiers tests. Enfin, nous exposerons le plan de sondage finalement retenu.

2 Présentation du plan de sondage initialement envisagé

L'objectif de notre étude est d'estimer le nombre d'objets, $N_P(T)$, de catégorie P distribués en France métropolitaine (hors Corse) sur une période donnée T. Dans la suite de ce document, nous appellerons "Sortie Jour" l'ensemble des adresses distribuées par le même facteur, un jour donné. Soit U_T l'ensemble des jours ouvrés de la période d'observation T, soit U_{C_t} l'ensemble des Sorties Jour le jour t. On note U_{i,C_t} l'ensemble des objets distribués le jour t dans la Sortie Jour i. Si l'objet k est de catégorie P, on pose $\mathbb{1}_{k\in P} = 1$ et 0 sinon.

Ainsi, nous cherchons à estimer :

$$N_P(T) = \sum_{t\in U_T} \sum_{i\in U_{C_t}} \sum_{k\in U_{i,C_t}} \mathbb{1}_{k\in P} \tag{1}$$

Etant donné que le périmètre de distribution d'un facteur varie en fonction du jour de la semaine, nous avons dû chercher une base de sondage, qui soit fixe sur notre période d'étude, et qui nous permette de capter ces variations journalières. La solution que nous avons retenue est d'utiliser la base d'adresses qui permet d'alimenter les plans de tri des machines. Ainsi, à partir d'un échantillon d'adresses, nous allons capter notre population cible, l'ensemble des Sorties Jour de notre période d'étude T. Nous allons donc mettre en place un plan de sondage indirect (Lavallée (2002), Lavallée (2007) et Deville & Lavallée (2006)). Soit U_A la population des adresses jour de taille N_A, où N_A est le nombre d'adresses en France métropolitaine multiplié par le nombre de jours ouvrés dans notre période d'étude T.

Plan de sondage initial (cf. figure 1)

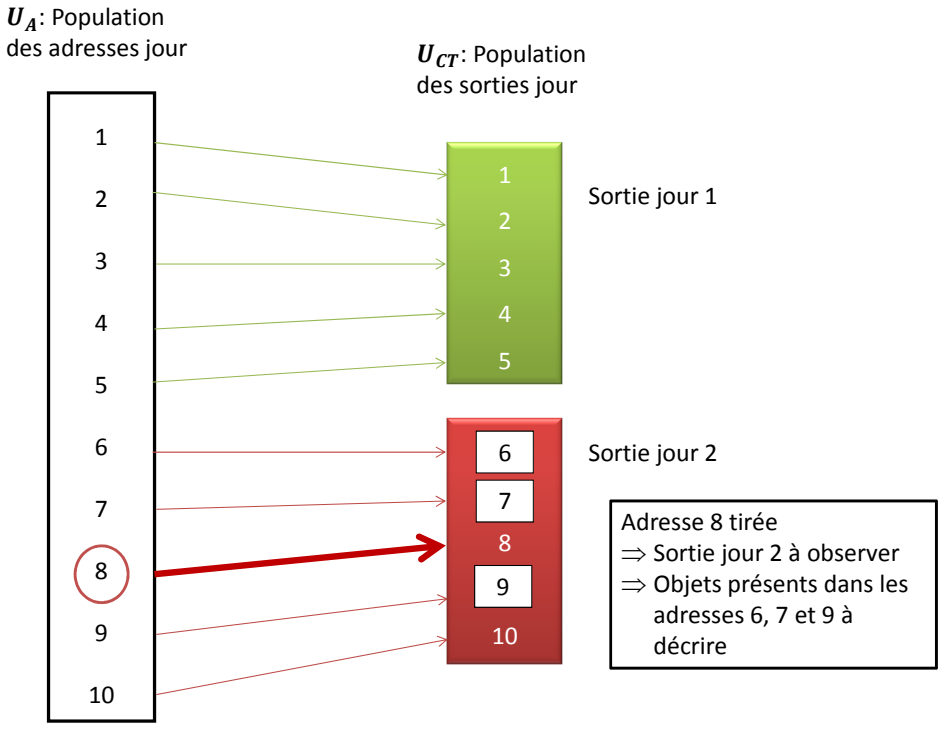

FIGURE 1 – Plan de sondage initialement envisagé.

1. Tirage en début de mois d'un échantillon d'adresses. Afin d'améliorer la précision de nos estimations et de répartir la charge de travail sur l'ensemble du territoire, nous allons mettre en place une stratification. Celle-ci sera construite à partir des données communales mise à disposition par l'INSEE (base logement, base couple-famille-ménage, base appartenance géographique des communes).
2. Affectation de manière aléatoire d'une date d'observation. Un équilibrage temporel sera effectué afin de lisser la charge de travail et de capter les phénomènes temporels.
3. Pour chaque adresse tirée, des enquêteurs vont être envoyés dans le bureau où se trouve l'adresse tirée le jour de l'observation.
4. Les enquêteurs identifient la Sortie Jour à observer.
5. Afin de pouvoir reconstituer la matrice de liens entre la population d'adresses jour U_A et la population des Sorties Jour U_{C_T}, les enquêteurs vont recenser, à l'aide d'un outil développé en interne, l'ensemble des adresses distribuées par la Sortie Jour captée à l'étape précédente. Ces adresses figurent sur la façade du casier de tri des facteurs.
6. Parmi les adresses validées, un échantillon d'adresses est sélectionné à l'aide d'un outil développé en interne. Le type de plan (tirage systématique, sondage aléatoire simple, sondage stratifié, etc.) sera défini en fonction de l'information auxiliaire disponible sur les adresses (productivité, largeur de la séparation sur le casier de tri, etc.). En l'absence d'information auxiliaire, un sondage aléatoire simple sera mis en place.
7. Les enquêteurs décrivent l'ensemble des objets présents dans les adresses sélectionnées.

La correspondance entre notre population des adresses jour U_A de taille N_A, et notre population des Sorties Jour U_{C_T} sur la période d'observation T, peut être représentée par une matrice de liens $\Theta_{AC_T} = [\theta_{ji}^{AC_T}]$ de taille $N_A \times N_{C_T}$. Dans le cas où deux unités $j \in U_A$ et $i \in U_{C_T}$ ne sont pas liées, nous fixons $\theta_{ji}^{AC_T} = 0$. Dans le cas contraire, nous avons choisi de fixer $\theta_{ji}^{AC_T} = 1$. Ce choix influe sur la précision de nos estimations (pour de plus amples détails, se référer à Deville & Lavallée (2006) et Lavallée & Labelle-Blanchet (2013)).

Nous sélectionnons l'échantillon s_A d'adresses jour de taille n_A à l'aide d'un plan de sondage stratifié. Soit $\mathbf{\Pi}_A = diag(\pi^A)$ la matrice diagonale de taille $N_A \times N_A$ contenant le vecteur $\pi^A = \{\pi_1^A, ..., \pi_{N_A}^A\}$ de probabilités de sélection des unités j dans U_A. Soit $\mathbf{1}_A$ le vecteur colonne de 1 de taille N_A. Soit $\mathbf{T}_A = diag(\mathbf{t}^A)$ la matrice diagonale de taille $N_A \times N_A$ contenant le vecteur $\mathbf{t}^A = \{t_1^A, ..., t_{N_A}^A\}$, où $t_j^A = 1$ si $j \in s_A$, et 0 autrement.

Ainsi, le trafic total distribué sur la période d'observation T, $N_P(T)$, sera estimé par :

$$\hat{N}_P^1(T) = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\Theta}_{AC_T} \hat{N}_P^{C_T} \tag{2}$$

où $\tilde{\mathbf{\Theta}}_{AC_T} = \mathbf{\Theta}_{AC_T}[diag(\mathbf{1}'_A\mathbf{\Theta}_{AC_T})]$ et où $\hat{N}_P^{C_T}$ est le vecteur des trafics de catégorie P estimés dans les sorties jours par le biais de l'échantillonnage d'objets du deuxième degré. Cette écriture matricielle mise en place par Deville & Lavallée (2006) est le reflet de la méthode généralisée du partage des poids (MGPP) présentées dans Lavallée (2002).

$\hat{N}_P^1(T)$ est un estimateur sans biais de $N_P(T)$ (Lavallée (2002) et Deville & Lavallée (2006)).

3 Présentation du plan de sondage final

Les premiers tests réalisés sur le terrain nous ont montré que le temps de validation des adresses est très long. Il faut en moyenne 40 minutes pour valider environ 400 adresses. Si nous choisissons cette méthodologie, nous allons trop retarder le départ du facteur en tournée, les observations ne se feront pas dans de bonnes conditions et nous augmenterons le risque d'un mauvais recensement des adresses présentes sur la Sortie Jour captée. Ce dernier point est très important. En effet, sans un recensement correct et exhaustif, nous serons dans l'incapacité de reconstituer notre matrice de liens $\mathbf{\Theta}_{AC_T}$ et donc de fournir une estimation de nos trafics. Nous avons donc dû trouver un compromis : nous allons utiliser une population intermédiaire U_B qui nous permettra de faire le lien entre les populations U_A et U_{C_T}. Etant donné qu'un casier de tri est constitué de cases et qu'à l'intérieur de chaque case se trouve une ou plusieurs adresses, nous avons décidé de prendre la population des cases de tri comme population intermédiaire.

Pour reconstituer le trafic total estimé sur une période donnée, nous aurions pu envisager de mettre en place d'autres méthodes :

- Mettre en place un sondage direct d'adresses et observer directement l'ensemble du courrier présents dans ces adresses.
- Mettre en place un sondage indirect : à partir d'un échantillon d'adresses, nous allons observer le courrier présent uniquement dans les cases identifiées par notre échantillon d'adresses.

Ces solutions sont certes beaucoup plus simples à mettre en œuvre mais leur mise en place nous auraient conduit à des coûts d'observation très élevés (très peu d'objets à observer par adresse ou par case) et à une très grande variance pour les objets rares dans notre population d'étude. La solution que nous allons mettre en place va nous permettre de mieux capter ces objets rares et de pouvoir continuer à sortir des statistiques sur les Sorties Jour (indicateur sur les coûts de distribution, sur les organisations mise en place à la distribution, etc.).

Principe du plan de sondage définitif (cf. Figure 2).

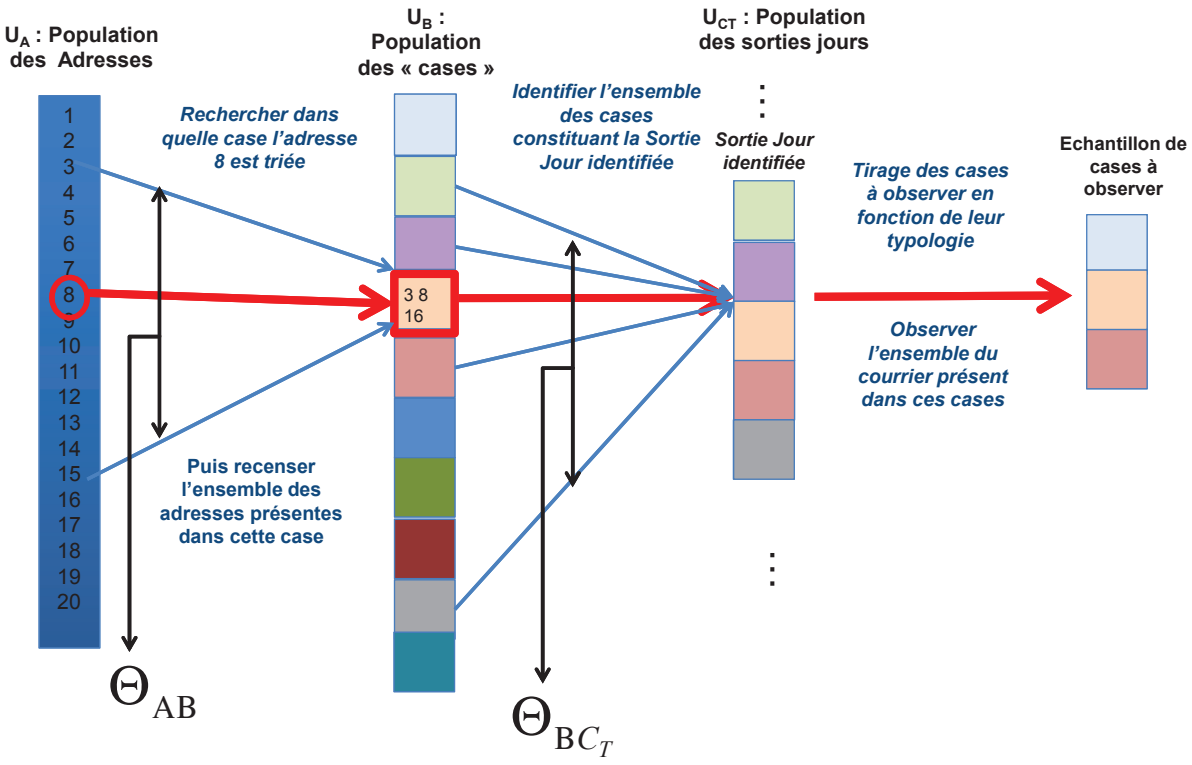

FIGURE 2 – Plan de sondage définitif

1,2 et 3. Les étapes 1, 2 et 3 sont identiques à celles présentées dans le plan de sondage initialement prévu.

4. Les enquêteurs identifient la case où se trouve l'adresse tirée j.
5. Afin de pouvoir reconstituer la matrice de liens Θ_{AB} entre la population d'adresses jour U_A et la population des cases U_B, les enquêteurs vont recenser l'ensemble des adresses présentes dans la case où se trouve l'adresse j.
6. Les enquêteurs identifient sur les casiers de tri, l'ensemble des cases qui vont sortir en distribution en même temps que la case identifiée à l'étape 4 le jour de l'observation. Cette identification va nous permettre de reconstituer la matrice de liens Θ_{BC_T}

entre la population des cases U_B et la population des Sorties Jour U_{C_T}. De plus, les enquêteurs renseigneront la typologie de chaque case (case de tri horizontale, verticale, productivité de la case, etc.) par le biais d'un outil informatique développé par nos soins. Ces informations nous permettrons de mettre en place la stratification du deuxième degré et d'améliorer nos précisions en utilisant les méthodes développées dans Lavallée & Labelle-Blanchet (2013).

7. Un module de l'outil informatique utilisé à l'étape 6 permettra de tirer un échantillon de cases à observer. Puis, les enquêteurs décrivent l'ensemble des objets présents dans les cases sélectionnées. Dans les cases très productives, un troisième degré pourra être mis en place.

Le trafic total distribué sur la période d'observation T, $N_P(T)$, sera alors estimé par :

$$\hat{N}_P(T) = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\Theta}_{AB} \tilde{\Theta}_{BC_T} \hat{N}_P^{C_T} \tag{3}$$

où $\tilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB}[diag(\mathbf{1}'_A \mathbf{\Theta}_{AB})]$, $\tilde{\mathbf{\Theta}}_{BC_T} = \mathbf{\Theta}_{BC_T}[diag(\mathbf{1}'_B \mathbf{\Theta}_{BC_T})]$ et où $\hat{N}_P^{C_T}$ est le vecteur des trafics de catégorie P estimés dans les Sorties Jour. Cet estimateur repose sur la propriété de transitivité exposée dans Deville & Lavallée (2006). Les outils informatiques que nous développons vont nous permettre de reconstituer l'ensemble des liens entre nos populations et ainsi d'obtenir des matrices de liens normalisées. Etant donné que la matrice $\tilde{\Theta}_{AB}\tilde{\Theta}_{BC_T}$ est une matrice de liens normalisée, $\hat{N}_P(T)$ est un estimateur sans biais de $N_P(T)$ (Deville & Lavallée (2006)).

Cette étude est actuellement en cours de déploiement auprès des 220 enquêteurs et devrait commencer début 2015. Certains points (stratification des adresses dans U_A, stratification des cases, etc.) seront affinés d'ici la fin de l'année 2014, en fonction des retours des premiers tests effectués actuellement sur le terrain.

Bibliographie

[1] Dessertaine, A. & Fluteaux, L. (2004), *Utilisation de la méthode généralisée du partage des poids dans le cadre des estimations de flux de courrier à La Poste* , Echantillonnage et méthodes d'enquêtes (Ed. P. Ardilly), Dunod, Paris.

[2] Deville, J.-C. & Lavallée, P. (2006), *Sondage indirect : Les fondements de la méthode généralisée du partage des poids*, Techniques d'enquête, 32, 185–196.

[3] Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids.* Éditions de l'Université de Bruxelles, Brussels.

[4] Lavallée, P. (2007). *Indirect Sampling.* New York : Springer.

[5] Lavallée, P. & Labelle-Blanchet, S. (2013), *Le sondage indirect appliqué aux populations asymétriques*, Techniques d'enquête, 39, 207–241.