Journal de la Société Française de Statistique
Vol. 152 No. 4 (2011)

Allocation optimale d'unités primaires pour un plan à plusieurs degrés : application à l'estimation de la fraude tarifaire grandes lignes à la SNCF

Title: Optimal allocation of primary units for multistage sampling: an application to tariff fraud for SNCF

Maguelonne Chandesris[1], Guillaume Chauvet[2] et Jean-Claude Deville[2]

Résumé : La SNCF réalise quotidiennement des contrôles auprès de ses clients afin de régulariser les voyageurs en situation irrégulière. Même si les contrôles sont très fréquents à bord des trains à réservation obligatoire, ils ne sont pas exhaustifs ; dans ce cas la fraude peut être opérante et non mesurée. Ce phénomène génère une perte financière importante que la SNCF souhaite évaluer avec précision en s'appuyant sur les données collectées lors des opérations de contrôle à bord des trains.
Le premier objectif de ce travail est de proposer une modélisation du mécanisme d'échantillonnage conduisant au contrôle, afin d'estimer la perte liée à la fraude tarifaire. La sélection de l'échantillon de voyageurs contrôlés est réalisée selon un plan à plusieurs degrés ; une difficulté particulière réside dans la définition des unités échantillonnées. Cette étape de modélisation permet de produire une estimation de la précision des estimations réalisées à l'aide des données historiques de régularisations. Le second objectif est de proposer un mécanisme de contrôle ultérieur optimisé, qui passe par le calcul d'une allocation de type Neyman pour un plan de sondage à plusieurs degrés.

Abstract: The SNCF performs daily controls on customers to identify and regulate the non-regular travelers. The controls are quite common on trains with compulsory reservation but are not exhaustive, so that the fraud may not be precisely measured. This phenomenon leads to significant financial loss that the SNCF wants to accurately estimate, based on data collected during controls on trains.
The first goal of this work is to propose a modeling of the sampling mechanism leading to the controled travelers, so as to estimate the loss associated with fraud. The controled passengers are selected through multistage sampling: a particular difficulty lies in the definition of the units sampled at the various stages. This modeling step enables to measure the accuracy of estimates, based on historical data collection. The second goal is to propose an optimized selection scheme for further controls, which requires the computation of a Neyman-type allocation for multistage sampling.

Mots-clés : allocation optimale, plan à plusieurs degrés
Keywords: multistage sampling, optimal allocation
Classification AMS 2000 : 62D05, 49M37

1. Introduction

A la SNCF, des moyens visant à lutter contre la fraude tarifaire sont déployés quotidiennement, pour autant ce phénomène reste par nature difficilement mesurable. Ce travail vise à fournir une méthodologie scientifique solide pour mieux quantifier ce phénomène à la fois en nombre (taux

1. Optimisation des Revenus, Direction Innovation et Recherche, SNCF, Paris, France.
E-mail : maguelonne.chandesris@sncf.fr
2. CREST (Ensai), Campus de Ker Lann, 35170 Bruz, France.
E-mail : chauvet@ensai.fr and E-mail : deville@ensai.fr

 ISSN: 2102-6238

de fraude) et en montant (perte financière associée). Pour cela, l'idée est de s'appuyer sur les données issues des contrôles quotidiens effectués à bord des trains.

Les données collectées consistent en un échantillon de voyageurs ayant fait l'objet d'un contrôle à bord, et sélectionnés selon un mécanisme de tirage à plusieurs degrés. La procédure de sélection de cet échantillon n'étant pas connue de façon détaillée, elle a fait l'objet d'une modélisation réaliste sous forme d'un plan de sondage probabiliste, sur la base des éléments connus de la procédure de sélection. Les principales notations ainsi que la modélisation du plan de sondage sont présentées en section 2. Dans le cas où un voyageur fait l'objet d'une régularisation pour fraude tarifaire, plusieurs variables le concernant sont relevées (origine et destination du voyage, tarif, classe, motif éventuel de fraude, ...). Cette information permet d'estimer les deux principaux indicateurs d'intérêt que sont le taux de fraude et le total de la somme non réglée par les fraudeurs (manque à gagner). Elle permet également de produire des estimations sur des domaines, afin notamment d'évaluer l'importance de la fraude selon les trains. L'estimation de ces indicateurs est discutée en section 3.

Si la précision de ces indicateurs est jugée insuffisante, notamment pour les domaines d'intérêt, la SNCF envisage de sélectionner un sur-échantillon afin de disposer d'estimations plus fiables. Une méthode permettant de calculer une allocation optimale pour cet échantillon complémentaire est présentée en section 4, avec une application sur la base des données collectées par la SNCF.

2. Notation et sélection de l'échantillon de voyageurs

Le parcours d'un voyageur est caractérisé par la donnée d'un train, circulant à une date et à une heure fixées. On appellera train-date la donnée d'un train et d'une date spécifique. En cas d'arrêts intermédiaires, le parcours d'un train-date se décompose en tronçons (intervalle entre deux gares desservies successivement). On parle alors de train-tronçon-date (ttd). Même s'il n'emprunte physiquement qu'un seul train pour effectuer son voyage, un voyageur emprunte un ou plusieurs train-tronçon-date. A chaque train-tronçon-date, on peut associer un certain nombre de caractéristiques : durée et longueur du tronçon, nombre de voyageurs à bord, etc. C'est à ce niveau que sont effectuées les opérations de contrôle à bord. Chaque régularisation émise à bord l'est dans un train-tronçon-date donné. Un voyageur empruntant généralement plusieurs train-tronçon-date, il est en pratique contrôlé dans au moins un train-tronçon-date de son voyage.

Dans la sélection puis l'exploitation de l'enquête, on distingue deux types d'unités. D'une part, le voyageur constitue l'unité élémentaire ou unité secondaire (dans la suite, on parlera simplement d'unité voyageur), unité pour laquelle plusieurs variables sont relevées si elle est échantillonnée. D'autre part, la donnée d'un train, d'un tronçon et d'une date constitue l'unité agrégée ou unité primaire. Dans la suite, on parlera simplement d'unité ttd. Le processus de sélection des voyageurs contrôlés est assimilé à un sondage à deux degrés, avec sélection au premier degré d'unités ttd, et sélection au second degré d'unités voyageur. On note U_I l'ensemble des unités ttd pour la période de l'enquête de contrôle. Chaque unité ttd se compose d'unités voyageur, et leur réunion donne la population totale U des unités voyageurs ayant emprunté le réseau SNCF pour la période considérée. Notons en particulier qu'une unité voyageur ne correspond pas à une personne, mais à la donnée d'un train, d'un tronçon, d'une date et d'une personne. Un même individu utilisant le réseau SNCF à deux dates différentes, ou sur deux trajets différents, est donc associé à deux unités voyageur.

 ISSN: 2102-6238

On assimile les données relevées sur la période étudiée à un échantillon sélectionné selon un plan à deux degrés. La population U_I des unités ttd est découpée en H strates notées $U_{I1},\ldots,U_{IH}$ de tailles respectives $M_1,\ldots,M_H$. La stratification permet de prendre en compte l'hétérogénéité des train-tronçon-date. Les strates sont créées en tenant compte de critères métier, i.e. des règles utilisées par les agents SNCF pour sélectionner les unités contrôlées (durée du tronçon, caractéristiques du jour, heure, ...). Les strates forment des groupes homogènes d'unités ttd au regard du phénomène de fraude. On note également $N_1,\ldots,N_H$ le nombre respectif d'unités voyageur dans chacune des strates. La réservation étant obligatoire dans les trains étudiés, ces effectifs par strates sont connus, et constituent donc une information auxiliaire mobilisable au moment de l'estimation (cf. section 3). En pratique, ces effectifs n'intègrent pas les fraudeurs qui n'ont pas pris de billet et sont donc sous-estimés ; le phénomène de fraude restant limité en termes de fréquence (moins de 1 % des unités voyageur sont contrôlés sans billet), cette sous-estimation peut être négligée.

Lors du premier degré de tirage, un échantillon S_I d'unités ttd est sélectionné dans U_I par tirage stratifié, avec sondage aléatoire simple de taille m_{Ih} dans chaque strate U_{Ih}. Soit S_{Ih} l'échantillon d'unités ttd sélectionné dans U_{Ih}, u_{ih} une unité ttd de S_{Ih} et N_{ih} son nombre d'unités voyageur. Dans chacune de ces unités u_{ih} sélectionnées au premier degré, un échantillon S_{ih} d'unités voyageur est obtenu par sondage aléatoire simple de taille n_{ih}. On note

$$n_h = \sum_{u_{ih}\in S_{Ih}} n_{ih}$$

le nombre total de voyageurs contrôlés dans la strate U_{Ih}. On note t_{kih} la somme totale non réglée par l'unité voyageur k de u_{ih}, que l'on suppose relevée sans erreur si l'individu k est contrôlé. La somme non réglée t_{kih} est nulle si l'unité voyageur est en situation régulière. On notera également r_{kih} la régularisation associée à l'unité voyageur k de u_{ih} (nulle si l'unité k est en situation régulière), et ε_{kih} la variable indicatrice de la fraude, qui vaut 1 si l'individu k de u_{ih} est en situation de fraude, et 0 sinon.

On s'intéresse à deux indicateurs : la perte associée à la fraude et le taux de fraudeurs, que l'on souhaite estimer sur chaque strate et sur l'ensemble de la population. La perte associée à la fraude se définit comme la différence entre la somme totale non réglée par les fraudeurs et le montant des régularisations. Elle est égale à

$$\begin{array}{rcccl} P_h & = & \sum_{u_{ih}\in U_{Ih}} \sum_{k\in u_{ih}} t_{kih} & - & \sum_{u_{ih}\in S_{Ih}} \sum_{k\in S_{ih}} r_{kih} \\ & = & T_h & - & \tilde{R}_h \end{array} \tag{1}$$

dans la strate U_{Ih}, et à

$$\begin{array}{rcccl} P & = & \sum_{h=1}^{H} T_h & - & \sum_{h=1}^{H} \tilde{R}_h \\ & = & T & - & \tilde{R} \end{array} \tag{2}$$

dans la population totale. Dans (1) et (2), le terme retranché représente le montant des régularisations, et son calcul ne pose pas de difficultés particulières si on connaît les sommes qui ont effectivement été récupérées. Dans l'estimation de P_h (respectivement P), l'incertitude porte uniquement sur l'estimation du premier terme T_h (respectivement T) qui représente la somme totale non réglée par les fraudeurs ; on se limitera donc à l'estimation de ce total dans la suite. Le taux de fraudeurs est égal à

$$TF_h = \frac{\sum_{u_{ih}\in U_{Ih}} \sum_{k\in u_{ih}} \varepsilon_{kih}}{\sum_{u_{ih}\in U_{Ih}} N_{ih}} = \frac{E_h}{N_h} \tag{3}$$

 ISSN: 2102-6238

dans la strate U_{Ih}, et à

$$TF = \frac{\sum_{h=1}^{H} E_h}{\sum_{h=1}^{H} N_h} = \frac{E}{N} \tag{4}$$

dans la population totale.

Comme l'a souligné un référé, la validité des estimateurs proposés dans la section suivante (et celle des estimateurs de variance associés) dépend fortement de la validité de la modélisation du mécanisme de contrôle. Une hypothèse cruciale, sous-jacente dans la modélisation proposée, est que le mécanisme de contrôle ne va pas conduire à sélectionner des voyageurs ayant un profil particulier relativement à la fraude, ou que les voyageurs ne vont pas adapter leur comportement relatif à la fraude en fonction du mécanisme de contrôle. En d'autres termes, le plan de sondage doit être non informatif (voir par exemple [4]), i.e. la distribution des variables d'intérêt doit être la même dans l'échantillon sélectionné et dans la population totale. Si le mécanisme de contrôle des unités ttd et des unités voyageur est réellement aléatoire, alors le mécanisme de sélection est non informatif. Dans le cas contraire, les estimateurs proposés peuvent être fortement biaisés.

Pour illustrer la notion de mécanisme informatif, prenons l'exemple (fictif) d'un train partant le lundi à 03h du matin de Rennes, direction Paris, avec un arrêt au Mans (ce qui correspond donc à deux unités ttd). Si le mécanisme de contrôle n'est pas connu des voyageurs, la fraude se réalisera indépendamment de ce mécanisme et la sélection pourra être considérée comme non-informative. Supposons maintenant que les contrôles sont toujours réalisés sur la première partie du voyage (entre Rennes et Le Mans), et que les voyageurs disposent de cette information. Cela peut donner lieu à un comportement de fraude où certains voyageurs circulent entre Rennes et Paris, avec un billet valable pour le tronçon Rennes-Le Mans seulement. Dans ce cas, ces fraudeurs échappent au mécanisme de contrôle, ce qui conduit à une sous-estimation du total de la fraude.

3. Estimation de la perte et du taux de fraudeurs

3.1. Estimation par strate

La somme totale T_h non réglée par les fraudeurs dans la strate U_{Ih} peut être estimée sans biais sous le plan de sondage par

$$\hat{T}_h = \frac{M_h}{m_{Ih}} \sum_{u_{ih} \in S_{Ih}} \frac{N_{ih}}{n_{ih}} \sum_{k \in S_{ih}} t_{kih}. \tag{5}$$

La variance $V(\hat{T}_h)$ de cet estimateur s'obtient à l'aide des formules pour un échantillonnage à deux degrés, avec sondage aléatoire simple à chaque degré. Un estimateur de variance sans biais pour $V(\hat{T}_h)$ est donné par

$$v(\hat{T}_h) = M_h^2 \left(1 - \frac{m_{Ih}}{M_h}\right) \frac{s_{t(Ih)}^2}{m_{Ih}} + \frac{M_h}{m_{Ih}} \sum_{u_{ih} \in S_{Ih}} N_{ih}^2 \left(1 - \frac{n_{ih}}{N_{ih}}\right) \frac{s_{t(ih)}^2}{n_{ih}}, \tag{6}$$

avec d'une part

$$s_{t(ih)}^2 = \frac{1}{n_{ih} - 1} \sum_{k \in S_{ih}} (t_{kih} - \bar{t}_{ih})^2 \tag{7}$$

 ISSN: 2102-6238

la dispersion estimée de la somme non réglée dans l'unité ttd u_{ih}, où

$$\bar{t}_{ih} = \frac{1}{n_{ih}} \sum_{k \in S_{ih}} t_{kih} \tag{8}$$

donne la somme moyenne non réglée par unité voyageur dans l'unité ttd u_{ih}, et d'autre part

$$s^2_{t(Ih)} = \frac{1}{m_{Ih} - 1} \sum_{u_{ih} \in S_{Ih}} \left(\hat{T}_{ih} - \frac{\hat{T}_h}{M_h} \right)^2 \tag{9}$$

la dispersion estimée des sommes non réglées dans la strate U_{Ih}, où

$$\hat{T}_{ih} = N_{ih} \bar{t}_{ih} \tag{10}$$

donne l'estimation du montant total des sommes non réglées dans l'unité ttd u_{ih}, voir par exemple [5, p. 178]. Comme le nombre N_h d'unités voyageur de la strate est connu, on préfère à l'estimateur $\hat{T}_h$ l'estimateur par le ratio

$$\hat{T}_{Rh} = \frac{N_h}{\hat{N}_h} \hat{T}_h, \tag{11}$$

où $\hat{N}_h$ est un estimateur sans biais du nombre d'unités voyageur dans la strate U_{Ih}, qui s'obtient à partir de la formule (5) en remplaçant t_{kih} par $x_{kih} = 1$. A l'aide de la technique de linéarisation ([2]), on obtient l'estimateur de variance approximativement sans biais

$$v(\hat{T}_{Rh}) = M_h^2 \left(1 - \frac{m_{Ih}}{M_h} \right) \frac{s^2_{lt(Ih)}}{m_{Ih}} + \frac{M_h}{m_{Ih}} \sum_{u_{ih} \in S_{Ih}} N_{ih}^2 \left(1 - \frac{n_{ih}}{N_{ih}} \right) \frac{s^2_{lt(ih)}}{n_{ih}}, \tag{12}$$

où $s^2_{lt(ih)}$ (respectivement $s^2_{lt(Ih)}$) s'obtient à partir des formules (7) et (8) (respectivement (9) et (10)) en remplaçant la variable t_{kih} par la variable linéarisée estimée $lt_{kih} = \left(t_{kih} - \frac{\hat{T}_h}{\hat{N}_h} \right)$.

De façon analogue, le taux de fraudeurs dans la strate U_{Ih} est estimé par

$$\widehat{TF}_h = \frac{\hat{E}_h}{\hat{N}_h}, \tag{13}$$

où $\hat{E}_h$ s'obtient à partir de la formule (5) en remplaçant t_{kih} par ε_{kih}. Un estimateur approximativement sans biais de variance est donné par

$$v(\widehat{TF}_h) = M_h^2 \left(1 - \frac{m_{Ih}}{M_h} \right) \frac{s^2_{l\varepsilon(Ih)}}{m_{Ih}} + \frac{M_h}{m_{Ih}} \sum_{u_{ih} \in S_{Ih}} N_{ih}^2 \left(1 - \frac{n_{ih}}{N_{ih}} \right) \frac{s^2_{l\varepsilon(ih)}}{n_{ih}}, \tag{14}$$

où $s^2_{l\varepsilon(ih)}$ (respectivement $s^2_{l\varepsilon(Ih)}$) s'obtient à partir des formules (7) et (8) (respectivement (9) et (10)) en remplaçant la variable t_{kih} par la variable linéarisée estimée $l\varepsilon_{kih} = \frac{1}{N_h}(\varepsilon_{kih} - \widehat{TF}_h)$.

 ISSN: 2102-6238

3.2. Estimation sur la population entière

La somme totale T non réglée par les fraudeurs s'estime à partir de la formule (11) par sommation, en utilisant l'estimateur par le ratio séparé [1, p.164]. On obtient l'estimateur

$$\hat{T}_{sr} = \sum_{h=1}^{H} \hat{T}_{Rh},$$

et en utilisant l'indépendance (supposée) des tirages dans les différentes strates, un estimateur de variance est donné par

$$v\left(\hat{T}_{sr}\right) = \sum_{h=1}^{H} v(\hat{T}_{Rh}).$$

L'estimateur du taux de fraudeurs sur l'ensemble de la population est donné par

$$\begin{aligned} \widehat{TF}_{sr} &= \left(\sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{N}_h\right)^{-1} \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{E}_h \\ &= \sum_{h=1}^{H} \frac{N_h}{N} \widehat{TF}_h, \end{aligned}$$

et en utilisant l'indépendance des tirages dans les différentes strates, un estimateur de variance est donné par

$$v\left(\widehat{TF}_{sr}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 v(\widehat{TF}_h).$$

3.3. Application aux données SNCF

Grâce à la modélisation présentée ci-dessus et aux données récoltées quotidiennement à bord des trains, les estimateurs proposés ont été calculés mensuellement pour les années 2007 à 2009 pour des domaines et sous-domaines sectoriels afin de suivre et d'analyser le phénomène de fraude tarifaire à bord des trains. Les domaines et sous-domaines sont définis en fonction du degré d'analyse souhaité. Sauf quelques rares cas pathologiques identifiés d'un point de vue métier, les taux de sondages ainsi que les variances correspondants observées ont démontré que le mode de contrôle en vigueur était largement suffisant pour utiliser ces estimations dans les (sous-) domaines d'intérêts, et ce sans campagne d'échantillonnage (c'est-à-dire de contrôle) complémentaire. Pour des raisons de confidentialité, il n'est pas possible de fournir ici d'exemples détaillés.

Du fait que le mode de contrôle en vigueur conduise à un échantillonnage suffisant pour des estimations de qualité utilisables en pratique, ces travaux ont permis de proposer un suivi régulier et relativement fin de la fraude tarifaire sur l'ensemble des trains Grandes Lignes (TGV France et TEOZ). Cela s'est traduit concrètement par la mise en place du Système d'Analyse de la FraudE (SAFE), qui calcule les estimateurs présentés précédemment en étant directement alimenté par les systèmes de suivi des réservations et de régularisations (des centaines de milliers de données par jour). Ce système constitue une aide à la décision clé pour l'aide au pilotage économique de la Lutte Anti-Fraude.

 ISSN: 2102-6238

4. Optimisation de l'allocation d'individus supplémentaires

4.1. Déterminer les tailles d'échantillons optimales

En dépit du fait que l'échantillonnage se soit révélé en pratique de bonne qualité pour les objectifs visés, nous avons souhaité étudier comment optimiser une éventuelle allocation complémentaire, ne serait-ce que pour mesurer l'optimalité de l'échantillonnage réalisé et identifier les pistes pour étendre les échantillons actuels afin d'améliorer la précision des indicateurs.

On dispose initialement dans chaque strate U_{Ih} d'observations réparties sur m_{Ih} unités primaires. On souhaite calculer une nouvelle allocation d'unités ttd dans chaque strate U_{Ih}, notée m_h, permettant d'optimiser la précision. On considère, de façon réaliste, que l'allocation du second degré (nombre d'unités voyageur tirées dans les unités ttd sélectionnées) ne dépend que d'impératifs techniques qui la rend non modifiable. L'optimisation que nous décrivons ci-dessous est relative à une variable particulière. L'optimisation conduit à des allocations qui pourront donc différer selon que l'on privilégie d'optimiser l'estimation de la somme totale non réglée ou celle du taux de fraude.

Soit y_{kih} la valeur prise par une variable d'intérêt y sur l'unité voyageur k d'une unité primaire u_{ih}. On s'intéresse à $\hat{Y}_h$ l'estimateur du total de la variable y_{kih} sur la strate U_{Ih}, obtenu en remplaçant dans la formule (5) la variable t_{kih} par y_{kih} et l'allocation initiale m_{Ih} dans la strate U_{Ih} par l'allocation recherchée m_h. On note également $\hat{Y}$ l'estimateur du total sur la population U. Un estimateur de la variance de $\hat{Y}_h$, obtenu à l'aide de la formule (6), peut se réécrire sous la forme

$$\begin{aligned} v(\hat{Y}_h) &= \frac{M_h^2}{m_h}\left[s^2_{y(Ih)} + \frac{1}{M_h}\sum_{u_{ih}\in S_{Ih}} N_{ih}^2\left(1-\frac{n_{ih}}{N_{ih}}\right)\frac{s^2_{y(ih)}}{n_{ih}}\right] - M_h s^2_{y(Ih)} \\ &= \frac{M_h^2}{m_h}D_h^2 - C_h^2 \end{aligned} \tag{15}$$

avec

$$\begin{aligned} C_h^2 &= M_h s^2_{y(Ih)}, \\ D_h^2 &= s^2_{y(Ih)} + \frac{m_h}{M_h}B_h^2, \\ B_h^2 &= \frac{1}{m_h}\sum_{u_{ih}\in S_{Ih}} N_{ih}^2\left(1-\frac{n_{ih}}{N_{ih}}\right)\frac{s^2_{y(ih)}}{n_{ih}}. \end{aligned} \tag{16}$$

La quantité C_h^2 constitue l'estimation d'un paramètre relatif à la strate U_{Ih}, qu'on peut considérer en première approximation comme indépendant du nombre d'unités primaires sélectionné. L'optimisation peut donc être réalisée sur la base de l'estimateur de variance simplifié

$$v_{SIMP}(\hat{Y}) = \sum_{h=1}^{H}\frac{M_h^2 D_h^2}{m_h}. \tag{17}$$

Les quantités $s^2_{y(Ih)}$ et B_h^2 intervenant dans le terme D_h^2 constituent également des estimations de paramètres relatifs à U_{Ih}, que l'on peut considérer comme indépendants du nombre m_h d'unités

 ISSN: 2102-6238

primaires sélectionné. Quand le taux de sondage m_h/M_h dans la strate tend vers zéro, le premier terme $s^2_{y(Ih)}$ est donc dominant dans D^2_h, et ce dernier peut être considéré comme fixe, ce que nous supposerons dans la suite.

Notons $A_h = \left(\sum_{j=1}^{H} M_j D_j\right)^{-1} M_h D_h$. L'optimisation consiste à déterminer les tailles d'échantillon m_h qui minimisent :

$$\sum_{h=1}^{H} \frac{A_h^2}{m_h}$$

sous les contraintes :

C1 : $m_{Ih} \leq m_h \leq M_h$ pour $h = 1, \ldots, H$,
C2 : $\sum_{h=1}^{H} m_h = m$,

l'allocation globale m étant fixée et supérieure à $m_1 = \sum_{h=1}^{H} m_{Ih}$. On obtient ainsi un problème de programmation linéaire qui peut être résolu grâce à des procédures d'optimisation classiques.

Comme l'a souligné un référé, l'optimisation est réalisée sous l'hypothèse que le terme D^2_h ne dépend pas du nombre m_h d'unités primaires sélectionné, ce qui suppose que le taux de sondage par strate m_h/M_h est faible. En pratique, le taux de sondage peut être non négligeable dans certaines strates. Dans ce cas, notons que l'on peut toujours réécrire l'estimateur de variance simplifié sous la forme

$$v_{SIMP}(\hat{Y}) \quad = \quad \sum_{h=1}^{H} \left(\frac{M_h^2}{m_h} \, s^2_{y(Ih)} + M_h \, B_h^2 \right). \tag{18}$$

Dans (18), les termes $M_h \, B_h^2$ peuvent être considérés comme indépendants des tailles d'échantillons m_h. Si certains taux de sondage par strate sont potentiellement non négligeables, on peut donc alternativement chercher à minimiser la quantité

$$\sum_{h=1}^{H} \frac{(A'_h)^2}{m_h}$$

sous les contraintes (C1) et (C2), avec $A'_h = \left(\sum_{j=1}^{H} M_j s_{y(Ij)}\right)^{-1} M_h s_{y(Ih)}$.

4.2. Procédure d'optimisation adaptée

Nous proposons d'utiliser une procédure plus adaptée à la nature du problème qui peut être utilisée comme guide dans le choix du nombre d'observations supplémentaire à collecter.

Si l'optimisation se faisait avec la contrainte (C2) uniquement, la solution obtenue serait de la forme $m_h = \mu \, A_h$ où μ a la dimension d'une taille d'échantillon. Avec les contraintes, on peut appliquer la méthode de Kuhn et Tucker ([3]). L'idée de base consiste à initialiser avec une allocation de la forme $m_h = \mu \, A_h$, pour un certain paramètre μ. Le problème est que cette allocation ne peut être utilisée dans certaines strates. Il y a d'une part les strates où elle conduirait

 ISSN: 2102-6238

Strate h	M_h	m_{Ih}	A_h	$\mu_h = m_{Ih}/A_h$	$\lambda_h = M_h/A_h$
1	200	25	0.3	83.33	666.67
2	100	20	0.2	100	500
3	300	15	0.1	150	3 000
4	200	10	0.25	40	800
5	400	30	0.15	200	2 666.67

TABLEAU 1. *Tailles de strate, tailles d'échantillon et paramètres relatifs à une population d'unités primaires*

à une taille d'échantillon inférieure à l'allocation initiale m_{Ih}. Ce cas se produit si $\mu\, A_h \leq m_{Ih}$, ce qui correspond donc à la classe

$$H^-(\mu) = \{h\,;\, \mu \leq \mu_h\},$$

avec $\mu_h = m_{Ih}/A_h$. Pour les strates de $H^-(\mu)$, on impose $m_h = m_{Ih}$. Il y a d'autre part les strates où l'allocation conduirait à une taille d'échantillon supérieure à la taille de la strate M_h. Ce cas se produit si $\mu\, A_h \geq M_h$, ce qui correspond à la classe

$$H^+(\mu) = \{h\,;\, \mu \geq \lambda_h\},$$

avec $\lambda_h = M_h/A_h$. Pour les strates de $H^+(\mu)$, on impose $m_h = M_h$.

La méthode de Kuhn et Tucker ([3]) conduit donc au résultat suivant : on obtient $m_h = M_h$ pour les A_h les plus gros, i.e. $h \in H^+(\mu)$; $m_h = m_{Ih}$ pour les strates h telles que les $\mu_h = m_{Ih}/A_h$ soient les plus élevés, i.e. $h \in H^-(\mu)$; $m_h = \mu\, A_h$ pour les autres strates, avec un paramètre μ à déterminer. Il est clair que les classes $H^+(\mu)$ et $H^-(\mu)$ sont disjointes et à déterminer en fonction de μ. On va donc, en fait, déterminer l'allocation en fonction de ce paramètre ce qui nous donnera l'allocation globale d'échantillon comme une fonction croissante de μ. Si les deux classes précédentes sont déterminées, l'équation

$$m = \sum_{h\in H^+} M_h + \sum_{h\in H^-} m_{Ih} + \mu \sum_{h\notin H^+\cup H^-} A_h,$$

conduit à :

$$\mu = \frac{m - \sum_{h\in H^+} M_h - \sum_{h\in H^-} m_{Ih}}{\sum_{h\notin H^+\cup H^-} A_h}.$$

En pratique, on pourra procéder de la façon suivante : on dispose de deux ensembles de nombres $\mu_h = m_{Ih}/A_h$ et $\lambda_h = M_h/A_h$, $h = 1,\ldots,H$, qui définissent $2H+1$ intervalles. Dans chacun d'eux les classes H^- et H^+ sont parfaitement définies. Il suffit de calculer m en fonction de μ pour chacune de ces valeurs ($m(\mu)$ est linéaire croissante dans chaque intervalle), puis, m étant donné, de déterminer l'intervalle auquel appartient μ et donc l'allocation.

Pour fixer les idées, considérons le cas d'une population constituée de 1200 unités primaires, découpées en 5 strates. Les paramètres utilisés sont donnés dans le tableau 1. On note $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(5)}$ (respectivement, $\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(5)}$) les valeurs prises par les coefficients $\mu_1, \mu_2, \ldots, \mu_5$ (respectivement, $\lambda_1, \lambda_2, \ldots, \lambda_5$), rangées par ordre croissant. On a ici

$$\begin{aligned}(\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(5)}) &= (\mu_4, \mu_1, \mu_2, \mu_3, \mu_5)\\ &= (40, 83.33, 100, 150, 200)\end{aligned}$$

 ISSN: 2102-6238

et

$$\begin{aligned}(\lambda_{(1)},\lambda_{(2)},\ldots,\lambda_{(5)}) &= (\lambda_2,\lambda_1,\lambda_4,\lambda_5,\lambda_3)\\ &= (500, 666.67, 800, 2\,666.67, 3\,000).\end{aligned}$$

On obtient donc un découpage selon les intervalles

$$\begin{array}{llll}[\,0,40\,[\,, & [\,40,83.33\,[\,, & [\,83.33,100\,[\,, & [\,100,150\,[\,,\\ [\,150,200\,[\,, & [\,200,500\,[\,, & [\,500,666.67\,[\,, & [\,666.67,800\,[\,,\\ [\,800,2\,666.67\,[\,, & [\,2\,666.67,3\,000\,[\,, & [\,3\,000,+\infty\,[\,, & \end{array}$$

et la courbe de la fonction $m(\mu)$ est donnée dans le graphique 4.2. Par exemple :

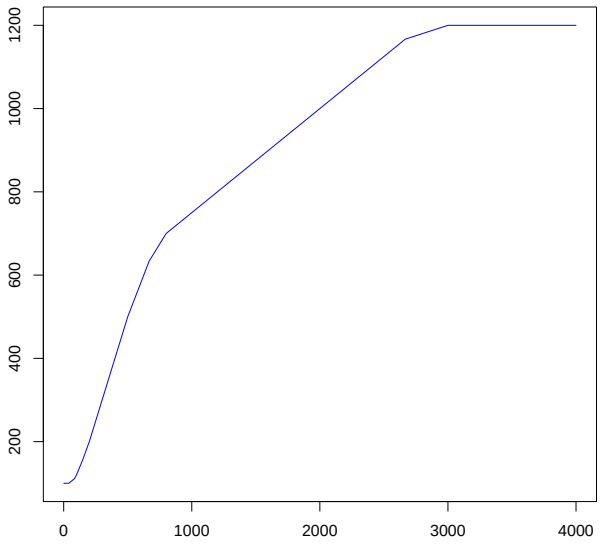

FIGURE 1. *Courbe de la fonction $m(\mu)$ pour μ variant de* 0 *à* 4000

– si $\mu \in [0,40[$, alors pour tout h on a $\mu \leq \mu_h$ et $\mu \leq \lambda_h$, d'où $H^- = \{1,2,3,4,5\}$. On obtient

$$m(\mu) = \sum_{h \in H^-} m_{Ih} = 100.$$

– si $\mu \in [150,200[$, alors

$$\mu_4 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu \leq \mu_5$$

et pour tout h, on a $\mu \leq \lambda_h$. On a donc $H^- = \{5\}$ et $\{h \notin H^- \cup H^+\} = \{1,2,3,4\}$. On obtient

$$\begin{aligned} m(\mu) &= \textstyle\sum_{h \in H^-} m_{Ih} + \mu \sum_{h \notin H^+ \cup H^-} A_h \\ &= 30 + 0.85\ \mu. \end{aligned}$$

 ISSN: 2102-6238

– si $\mu \in [800, 2\,666.67[$, alors pour tout h on a $\mu \geq \mu_h$ et d'autre part,

$$\lambda_2 \leq \lambda_1 \leq \lambda_4 \leq \mu \leq \lambda_5 \leq \lambda_3.$$

On a donc $H^+ = \{1,2,4\}$ et $\{h \notin H^- \cup H^+\} = \{3,5\}$. On obtient

$$\begin{array}{rcl} m(\mu) & = & \sum_{h\in H^+} M_h + \mu \sum_{h\notin H^+\cup H^-} A_h \\ & = & 500 + 0.25\ \mu. \end{array}$$

Supposons que l'on souhaite obtenir un échantillon de $m = 160$ unités primaires. On a $m(\mu) = 160$ pour une valeur de μ appartenant à l'intervalle $[150, 200[$, ce qui correspond à $H^- = \{5\}$ et $\{h \notin H^- \cup H^+\} = \{1,2,3,4\}$. On en déduit que

$$m_5 = m_{15} = 30,$$

et

$$\begin{array}{rcl} \mu & = & \left(\sum_{h\notin H^+\cup H^-} A_h\right)^{-1} \left(m - \sum_{h\in H^+} M_h - \sum_{h\in H^-} m_{Ih}\right) \\ & = & (0.85)^{-1}(160 - 0 - 30) = 152.94 \end{array}$$

pour $h \neq 5$, on a $\mu_{uh} = \mu\ A_h = 152.94\ A_h$, ce qui donne après arrondi

$$m_1 = 46 \quad m_2 = 31 \quad m_3 = 15 \quad m_4 = 38.$$

4.3. Application aux données SNCF

Comme mentionné en introduction de ce paragraphe, la procédure d'optimisation décrite ici n'a pas été utilisée à des fins opérationnelles. A la fois parce que l'échantillonnage actuel conduisait à des estimateurs de qualité suffisante en pratique mais également parce que l'allocation des moyens de contrôle dépend de nombreux impératifs autres que celui de précision des estimateurs (emploi du temps du personnel, service à bord, lutte anti-fraude, coûts, ...).

Nous avons néanmoins réalisé des applications numériques pour étudier de manière plus fine où une réallocation des moyens de contrôle s'avérerait le plus efficace pour augmenter la précision des indicateurs - en particulier lorsque la précision initiale était la moins bonne. On rappelle que l'on note m_{Ih} le nombre d'unités ttd sélectionnées initialement dans la strate U_{Ih}. On note également $m_{1.} = \sum_{h=1}^{H} m_{Ih}$ le nombre total d'unités ttd sélectionné initialement. Un exemple de résultat est fourni ci-dessous en comparant quatre allocations :

– allocation initiale : $m_h^* = m_{Ih}$ dans chaque strate U_{Ih} ;
– réallocation complète : on sélectionne un échantillon d'unités ttd de même taille globale $\sum_{h=1}^{H} m_h^* = m_{1.}$, mais avec une allocation par strate visant à minimiser la variance donnée en (17), selon la procédure décrite en section 4.2 ;
– allocation complémentaire : en plus des unités ttd échantillonnées initialement, un échantillon complémentaire de taille globale m_0 (fixé d'après une contrainte de budget) est sélectionné ; les allocations par strate $m_h^* \geq m_{Ih}$ sont déterminées afin de minimiser la variance donnée en (17), selon la procédure décrite en section 4.2 ;
– allocation totale : toutes les unités ttd sont sélectionnées ($m_h^* = M_h$ dans chaque strate U_{Ih}).

 ISSN: 2102-6238

On utilise dans cet exemple les données collectées au mois de septembre 2008 pour le TGV Méditerranée - Paris Côte d'Azur, ce qui correspond à 4 strates d'unités ttd. Les tableaux 2 et 3 donnent, dans le cas d'une optimisation pour l'estimation du taux de fraude (respectivement, de la perte financière) et pour chacune des quatre allocations, le nombre d'unités ttd sélectionnées dans chaque strate ainsi que la valeur obtenue pour $S = \sum_{h=1}^{H} \frac{A_h^2}{m_h}$. Dans le cas de l'allocation complémentaire, l'échantillon d'unités ttd de taille finale $m^* = 1165$ correspond à un complément d'échantillon de taille $m_0 = 485$. Comme prévu, la réallocation complète conduit à des estimations plus précises que l'allocation initiale ; l'amélioration est particulièrement sensible dans le cas de l'optimisation pour l'estimation du taux de fraude.

 ISSN: 2102-6238

	m_1^*	m_2^*	m_3^*	m_4^*	m^*	S
Allocation initiale	147	81	231	221	680	$2.2\ 10^{-3}$
Réallocation complète	63	44	453	119	680	$1.0\ 10^{-3}$
Allocation complémentaire	212	226	993	225	1 165	$0.6\ 10^{-3}$
Allocation totale	212	240	2 249	225	2 926	$0.4\ 10^{-3}$

TABLEAU 2. *Quatre allocations d'unités ttd et précision associée dans le cas d'une optimisation de l'estimation du taux de fraude*

	m_1^*	m_2^*	m_3^*	m_4^*	m^*	S
Allocation initiale	147	81	231	221	680	$1.6\ 10^{-3}$
Réallocation complète	108	59	334	179	680	$1.5\ 10^{-3}$
Allocation complémentaire	212	136	592	225	1 165	$0.9\ 10^{-3}$
Allocation totale	212	240	2 249	225	2 926	$0.5\ 10^{-3}$

TABLEAU 3. *Quatre allocations d'unités ttd et précision associée dans le cas d'une optimisation de l'estimation de la perte financière*

Références

[1] W.G. COCHRAN : *Sampling Techniques*. Wiley, 1977.

[2] J-C. DEVILLE : Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25:193–203, 1999.

[3] H.W. KUHN et A.W. TUCKER : Nonlinear programming. *In Proceedings of 2nd Berkeley Symposium*, pages 481–492. University of California Press, 1951.

[4] D. PFEFFERMANN et M. SVERCHKOV : Inference under informative sampling. *In Handbook of Statistics*, volume 29B, pages 455–487. Pfeffermann, D., and Rao, C.R., 2009.

[5] Y. TILLÉ : *Théorie des sondages : échantillonnage et estimation en population finie*. Dunod, 2001.

 ISSN: 2102-6238