Journal de la Société Française de Statistique
Vol. 155 No. 4 (2014)

Estimation de la consommation d'eau d'un secteur hydraulique à partir d'un échantillon d'usagers télérelevés

Title: Water consumption estimation of an hydraulic district from a sample of users equipped with automatic meter reading

Karim Claudio[1,2], Vincent Couallier[2], Yves Le Gat[3] et Jérôme Saracco[2,4]

Résumé : Le télérelevé des compteurs d'eau potable est, à l'heure actuelle, la meilleure technologie permettant de connaître en temps réel la consommation en eau d'un usager. A l'échelle d'un secteur hydraulique, un télérelevé complet permet de connaître la consommation totale d'une population de taille finie, au pas de temps aussi fin que quelques heures. Cependant son coût de déploiement rend sa généralisation parfois impossible sur certaines communes, pour lesquelles des techniques d'échantillonnage doivent être mises en place. Dans un objectif d'estimation du total des consommations, cet article décrit et compare des techniques de sondage et propose de retenir une méthodologie de mise en place d'un échantillon opérationnel et de calage de l'estimateur du total correspondant.

Abstract: Automatic water meter reading is currently the best technology supplying real-time water consumption data. A comprehensive meter reading enables to know precisely the total consumption, on a time scale as short as an hour. However the generalization for some municipalities is sometimes too expensive, in which case sampling techniques are needed. So as to estimate precisely the total consumption, we describe and compare standard techniques, and we propose a methodology for selecting such sample, and an estimation strategy for the total consumption.

Mots-clés : Télérelevé des compteurs, Estimation de la consommation en eau, Sondage stratifié, Redressement par information auxiliaire
Keywords: Automatic meter reading, Estimation of water consumption, Stratified sampling, Calibration with auxiliary information
Classification AMS 2000 : 35L05, 35L70

Introduction

Les problématiques liées à l'eau ont évolué avec le temps, en France comme dans le reste du monde, et l'enjeu actuel pour les acteurs de la gestion de l'eau est la préservation de la ressource. Les pertes sur les réseaux d'eau potable représentent en moyenne 20% du volume annuel distribué en France, c'est pourquoi l'accent est mis sur la minimisation des volumes perdus. Pour cela il

[1] LyRE - centre de R&D Lyonnaise des Eaux
E-mail : karim.claudio@lyonnaise-des-eaux.fr
[2] Institut de Mathématiques de Bordeaux
E-mail : vincent.couallier@u-bordeaux2.fr
[3] IRSTEA Bordeaux - Équipe Réseaux, épuration et qualité des eaux.
E-mail : yves.legat@irstea.fr
[4] INRIA Bordeaux Sud-Ouest - Équipe CQFD
E-mail : jerome.saracco@math.u-bordeaux1.fr

 ISSN: 2102-6238

convient d'améliorer et d'affiner la connaissance que l'on a sur le rendement de réseau (défini comme le ratio entre le volume consommé sur un intervalle de temps Δt et le volume qui a été distribué sur ce même intervalle). Concernant le volume distribué (dénommé l'*entrée*), les données disponibles sont complètes : l'instrumentation des réseaux d'eau potable permet de connaître le volume distribué sur un pas de temps $\Delta t = 5$ min. L'incertitude sur la connaissance du rendement vient du volume consommé (la *sortie*). Puisqu'en général, la relève manuelle des compteurs d'eau se fait chaque année, et à des dates étalées dans le temps, on peut au mieux estimer le volume consommé uniquement sur un pas de temps annuel.

Cette connaissance de la consommation ne permet pas de réaliser un suivi régulier du rendement, ce qui a conduit au déploiement d'une nouvelle technologie : associé au compteur d'eau individuel, un émetteur de télérelève récolte et retransmet automatiquement les index de consommation (consommation cumulée) sur un pas de temps variant entre une et six heures. Aujourd'hui l'étendue du télérelevé dépend des contrats engagés par les collectivités : pour certaines le télérelevé est exhaustif (tous les compteurs sont équipés), pour d'autres il est partiel voire inexistant. Dans le premier cas, la connaissance de la consommation est totale, il est donc facile de calculer le rendement de réseau ou encore les pertes (différence entre l'*entrée* et la *sortie*). Dans le second cas, il n'est pas possible de calculer directement ces indicateurs de performance. La généralisation du télérelevé engendrant un coût que certaines collectivités ne peuvent pas supporter, on cherche alors à estimer, à partir d'un échantillon de la population, la série chronologique des consommations totales (journalières ou hebdomadaires).

Une information fiable sur la consommation d'un secteur a de grandes implications sur la gestion du réseau. En particulier, l'estimateur de la consommation de la population doit être suffisamment précis afin de permettre la détection de fuites potentielles. En conséquence, l'échantillon à constituer (taille d'échantillon, tirage des individus) doit conduire à un estimateur du total dont la précision est contrôlée. L'application de la théorie des sondages en population finie répond à cette problématique.

S'il n'existe pas, dans la littérature en théorie des sondages, de méthodes complètes pour constituer un échantillon dans la cas d'une estimation de la consommation d'eau, les différentes étapes permettant la construction d'un panel d'usagers à télérelever y sont développées : la construction des strates (voir par exemple Dalenius and Hodges, 1959 ou Serfling, 1968), le taux optimal d'échantillonnage (Tillé, 2001), l'allocation de l'échantillon dans chaque strate (Cochran, 1977,Ardilly, 2006). A partir de ces différents travaux, il est alors possible de définir un échantillon permettant d'estimer, de manière fiable, la consommation en eau d'une population.

Une fois l'échantillon constitué, des informations auxiliaires sont aussi souvent disponibles. Ainsi, si l'estimateur initial ne permet pas de détecter de manière suffisamment précise les fuites sur le réseau d'eau potable, ces informations vont nous permettre de redresser cet estimateur, afin d'en améliorer potentiellement la précision.

Dans cet article, on décrit la mise en oeuvre d'un plan d'échantillonnage stratifié et on évalue, sur des données réelles, différentes méthodes de redressement à partir d'informations auxiliaires mises à jour temporellement. En particulier, on illustre la faible performance d'une post-stratification consécutive à une stratification, ainsi que le gain attendu des méthodes de redressement par régression. Cette étude est rendue possible par l'exploitation de données réelles d'un secteur entièrement télérelevé qui permet comme cas-test d'évaluer la performance des

 ISSN: 2102-6238

méthodes de sondage proposées.

La section 1 rappelle les notations et définit le plan d'échantillonnage stratifié avec tirage aléatoire simple et sans remise. Diverses méthodes de constitution de strates et de choix de taille d'échantillon sont ensuite comparées afin d'en valider une pour ce type de données. Après avoir illustré le fait qu'entre le moment de la constitution de l'échantillon et son utilisation pour l'estimation du total, de nouvelles informations auxiliaires sont disponibles, nous évaluons, à la section 2, la performance de différentes méthodes permettant de redresser un estimateur issu d'un sondage stratifié. La section 3 conclut l'article et donne quelques perspectives.

1. Constitution d'un échantillon et estimation du total

Soit une population U de taille N dont on extrait un échantillon s de taille n. Cet échantillon, une fois constitué et équipé en télérelevé, doit permettre de récolter des données sur la consommation en eau potable des usagers. La variable d'intérêt est la consommation individuelle (notée Y_i, $\forall i \in [\![1,N]\!]$) sur un pas de temps défini (le jour ou la semaine). Nous nous intéressons à son total $T_Y = \sum_{i\in U} Y_i$.

L'aspect temporel de la variable d'intérêt conduit à considérer que l'estimation du total peut être faite à chaque pas de temps. La variable d'intérêt sera donc notée $Y_i(t)$, la consommation de l'individu i au temps t, dont nous voulons estimer le total $T_Y(t)$.

Au vu des données disponibles et dans la mesure où l'objectif est d'obtenir une information infra-annuelle, nous portons notre choix sur un pas de temps hebdomadaire. Pour reconstruire les données erronées par interpolation linéaire, il est plus sûr de travailler sur un maximum de données valides. Le pas de temps hebdomadaire permet de disposer de 28 index (contre 6 index sur un pas de temps journalier) pour calculer les consommations et parer plus facilement les éventuelles erreurs de télé-transmission. Nous notons $\mathscr{T}$ le nombre de semaines pour une période d'étude considérée, nous allons estimer $\mathscr{T}$ consommations totales hebdomadaires.

Chaque année, un relevé manuel des compteurs est réalisé, permettant ainsi de connaître la quantité d'eau consommée par chaque individu de la population durant l'année qui vient de s'écouler. Il semble raisonnable de réaliser un sondage stratifié à partir de cette variable connue exhaustivement. Pour une estimation de la consommation durant l'année A, la variable de stratification la plus adéquate (et disponible) est la consommation annuelle individuelle en $A-1$. Nous noterons par la suite la variable de stratification X.

1.1. Stratification de la population

Dans une population, on distingue deux types d'usagers : les ménages et les gros consommateurs. Cette dernière catégorie de la population (essentiellement composée d'industriels et d'institutions) est définie, selon les experts métier, comme les usagers dont la consommation annuelle individuelle est supérieure ou égale à 1000 m^3. Ce groupe de consommateurs, malgré son faible effectif (ils sont environ 1% de la population) représente une part importante des consommations annuelles totales (leurs consommations en 2010 représentent 10% de la consommation annuelle totale). Afin de limiter l'influence de ces valeurs extrêmes, nous attribuons à ces individus un poids de sondage égal à 1 (voir Beaumont et al., 2013) : ils sont donc regroupés dans une même strate enquêtée exhaustivement.

 ISSN: 2102-6238

On se fixe pour des raisons opérationnelles un nombre L de strates à créer. La borne des *gros consommateurs* étant imposée, il reste $L^* = L - 1$ strates à définir. Plus le nombre de strates est important, pour l'homogénéité de ces dernières augmente. Cependant, ce gain en précision devient marginal à un certain point et il peut arriver dans certains cas qu'une augmentation du nombre de strates nuisent à cette homogénéité (voir Kpedekpo, 1973). Nous utilisons comme indicateur pour définir le nombre de strates la somme des variances intra-strates de la variable X, notée $V_W(X)$:

$$V_W(X) = \sum_{h=1}^{L^*} \frac{N_h}{N} S_{Xh}^2$$

où N_h et S_{Xh}^2 sont respectivement l'effectif et la variance de X au sein de la strate h. Le nombre de strates est choisi de telle sorte que le gain en précision à l'ajout d'une strate soit significatif. Nous choisissons alors le nombre L^* de strates tel que le gain en précision pour $L^* + 1$ strate soit inférieur à 1% (ce seuil a été fixé arbitrairement).

$$L^* = \arg\max_l \left\{ \frac{V_W(X|=l-1) - V_W(X|=l)}{V_W(X|=l-1)} \geq 1\% \right\}, \tag{1}$$

où $V_W(X|=l)$ est la somme des variances intra-strates de X en considérant l strates. Une fois le nombre de strates sélectionné, il convient de choisir les bornes des strates pour compléter leur définition.

Le découpage optimal de la population est celui qui permettrait d'obtenir des strates homogènes, c'est-à-dire obtenir une variance minimale dans chaque strate. D'après Dalenius (Dalenius, 1950), cela revient à trouver les bornes x_1,..., x_{L^*-1} comme solutions du système suivant

$$\frac{S_{Xh}^2 + (x_h - \bar{X}_h)^2}{S_{Xh}} = \frac{S_{Xh+1}^2 + (x_{h+1} - \bar{X}_{h+1})^2}{S_{Xh+1}}, \qquad \forall h \in [\![1, L^* - 1]\!]$$

où $\bar{X}_h$ est la moyenne de la variable X au sein de la strate h. S'il n'existe pas de solutions analytiques à ce problème, plusieurs approximations ont été proposées (voir par exemple Serfling, 1968, Nicolini, 2001 et Lavallée and Hidiroglou, 1988), nous retenons toutefois la méthode de Dalenius and Hodges (1959) pour sa simplicité d'implémentation. Un exemple de mise en œuvre pratique est détaillé dans Cochran (1977, p.129). Considérons la fonction $\mathbf{C}_X(\xi)$ définie ainsi :

$$\mathbf{C}_X(\xi) = \sum_{x=1}^{\xi} \sqrt{\sum_{i=1}^{N} \mathbb{1}_{[x-1<X_i\leq x]}}, \qquad \forall \xi \in \mathbb{N}^*.$$

La borne supérieure de la strate h s'obtient en résolvant l'équation (2) :

$$x_h = \arg\max_{\xi} \left\{ \mathbf{C}_X(\xi) \leq \frac{h}{L^*} \times \mathbf{C}_X(X_M) \right\} \tag{2}$$

où X_M est la valeur maximale de la variable X. Les strates obtenues en appliquant la méthode de stratification de Dalenius et Hodges à nos données sont indiquées dans le Tableau 4 à la Section 1.3.

 ISSN: 2102-6238

1.2. Taille n de l'échantillon et répartition dans les strates

L'échantillon s est réparti dans chaque strate en sous-échantillons g_h de taille n_h, avec $n = \sum_{h=1}^{L} n_h$. On note $a_h = \frac{n_h}{n}$. On considère deux approches pour répartir l'échantillon dans chaque strate : la répartition proportionnelle et la répartition x-optimale. La dernière strate (celle des *gros consommateurs*) étant enquêtée exhaustivement, la répartition permettra donc de calculer n_h pour $h \in [\![1, L^*]\!]$.

La répartition proportionnelle, comme décrite dans Tillé (2001) est une méthode consistant à répartir l'échantillon dans chacune des L^* strates de manière proportionnelle à l'effectif de la strate :

$$a_h = \frac{n_h}{n^*} = \frac{N_h}{N^*}, \qquad \forall h \in [\![1, L^*]\!].$$

où $N^* = \sum_{h=1}^{L^*} N_h$ et $n^* = \sum_{h=1}^{L^*} n_h$. Une conséquence pratique de cette approche est que le taux de sondage dans chaque strate est constant quel que soit $h \in [\![1, L^*]\!] : f_h = n_h/N_h = n^*/N^*, \forall h \in [\![1, L^*]\!]$.

La répartition x-optimale est une méthode dérivée de la répartition de Neyman (voir par exemple Cochran, 1977) qui, en plus de prendre en compte l'effectif de chaque strate, inclut la dispersion de celle-ci. La répartition de Neyman est :

$$a_{h-Neyman} = \frac{N_h S_{Yh}}{\sum_{i=1}^{L^*} N_i S_{Yi}}, \qquad \forall h \in [\![1, L^*]\!],$$

où S_{Yh} est l'écart-type de Y au sein de la strate h. Ces valeurs étant inconnues, la répartition x-optimale préconise d'utiliser l'écart-type (S_{Xh}) de la variable de stratification X :

$$a_{h-x.opt} = \frac{N_h S_{Xh}}{\sum_{i=1}^{L^*} N_i S_{Xi}}, \qquad \forall h \in [\![1, L^*]\!].$$

On peut montrer (voir par exemple Ardilly, 2006) que la variance d'un estimateur d'un total ou d'une moyenne est plus petite si l'on utilise la répartition de Neyman que la répartition proportionnelle et cette différence est d'autant plus importante que les S_{Yh} varient d'une strate à l'autre. Par analogie, on préfère alors la répartition x-optimale dans notre cas d'étude où il y a de fortes disparités des variances intra-strates (voir Tableau 4).

Concernant la taille n de l'échantillon, le nombre d'individus à sonder a un impact direct sur la précision de tout estimateur, en particulier celui de Horvitz-Thompson que l'on utilisera ($\hat{T}_Y$). Il est bien connu que la variance de l'estimateur de Horvitz-Thompson (dans le cadre d'un sondage stratifié) s'écrit :

$$\mathbb{V}(\hat{T}_Y) = \sum_{h=1}^{L} N_h^2 \frac{(1 - f_h)}{n_h} S_{Yh}^2,$$

 ISSN: 2102-6238

ce qui permet de trouver une taille n^* pour un écart-type σ de $\hat{T}_Y$ que l'on se fixe (*cf.* Tillé, 2001)

$$n^* = \frac{\sum_{h=1}^{L^*} \frac{N_h^2}{a_h} S_{Yh}^2}{\sigma^2 + \sum_{h=1}^{L^*} N_h S_{Yh}^2}. \tag{3}$$

Cette formule est inexploitable en pratique car les S_{Yh}^2 sont inconnus. Nous aurions pu, tout comme pour la répartition de l'échantillon, contourner ce problème en utilisant les variances S_{Xh}^2.Cependant, cette option pose ici différents problèmes :

- il est difficile de transcrire l'erreur tolérée à un pas de temps hebdomadaire vers un pas de temps annuel,
- l'exploitation de données annuelles ne permet pas de mettre en évidence la saisonnalité des données et notamment la forte hétérogénéité des consommations en périodes estivales,
- l'application de la formule (3) avec des données annuelles n'a pas renvoyé des résultats satisfaisants.

Comme il est recommandé dans Fellegi (2010), le calcul de n est "difficile à obtenir et une approximation est fréquemment faite à partir de populations similaires". Un calcul préliminaire sur des secteurs entièrement télérelevés pourra permettre de calculer la taille d'échantillon nécessaire pour obtenir une précision σ requise par des experts métier et servir de référence pour d'autres applications.

1.3. Mise en application de la méthode

Les résultats numériques de cet article proviennent de la commune de Canéjan (33) : il s'agit d'une population de 1822 usagers dont les compteurs d'eau sont tous équipés d'un émetteur de télérelevé depuis fin 2009. Sur cette population, les index de consommation sont retransmis toutes les 6 heures (soit 4 index/jour). La période d'étude s'étend du 01/01/2011 au 31/12/2011, 52 estimateurs de la consommation totale hebdomadaire sont calculés et peuvent être comparés aux valeurs mesurées de la consommation hebdomadaire totale. La stratification est réalisée à partir des consommations annuelles individuelles de l'année 2010, disponibles à partir des bases de données clientèles.

On présente dans cette section l'application numérique de la méthodologie préconisée. On considère ici l'estimateur de Horvitz-Thompson du total de la consommation à la semaine t :

$$\hat{T}_Y(t) = \sum_{i \in s} \frac{Y_i(t)}{\pi_i} = \sum_{h=1}^{L} N_h \hat{\bar{y}}_h(t) \qquad \text{où} \qquad \hat{\bar{y}}_h(t) = \frac{1}{n_h} \sum_{i \in g_h} Y_i(t) \tag{4}$$

avec $\pi_i = \frac{n_h}{N_h}$ si $i \in G_h$. Cet estimateur est sans biais : $\mathbb{E}[\hat{T}_Y(t)] = T_Y(t)$.

La stratification de la population suivant la méthode proposée en Section 1.2 préconise un découpage de la population en 11 strates comme le montre les résultats du Tableau 1.
Le gain en précision en créant 11 strates étant marginal, nous décidons de segmenter la population en 10 groupes, auxquels il faut rajouter la strate des "*gros consommateurs*", soit un total de 11

 ISSN: 2102-6238

strates.

Concernant à présent les bornes de ces strates, sur notre cas d'étude, nous avons $\mathbf{C}_X(1000)$=655 (la variable X a été bornée à 1000 pour exclure les *gros consommateurs*). Le calcul de la borne supérieure de la strate une revient d'après l'équation (2) à résoudre le problème suivant

$$x_1 = \arg\max_{\xi} \left\{ \mathbf{C}_X(\xi) \leq \frac{1}{10} \times 655 \right\}$$

Comme indiqué au Tableau 2, la borne supérieure de la strate 1 est 31 m^3/an. De la même façon pour la strate 2, il faut résoudre le problème

$$x_2 = \arg\max_{\xi} \left\{ \mathbf{C}_X(\xi) \leq \frac{2}{10} \times 655 \right\}$$

Ainsi la borne supérieure de la strate 2 est 50 m^3. L'opération est par la suite réitérée pour chacune des autres strates, les bornes finales des strates de consommation sont présentées au Tableau 4.

TABLEAU 1. *Choix du nombre de strates**

l	$V_W(X\|=l)$	$\frac{V_W(X\|=l-1)-V_W(X\|=l)}{V_W(X\|=l-1)}$
1	16 019	
2	13 372	16.50%
...	...	...
9	11 088	1.10%
10	10 972	**1.10 %**
11	10 869	0.90%
12	10 772	0.90%

* hors "*gros consommateurs*"

TABLEAU 2. *Calcul des bornes de strates*

ξ (m^3/an)	$\mathbf{C}_X(\xi)$	Strate
1	5.47	1
...	...	...
30	62.51	1
31	**64.69**	**1**
32	67.79	2
...	...	...
50	**128.93**	**2**
51	132.67	3
...	...	...
1000	655	10

Après discussion avec les experts métier, nous nous fixons comme critère d'obtenir un écart-type moyen sur les 52 estimateurs égal à 91 m^3, ce qui correspond au volume hebdomadaire estimée d'une fuite. Sur notre période d'étude et pour ce cas-test entièrement télérelevé, en utilisant la formule (3), il est possible de calculer 52 valeurs de la taille d'échantillonnage $n(t)$ optimale, que l'on peut analyser avec le Tableau 3.

TABLEAU 3. *Fractiles pour la taille d'échantillon et le taux de sondage pour les 52 semaines*

	1er Quart.	Médiane	Moyenne	3ème Quart.
$n(t)$	467	638	934	1 116
$n(t)/N$	26%	35%	51%	61%

Comme il a été dit précédemment, l'échantillon ne peut être modifié une fois créé, c'est pourquoi $n(t)$ doit être constant et fixé à une valeur n. Nous nous servons alors de la médiane (écartant ainsi tout résultat extrême) pour décider du nombre d'individus à échantillonner : $n = 638$, ce qui

 ISSN: 2102-6238

correspond à un taux de sondage $f = 35\%$. Les individus sont ensuite sélectionnés par sondage aléatoire simple dans chaque strate suivant la répartition x-optimale, qui est préférable dans un cas comme le nôtre où il y a une forte disparité des S_{Xh} (voir le plan de sondage du Tableau 4).

TABLEAU 4. *Définition du plan de sondage (stratification et échantillonnage) pour le calcul de l'estimateur* $\hat{T}_Y(t)$

strate	Bornes	N_h	S^2_{Xh}	N_h/N	n_h	n_h/N_h
1	[0 ; 31 [	180	100	10%	86	48%
2	[31 ; 50 [	173	33	9%	45	26%
3	[50 ; 65 [	205	19	11%	44	21%
4	[65 ; 79 [	200	20	11%	40	20%
5	[79 ; 94 [	198	19	11%	42	21%
6	[94 ; 109 [	191	19	10%	38	20%
7	[109 ; 129 [	180	34	10%	50	28%
8	[129 ; 150 [	174	31	10%	50	29%
9	[150 ; 185 [	159	111	9%	81	51%
10	[185 ; 1000 [	149	22 071	8%	149	100%
11	[1000 ; +∞ [	13	1 277 965	1%	13	100%

Pour pouvoir évaluer numériquement ce plan d'échantillonnage, nous avons opté pour une approche de type Monte Carlo. Nous aurions pu éviter ce type d'approche dans la mesure où l'estimateur de Horvitz-Thompson est sans biais et que sa variance théorique est exactement calculable. Cependant, pour des personnes non initiées à la statistique, une approche empirique par simulation est plus facile à appréhender qu'une approche "théorique". A partir des données de la commune de Canéjan entièrement télérelevée, nous avons répliqué 25 000 fois le tirage d'un échantillon suivant ce plan. Les 25 000 simulations permettent de calculer la moyenne des estimations hebdomadaires ainsi que l'écart-type correspondant.

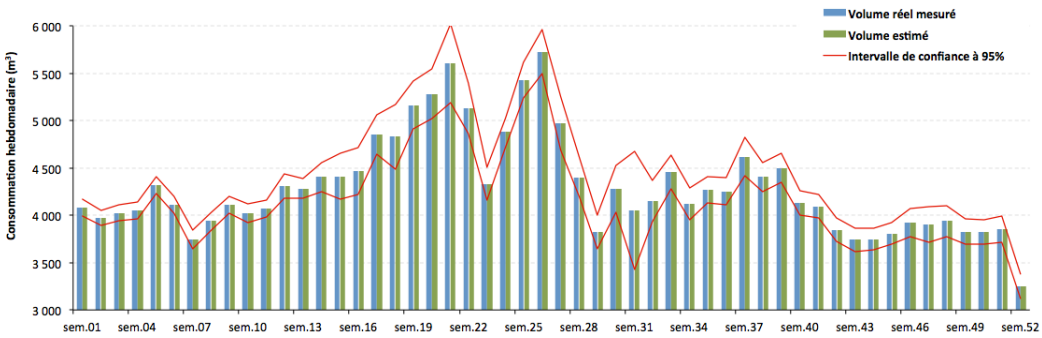

FIGURE 1: Estimation du total de la consommation hebdomadaire

Les simulations illustrent clairement l'absence de biais de l'estimateur comme le montre la Figure 1. Une majorité (64%) des écart-types sont inférieurs au seuil imposé (*cf.* Tableau 5), l'écart-type moyen (resp. médian) étant de 87 m^3 (resp. 74 m^3). Les résultats actuels apparaissent donc satisfaisants.

 ISSN: 2102-6238

TABLEAU 5. *Fractiles pour l'écart-type de l'estimation de la consommation journalière totale pour les 52 semaines.*

1^{er} Quartile des écarts-types hebdomadaires	58
Médiane des écarts-types hebdomadaires	74
Moyenne des écarts-types hebdomadaires	87
$3^{ème}$ Quartile des écarts-types hebdomadaires	106
% d'écarts-types ≤ 91 m^3	64%

2. Redressement d'un échantillon

L'échantillon construit précédemment à l'aide de la consommation annuelle 2010 répond aux exigences imposées : les estimateurs de la consommation hebdomadaire en 2011 obtenus ont une précision moyenne inférieure à 91 m^3. Le coefficient de corrélation entre la variable de stratification et la variable d'intérêt est d'autant plus fort que les deux variables sont proches temporellement, comme le montre la Figure 2.

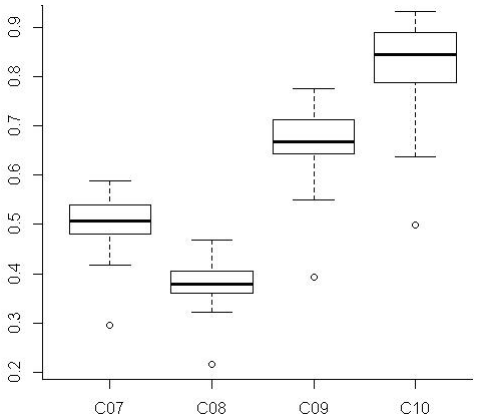

FIGURE 2: Boîte à moustaches des corrélations linéaires entre les 52 consommations hebdomadaires individuelles en 2011 et les consommations annuelles individuelles 2007, 2008, 2009 et 2010

Autrement dit, la qualité de l'estimateur, fondé sur une stratification à l'année A, devrait avoir tendance à se dégrader au fur et à mesure de son utilisation au cours du temps. De plus, l'évolution des comportements des usagers entraîne des mouvements de strates (*stratum jumper*, Rivest, 1999) qui dégrade l'homogénéité initiale des strates. Il convient de prendre en compte toute nouvelle information auxiliaire actualisée disponible, qui sera donc mieux corrélée à notre variable d'intérêt.

Supposons qu'un échantillon ait été créé suivant les préconisations précédemment présentées, afin d'estimer les consommations hebdomadaires durant l'année 2011. Ce même échantillon est réutilisé ultérieurement, pour estimer les consommations hebdomadaires des six premiers mois de l'année 2012. Ce plan de sondage ne semble pas de prime abord être optimal, puisque la variable de stratification est la consommation annuelle en 2010 pour une estimation en 2012. Nous disposons cependant d'une nouvelle information qui est la consommation annuelle individuelle en 2011. Nous proposons d'utiliser cette information auxiliaire, disponible sur toute la population, pour redresser l'estimateur. Nous notons Z cette variable auxiliaire "consommation annuelle

individuelle en 2011". Nous nous intéressons à présent au total des consommations hebdomadaires en 2012. Dans la suite, nous présentons dans un premier temps trois méthodes permettant de redresser l'estimateur d'Horvitz-Thompson, puis évaluons leur performance sur le cas-test de la commune de Canéjan.

2.1. Redressement par post-stratification

Cette méthode permet de stratifier la population, et l'échantillon déjà extrait, selon la variable auxiliaire actualisée. A priori, cette opération est appropriée compte tenu de la variable d'intérêt considérée. "*Toute opération de post-stratification consécutive à un sondage aléatoire simple [...] améliore l'estimation par rapport à la moyenne simple*" (voir Ardilly, 2006). Nous illustrons dans cette section l'effet de la post-stratification sur l'estimateur initial (qui est ici issu d'un sondage stratifié). La méthodologie développée en Section 1 est ré-appliquée afin de découper la population selon la variable Z. On note Γ_k, $k = 1, \ldots, K$, les post-strates de taille respective M_k. On note $A_{kh} = \Gamma_k \bigcap G_h$, de taille Θ_{kh}, l'intersection entre la strate G_h et la post-strate Γ_k. On note γ_k, de taille aléatoire m_k, l'intersection de l'échantillon s et de la post-strate Γ_k. Enfin, on note $\alpha_{kh} = \gamma_k \bigcap g_h$, de taille aléatoire θ_{kh}, l'intersection entre le sous-échantillon g_h et la post-strate Γ_k.

L'estimateur du total post-stratifié s'écrit de la façon suivante (*cf.* Särndal et al., 1992 ou Deville et al., 1993) :

$$\hat{T}_{Y\,post} = \sum_{k=1}^{K} M_k \frac{\sum_{i\in\gamma_k} d_i Y_i}{\sum_{i\in\gamma_k} d_i}, \tag{5}$$

où $d_i = \frac{1}{\pi_i}$. Ainsi, deux estimateurs redressés par post-stratification sont envisageables, si l'on considère un sondage stratifié effectué au préalable.

On peut considérer un premier estimateur post-stratifié, obtenu en réalisant un redressement dans les post-strates Γ_k, ce qui nécessite de connaître les tailles M_k :

$$\hat{T}_{Y\,post\,1} = \sum_{k=1}^{K} M_k \frac{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i\in\alpha_{hk}} Y_i}{\sum_{h=1}^{L} \frac{N_h}{n_h} \theta_{hk}}. \tag{6}$$

Le second estimateur est lui obtenu en réalisant un redressement dans les post-strates A_{kh} ce qui nécessite de connaître les tailles Θ_{hk} :

$$\hat{T}_{Y\,post\,2} = \sum_{h=1}^{L} \sum_{k=1}^{K} \frac{\Theta_{hk}}{\theta_{hk}} \sum_{i\in\alpha_{hk}} Y_i. \tag{7}$$

Nous pouvons remarquer qu'il suffit qu'un sous-échantillon α_{hk} ne soit pas vide dans la post-strate k pour que l'estimateur présenté à l'équation (6) soit calculable alors que le second estimateur ne tolère aucun sous-échantillon vide ($\alpha_{hk} \neq \emptyset, \forall h, \forall k$).
Des méthodes permettent de résoudre ce problème en regroupant des "strates" adjacentes. Il est, par exemple, préconisé dans Jay et al. (2007) de regrouper les intersections pour éviter d'avoir des sous-échantillons vides.

 ISSN: 2102-6238

2.2. *Redressement par régression*

Etant donné la forte corrélation linéaire entre la variable d'intérêt Y et la variable Z, il est raisonnable de supposer l'existence d'une relation de type affine entre ces deux variables :

$$Y_i(t) = \alpha(t) + \beta(t) Z_i + \varepsilon_i(t), \qquad \forall i \in [\![1, N]\!]$$

avec $\sum_{i \in U} \varepsilon_i(t) = 0$. Ainsi, une piste à exploiter est le redressement par régression (voir par exemple Särndal et al., 1992). Notons $\widetilde{\mathbf{Z}}_i$ un vecteur de variables auxiliaires (défini par la suite). Nous considérons $\hat{T}_{\widetilde{Z}}$ l'estimateur du total $T_{\widetilde{Z}}$. Le redressement par régression de $\hat{T}_Y(t)$ s'effectue de la façon suivante :

$$\hat{T}_{Yr}(t) = \hat{T}_Y(t) + \hat{\mathbf{B}}'(t) \left(\mathbf{T}_{\widetilde{\mathbf{Z}}} - \hat{\mathbf{T}}_{\widetilde{\mathbf{Z}}} \right) = \hat{T}_Y(t) + \hat{\mathbf{B}}'(t) \left(\mathbf{T}_{\widetilde{\mathbf{Z}}} - \sum_{i \in s} d_i \widetilde{\mathbf{Z}}_i \right), \tag{8}$$

où

$$\hat{\mathbf{B}}(t) = \left(\sum_{i \in s} d_i \widetilde{\mathbf{Z}}_i \widetilde{\mathbf{Z}}_i' \right)^{-1} \left(\sum_{i \in s} d_i \widetilde{\mathbf{Z}}_i Y_i(t) \right). \tag{9}$$

Deux approches sont possibles, comme dans le cas de la post-stratification : soit on redresse l'estimateur sur toute la population, soit sur chacune des strates.
La première option consiste à redresser, suivant la formule (8), l'estimateur du total $\hat{T}_Y(t)$. Dans ce cas, $\widetilde{\mathbf{Z}}_i' = (1, Z_i)$ et $\hat{\mathbf{B}}(t) = (\hat{\alpha}(t), \hat{\beta}(t))'$. Le paramètre de pente $\hat{\beta}(t)$ est un scalaire, commun à toute la population. L'estimateur redressé est alors noté $\hat{T}_{Yr_1}(t)$.
Dans la seconde option, nous considérons l'estimateur du total comme une somme pondérée d'estimateurs des moyennes par strate (formule (4)). Ainsi, nous redressons par régression tous les estimateurs des moyennes $\hat{\bar{y}}_h(t)$; $\widetilde{\mathbf{Z}}_i' = (I_{1i}, I_{1i}Z_i, I_{2i}, I_{2i}Z_i, \cdots, I_{Li}, I_{Li}Z_i)$ où I_h ($h = 1, \cdots, L$) sont des variables indicatrices indiquant l'appartenance aux strates. Dans ce cas, $\hat{\mathbf{B}}(t) = (\hat{\alpha}_1(t), \hat{\beta}_1(t), \cdots, \hat{\alpha}_L(t), \hat{\beta}_L(t))'$, chaque strate ayant alors sa propre pente $\hat{\beta}_h(t)$. L'estimateur correspondant est noté $\hat{T}_{Yr_2}(t)$.

Remarque : Le redressement par post-stratification est un cas particulier de redressement par régression pour lequel les variables auxiliaires sont des variables indicatrices de l'appartenance aux post-strates ($\widetilde{\mathbf{Z}}_i' = (I_1, I_2, \cdots, I_K)$).

2.3. *Calage*

Une troisième approche de redressement est le calage. A chaque individu i de l'échantillon s est associé un poids de sondage d_i afin de construire l'estimateur de Horvitz-Thompson du total $\hat{T}_{Y\pi} = \sum_s d_i Y_i(t)$. L'idée du calage est de calculer, à partir des poids d_i, de nouveaux poids w_i en tenant compte du vecteur $\widetilde{\mathbf{Z}}$, préalablement défini en section 2. Plus précisément, l'objectif est de trouver les poids w_i solutions du problème d'optimisation suivant :

 ISSN: 2102-6238

$$\begin{cases} \min \sum_{i\in s} D(w_i, d_i) \\ \text{s.c.} \sum_{i\in s} w_i \widetilde{\mathbf{Z}}_i = \sum_{i\in U} \widetilde{\mathbf{Z}}_i, \end{cases}$$

où D est une mesure de distance. A noter que par abus de langage, D est appelée "distance" même si elle ne vérifie pas les 3 critères de définition d'une distance au sens mathématique du terme, notamment le fait que $D(d_i, w_i) = D(w_i, d_i)$. Les poids w_i doivent être proches des d_i, l'usage des poids initiaux garantissant un estimateur sans biais.
Si on note $\delta(w_i, d_i) = \partial D(w_i, d_i)/\partial w_i$, il est alors possible de définir une fonction F, telle que $d_i F(.)$ soit l'inverse de la fonction $\delta(., d_i)$ vérifiant

$$w_i = d_i F(\lambda^\top \widetilde{\mathbf{Z}}_i), \tag{10}$$

Pour plus de détail, le lecteur peut se référer à Tillé (2001). Dans l'équation (10), λ correspond au multiplicateur de Lagrange solution de l'équation :

$$\sum_{i\in s} d_i F(\lambda^\top \widetilde{\mathbf{Z}}_i) \widetilde{\mathbf{Z}}_i = \sum_{i\in U} \widetilde{\mathbf{Z}}_i.$$

L'estimateur redressé par calage généralisé s'écrit :

$$\hat{T}_{Y\,CAL}(t) = \sum_{i\in s} w_i Y_i(t) = \sum_{i\in s} d_i F(\lambda^\top \widetilde{\mathbf{Z}}_i) Y_i(t). \tag{11}$$

Deville and Särndal (1992) définissent une forme généralisée de la fonction de calage F dépendant d'un réel α, forme à partir de laquelle il est possible de calculer la mesure de distance D associée (***cf.*** Tableau 6).

TABLEAU 6. *Exemples de distance et fonction de calage.*

α	$D^\alpha(w_i, d_i)$	$F^\alpha(t)$
$\mathbb{R}\backslash\{0,1\}$	$\dfrac{\frac{w_i^\alpha}{d_i^{\alpha-1}} + (\alpha-1)d_i - \alpha w_i}{\alpha(\alpha-1)}$	$\sqrt[\alpha-1]{1+t(\alpha-1)}$
0	$-d_i \ln(\frac{w_i}{d_i}) + w_i - d_i$	
1	$w_i \ln(\frac{w_i}{d_i}) + d_i - w_i$	$\exp(t)$

Il y a une infinité de possibilités de redressement par calage en fonction du choix de $\alpha \in \mathbb{R}$ intervenant dans la distance D^α, en plus des autres distances existantes (comme la distance euclidienne par exemple). Cependant seules les deux les plus utilisées en pratique sont comparées ici :

- le cas où $\alpha = 1$ (méthode dite du *Raking Ratio*),
- le cas où $\alpha = 2$ (D^α est alors la distance du χ^2).

Les distances correspondantes sont :

$$D^1(d_i, w_i) = w_i \ln(\frac{w_i}{d_i}) + d_i - w_i \qquad \text{et} \qquad D^2(d_i, w_i) = \frac{(w_i - d_i)^2}{2d_i},$$

 ISSN: 2102-6238

et les fonctions de calages associées sont les suivantes :

$$F^1(t) = \exp(t) \qquad \text{et} \qquad F^2(t) = 1 + t.$$

Un des avantages du choix $\alpha = 1$ ou $\alpha = 2$ est que le problème d'optimisation mène toujours à une solution (voir Deville and Särndal, 1992). Chacune des deux méthodes re-définit les poids initiaux d_i en :

$$\begin{aligned} w_i &= d_i F^1(\lambda^\top \widetilde{\mathbf{Z}}_i) = d_i e^{\lambda^\top \widetilde{\mathbf{Z}}_i} && \text{si } \alpha = 1, \\ w_i &= d_i F^2(\lambda^\top \widetilde{\mathbf{Z}}_i) = d_i (1 + \lambda^\top \widetilde{\mathbf{Z}}_i) && \text{si } \alpha = 2. \end{aligned}$$

Remarque : Le redressement par régression (avec une seule variable auxiliaire Z) correspond à un cas particulier du redressement par calage : celui du calage avec une distance du χ^2 et la variable auxiliaire $\widetilde{\mathbf{Z}} = (1, Z)$. La démonstration de cette relation est développée dans Deville et al. (1993) et plus de détails sont disponibles dans Ardilly (2006).

2.4. *Résultats et discussion*

Nous jugerons la pertinence d'un estimateur redressé à partir de deux critères : son biais et son écart-type. Ces deux critères permettent de calculer l'erreur quadratique moyenne (EQM) :

$$EQM(\hat{T}_{Yr}(t)) = \mathbb{V}(\hat{T}_{Yr}(t)) + \left(\mathbb{E}[\hat{T}_{Yr}(t)] - T_Y(t)\right)^2$$

où $\hat{T}_{Yr}(t)$ est l'estimateur redressé et $\mathbb{V}(\hat{T}_{Yr}(t))$ sa variance. Les résultats numériques proviennent des données de la commune test de Canéjan, où l'on cherche à estimer les consommations totales hebdomadaires pour les 6 premiers mois de l'année 2012. Nous réalisons 25 000 itérations du tirage de l'échantillon, selon le plan de sondage décrit au Tableau 4, pour calculer l'estimateur de Horvitz-Thompson et les estimateurs redressés en utilisant la consommation annuelle 2011 comme variable auxiliaire.

2.4.1. *Résultats du redressement par post-stratification*

Le découpage de la population suivant la variable auxiliaire Z et la méthodologie précédemment développée à la Section 1.1 renvoie les 18 post-strates définies dans le Tableau 7.

TABLEAU 7. *Post-strates définies sur la population selon la variable auxiliaire Z*

Post-str	M_k	M_k/N	Borne inf	**Post-str**	M_k	M_k/N	Borne inf
1	108	6%	0	**10**	116	6%	102
2	98	5%	21	**11**	114	6%	111
3	110	6%	36	**12**	106	6%	122
4	109	6%	48	**13**	104	6%	133
5	121	7%	58	**14**	98	5%	146
6	112	6%	67	**15**	92	5%	162
7	114	6%	75	**16**	89	5%	187
8	114	6%	84	**17**	85	5%	231
9	120	7%	93	**18**	12	1%	1000

Concernant les deux estimateurs redressés par post-stratification que nous avons présentés, seul le premier est réellement exploitable. En effet, le risque non négligeable d'avoir des échantillons

 ISSN: 2102-6238

d'effectif nul dans les intersections strates/post-strates (α_{hk}) biaise l'estimateur $\hat{T}_{Y\,post\,2}$ et rend le redressement inefficace. Sur nos 25 000 simulations, toutes se sont retrouvées dans un cas où au moins un effectif θ_{hk} était nul (pour un effectif $\Theta_{hk} \neq 0$). Nous avons donc dû appliquer la méthode de fusion des intersections strates-post-strates.
Par exemple, lors une simulation en particulier, l'intersection de la strate 1 et de la post-strate 8 a un effectif d'un individu ($\Theta_{1,8} = 1$) alors que $\alpha_{1,8}$ est vide ($\theta_{1,8} = 0$). L'intersection adjacente ($A_{1,9}$) est composée de 2 individus dont un faisant partie de l'échantillon ($\Theta_{1,9} = 2$ et $\theta_{1,9} = 1$). Ces deux groupes sont alors fusionnés pour former un groupe composé de 3 individus ($\Theta_{1,(8+9)} = 3$) dont 1 échantillonné ($\theta_{1,(8+9)} = 1$).

Nous ne nous attarderons pas ici sur cette méthode du fait des nombreux calculs qu'elle engendre (18×11 calculs d'estimateurs) et de la piètre qualité des estimateurs qui en découle. Le premier estimateur post-stratifié $\hat{T}_{Y\,post\,1}$ est quant à lui plus performant, du fait que l'on découpe l'échantillon en 18 post-strates et non pas 18×11 intersections strates/post-strates. Le redressement permet de réduire l'écart-type de l'estimateur initial comme on peut le voir dans le Tableau 8.

2.4.2. *Résultats du redressement par régression*

Comme dit précédemment, les deux méthodes de redressement par régression ont des écritures proches. Le fait d'effectuer une régression par strate ne se justifie que dans le cas où les valeurs des coefficients de pente $\hat{\beta}_h(t)$ sont différents d'une strate à l'autre. Analysons le comportement entre la variable d'intérêt et la variable auxiliaire dans chaque strate. La variable d'intérêt étant la consommation individuelle hebdomadaire étudiée les six premiers mois de l'année 2012, il existe 26 régressions entre Y et Z. Nous ne présentons ici qu'un exemple (la semaine 02), mais les résultats numériques traitent de l'ensemble des estimations hebdomadaires. La pente estimée $\hat{\beta}(02)$ de la droite de régression calculée sur la population est égale à 0.0193. L'étude par strate, illustrée par la Figure 3, suggère des relations linéaires entre la variable d'intérêt et la variable auxiliaire considérée.

L'étendue des $\hat{\beta}_h(02)$ va de 0.0113 à 0.0366, pour une valeur moyenne de 0.0194, proche de la valeur de $\hat{\beta}$. Si on regarde en particulier la strate 10, dont le coefficient de pente est nettement différent de $\hat{\beta}$, on remarque que cette différence est due à la présence d'un seul point atypique (distance de Cook = 52). Ainsi, la droite calée sur la strate est très fortement influencée par cet individu ayant une valeur importante pour Z (> 2 000 m^3). La pente de régression sur les individus de la strate 10 sans ce point est de 0.014 et elle est donc très similaire à celle de la régression sur toute la population.

Ainsi, il n'y a pas de raison évidente de faire une régression par strate, les deux méthodes de redressement par régression devraient fournir des résultats d'estimation équivalents, comme le corroborent les résultats du Tableau 8. Les deux méthodes réduisent l'écart-type par rapport à celui de l'estimateur initial, ce qui illustre l'efficacité du redressement par régression.

 ISSN: 2102-6238

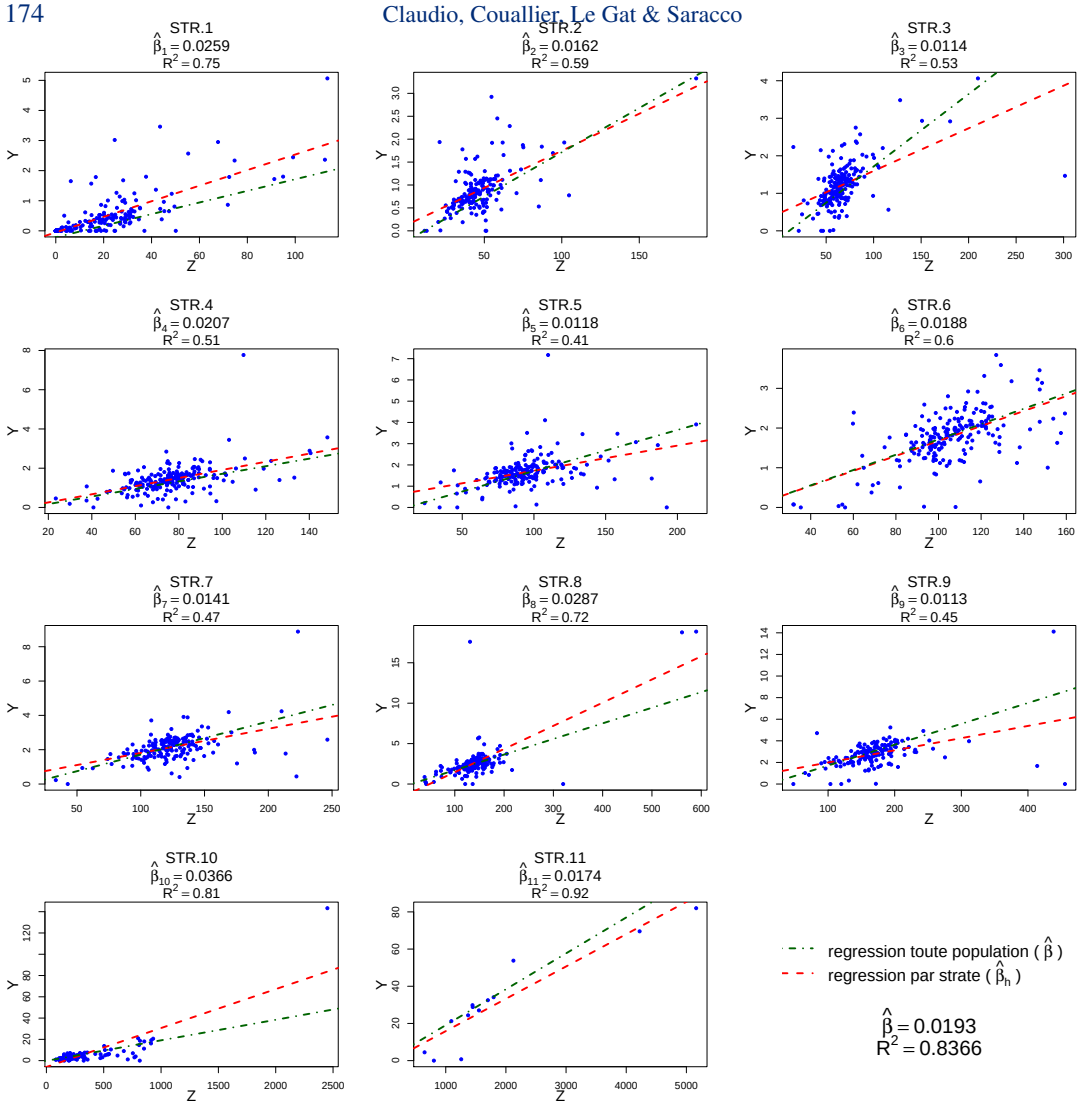

FIGURE 3: Régression linéaire entre la variable d'intérêt, pour la semaine 02, $Y(02)$ et la variable auxiliaire Z pour chaque strate (en m^3)

2.4.3. Résultats du calage

Comparons les deux cas de calage avec $\alpha = 1$ et $\alpha = 2$, les calculs des écart-types des deux estimateurs correspondants sont présentés dans le Tableau 8. Quelle que soit la distance choisie (parmi les deux présentées), les performances du calage sont identiques et permettent de réduire l'écart-type de l'estimateur initial. Cela confirme que les méthodes de calage, suivant une fonction F comme définie au Tableau 6, sont asymptotiquement équivalentes (voir Deville and Särndal, 1992). Il n'y a donc pas de raison particulière de préférer une méthode plutôt qu'une autre.

 ISSN: 2102-6238

2.4.4. *Choix de la méthode de redressement*

Un premier regard sur les résultats du redressement (fournis au Tableau 8) montre que la post-stratification sur post-strates A_{kh} (méthode 2) est la moins performante des différentes méthodes dans la mesure où, contrairement aux autres techniques de redressement, elle augmente l'erreur quadratique moyenne (augmentation de 27%). .

TABLEAU 8. *Comparaison des méthodes de redressement : analyse des estimateurs du total*

Méthode	EQM	Biais moyen	Ecart-type median	$\sigma \leq 91$ m^3 (%)	$f^{(a)}$	Δ coût (k€)$^{(b)}$
Stratification	1.56 10^4	0.3	94	48%	35%	
Post-strat. (méth.1)	1.48 10^4	1.1	88	52%	36%	1.3
Post-strat. (méth.2)	1.98 10^4	22.5	103	33 %	31%	-5.1
Régression (méth.1)	1.45 10^4	0.4	84	56%	37%	2.5
Régression (méth.2)	1.50 10^4	19.4	84	52%	37%	2.5
Calage ($\alpha = 1$)	1.46 10^4	0.5	84	56%	37%	2.5
Calage ($\alpha = 2$)	1.46 10^4	0.6	84	56%	37%	2.5

a. Taille d'échantillon nécessaire pour obtenir la même précision sans redressement.
b. Investissements nécessaires (<0) ou économisés (>0), pour un coût unitaire de 70€ par compteur équipé.

De nouveau, si nous comparons les méthodes par rapport à l'EQM, la post-stratification sur les post-strates Γ_k (méthode 1) et la régression par strate sont les méthodes de redressement qui donnent les résultats les moins performants (réduction respective de l'EQM de 5% et 4% contre 7% pour les autres). Ces deux méthodes ne sont donc pas retenues dans notre étude. Concernant les méthodes restantes (la régression sur la population et les deux méthodes de calage), on constate que les méthodes sont équivalentes. Même si le biais diffère d'une méthode à l'autre, compte tenu des ordres de grandeur du biais et de l'écart-type, le premier est négligeable face au second et l'EQM dépend essentiellement de l'écart-type de l'estimateur.

Nous avons déjà vu l'équivalence des méthodes de calage quelle que soit la fonction de calage choisie, analysons alors en détail les résultats des méthodes de calage ainsi que du redressement par régression (méthode 1). La Figure 4 illustre les valeurs des écarts-types hebdomadaires pour l'estimateur stratifié et les estimateurs redressés.

Aucune méthode ne se détache des autres. L'écart moyen (en valeur absolue) entre l'écart-type des méthodes de redressement par la régression et par calage étant de 0.6 m^3 (pour un écart-type moyen de 200 m^3), les trois méthodes ont sensiblement les mêmes performances en termes de précision, et apparaissent clairement meilleures que l'estimateur stratifié non redressé. Cela confirme empiriquement le fait que ces méthodes sont asymptotiquement équivalentes comme démontré dans Deville et al. (1993).

Cependant s'il fallait choisir entre ces deux techniques (redressement par régression ou calage), le choix se porterait sur le redressement par régression (globale). D'une part, cette méthode exploite la relation linéaire qui existe entre la variable auxiliaire et la variable d'intérêt ; d'autre part, pour des personnes non initiées à la statistique, l'approche par régression est plus facilement compréhensible.

 ISSN: 2102-6238

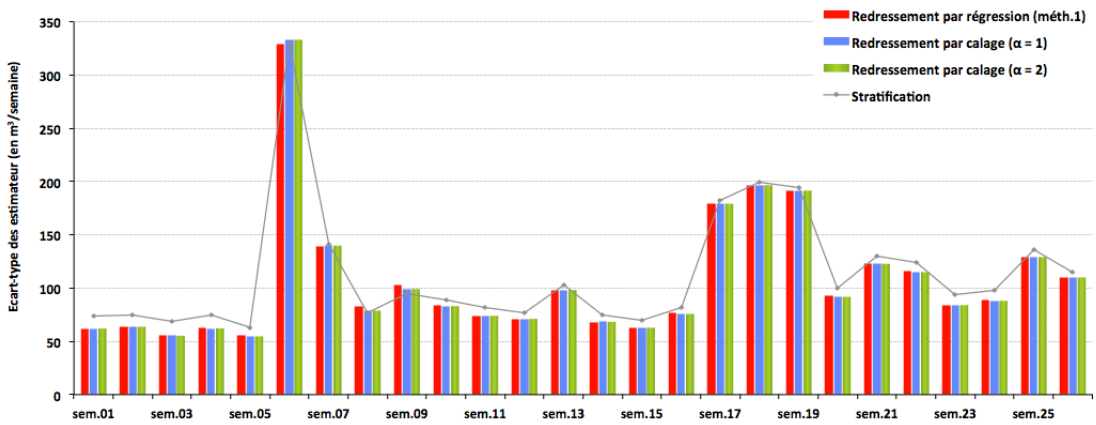

FIGURE 4: Estimation par Monte Carlo de l'écart-type σ hedomadaire de l'estimateur stratifié et des estimateurs redressés par calage et régression.

2.4.5. Impact économique du redressement

Un autre moyen d'améliorer la précision de l'estimateur sans le redresser aurait été d'augmenter la taille de l'échantillon. Cependant l'équipement supplémentaire de compteur en télérelevé engendre un coût opérationnel, alors que le redressement est une méthode qui ne nécessite ici aucun investissement financier dans la mesure où les données auxiliaires sont déjà à disposition par l'opérateur. Nous décidons ainsi d'évaluer l'impact économique du redressement (voir Tableau 8) en calculant la taille d'échantillon nécessaire pour atteindre les mêmes précisions pour les estimateurs.

Mis à part le cas de la post-stratification sur les sous-totaux par strates (méthode 2), le redressement permet de réaliser des économies d'investissement allant de 1.3 à 2.5 k€. Ces montants paraissent faibles compte tenu de la valeur de certains contrats de délégation des services d'eau mais en transposant ces résultats à des contrats plus importants comme celui de la Communauté Urbaine de Bordeaux, ces économies peuvent aller jusqu'à 308 k€, pour un écart de seulement un an entre la pose des émetteurs télérelevés (*i.e.* la constitution de l'échantillon stratifié) et l'usage d'un estimateur redressé. Ces montants tendent à augmenter naturellement au cours du temps.

Conclusion

La méthodologie, décrite en première partie de cet article, permet d'estimer efficacement le total des consommations hebdomadaires, au vu des objectifs fixés en termes de précision ; l'estimation des volumes consommés devant permettre d'évaluer efficacement les pertes d'eau en réseau (*pertes = volumes distribués - volumes consommés*). Le plan de sondage, aussi efficace soit-il, devient rapidement obsolète à cause de la dégradation de la corrélation entre la variable d'intérêt et la variable de stratification. L'estimateur initial restant malgré tout sans biais, c'est essentiellement la précision de cet estimateur qui se dégrade et qu'il est nécessaire d'améliorer. Les techniques de redressement, présentées en seconde partie, permettent quasiment toutes de répondre à cette

 ISSN: 2102-6238

problématique (mise à part la post-stratification qui est délicate à exploiter dans le cadre d'un sondage initial stratifié). Le redressement par régression est la méthode la plus pertinente ici ; elle fait appel à la relation quasi-linéaire qu'il y a entre la variable d'intérêt et une variable auxiliaire actualisée. De manière globale, le redressement est une technique qui augmente la précision de l'estimateur initial sans pour autant requérir ici un quelconque investissement d'un point de vue économique et financier.

L'estimation fiable des consommations totales permet alors de calculer sur un pas de temps hebdomadaire (ou journalier, la méthodologie décrite ici restant la même) les pertes d'eau en réseau. En travaillant en temps réel, il serait alors possible de détecter et de quantifier les fuites d'eau.

Remerciement : Les auteurs souhaitent remercier l'éditeur en chef, les éditeurs associés ainsi que les deux relecteurs anonymes pour leurs commentaires et leurs remarques constructives qui ont permis une amélioration substantielle de l'article.

Références

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip.

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100(3) :555–569.

Cochran, W. (1977). *Sampling Techniques, 3rd Edition*. Wiley Series.

Dalenius, T. (1950). The problem of optimum stratification. *Skand. Akt. Tidskrift*, page 203.

Dalenius, T. and Hodges, J. (1959). Minimum variance stratification. *American Statistical Society*, 54 :88–101.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382.

Deville, J., Särndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423) :1013–1020.

Fellegi, I. (2010). Méthodes et pratiques d'enquête. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-fra.pdf. Statistique Canada.

Jay, J. K., Li, J., and Valliant, R. (2007). Regroupement de cellules lors de la post-stratification. *Techniques d'enquête*, 33(2) :157–170.

Kpedekpo, G. (1973). Recent advances on some aspects of stratified sample design. A review of the literature. *Metrika*, 20(1) :55–64.

Lavallée, P. and Hidiroglou, M. (1988). On the stratification of skewed population. *Techniques d'enquête*, 14 :35–45.

Nicolini, G. (2001). A method to define strata boundaries. *Università degli Studi di Milano*.

Rivest, L. (1999). Stratum jumpers : can we avoid them. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Särndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Serfling, R. (1968). Approximately optimal stratification. *Journal of the American Statistical Association*, 63(324) :1298–1309.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en population finie*. Dunod.

 ISSN: 2102-6238