Journal de la Société Française de Statistique
Vol. 156 No. 1 (2015)

Estimation de synchrones de consommations électriques avec calage dynamique sur des données de relevés d'index asynchrones

Title: Aggregated load curves estimation in presence of asynchronous individual meter data

Anne De Moliner [1]

Résumé : Afin d'appuyer les directions opérationnelles d'ERDF qui doivent équilibrer l'offre et la demande d'électricité à tout instant, les statisticiens d'EDF R&D ont besoin d'estimer le plus précisément possible la consommation d'électricité au pas demi-horaire, au périmètre de différents groupes de clients. Ces estimations sont réalisées à partir d'échantillons de clients, et recalées à l'aide des relevés d'index de consommation électriques disponibles pour l'exhaustivité de la population : la différence entre les index relevés à deux dates correspond en effet à la consommation électrique totale du client entre ces dates. Cependant ces relevés de consommation, bientôt disponibles à un rythme mensuel, sont avant tout destinées à la facturation et donc sont réalisés à des dates différentes pour chacun. On sort donc du cadre classique du calage, où l'information auxiliaire recouvre la même réalité pour toute la population. Dans cet article, nous proposons d'adapter les techniques usuelles du calage à notre contexte où on estime des courbes à partir d'une information auxiliaire non synchrone qui évolue au cours du temps. Les méthodes proposées seront testées sur un jeu de données réelles.

Abstract: The French Company Electricité de France needs to estimate as precisely as possible the electricity consumption at an aggregated level at an half-hourly timestep. These estimations are carried out based on customers samples and can be enhanced by using individual metering data collected every six months (every month in the near future) for each client of the population. However, these metering data are gathered for billing purposes so the metering does not occur simultaneously for all the population. For this reason, we cannot apply directly the usual calibration techniques so, in this article, we propose to adapt them to our context where we estimate an aggregated curve using an asynchronous auxiliary information evoluting over time. The performances of our methods will be assessed on a real dataset.

Mots-clés : Redressement, Données fonctionnelles, Courbes de charge, Industrie
Keywords: Calibration, Functional data, Industry
Classification AMS 2000 : 62D05

1. Introduction et contexte

EDF a besoin d'estimer de manière très précise la consommation de groupes de clients à chaque demi-heure sur de longues périodes. Ces données agrégées, aussi appelées courbes de charge agrégées, permettent ensuite de mener des études de connaissances client et d'évaluer l'impact de nouveaux usages ou équipements électriques. Elles permettent également à ERDF d'assurer sa mission d'opérateur régulé dans le cadre des mécanismes de marché. Afin de construire nos estimations, nous disposons de panels de quelques milliers de ménages dont la consommation électrique est mesurée toutes les demi-heures pendant plusieurs années.

[1] EDF R&D ICAME
E-mail : anne.de-moliner@edf.fr

 ISSN: 2102-6238

Dans un futur proche, avec le déploiement massif de compteurs communicants, nous disposerons de relevés d'index de consommations électriques totales au pas de temps mensuel pour l'ensemble des clients de la population, contre un pas de temps semestriel actuellement. De plus, la taille de nos échantillons sera amenée à croître significativement. En outre, les relevés d'index (que nous appellerons également par la suite "données de relève") représentent une source d'information extrêmement riche que l'on cherchera à exploiter au maximum pour améliorer les estimations de courbes de charge agrégées. Néanmoins, elles sont collectées avant tout dans une optique de facturation et non spécifiquement pour améliorer les estimations, c'est pourquoi elles seront très régulières mais pas synchrones : les index de consommation de certains clients seront relevés tous les six du mois, d'autres le quinze, etc...

Cet asynchronisme risque de détériorer l'efficacité des techniques de redressement comme le calage sur marges par exemple : en effet, pour estimer a posteriori la consommation d'un agrégat de clients à un instant donné, on cherchera à utiliser la consommation de la période de relève en cours (i.e. la consommation déduite de la différence entre le relevé d'index situé immédiatement avant l'instant et celui situé immédiatement après) comme variable de calage car cette donnée fournit une information précieuse sur le niveau moyen de consommation du client. Toutefois les périodes de relèves en cours sont différentes d'un individu à l'autre, ce décalage pouvant atteindre plusieurs semaines si les relevés sont mensuels, or en raison notamment de la forte prévalence du chauffage électrique en France, la consommation est très sensible aux températures et un décalage de quelques jours dans les dates de relevés peut influer fortement sur la consommation mensuelle si par exemple un pic de froid survient dans cet intervalle de temps. Le décalage temporel des dates de relevés vient donc "parasiter" l'information fournie par les données de relevés d'index et dégrade donc l'efficacité du calage. Dans cet article, nous allons donc proposer une méthode permettant d'adapter au mieux les techniques de calage usuelles afin de limiter l'impact de l'asynchronisme dans notre contexte.

Après avoir présenté la technique naïve de calage, consistant à ignorer le caractère asynchrone des relevés, et mis en lumière ses insuffisances, nous présenterons la méthode que nous préconisons, consistant à découper la population en différents groupes en fonction des dates de relevés et à tenir compte de ces groupes dans le calcul des poids. Lorsque l'échantillon est très conséquent, il est possible de découper la population en autant de groupes qu'il y a de dates de relevés possibles (trente pour des relevés au pas mensuel, en considérant qu'aucun index n'est relevé le 31 du mois). Nous proposerons également une méthode alternative, basée sur moins de groupes de dates de relevés qui sera donc pertinente sur les échantillons de taille plus modérée. Enfin, nous appliquerons ces techniques sur un jeu de données de courbes de charges réelles et évaluerons leur performance.

2. Calage sur information auxiliaire dynamique

2.1. Hypothèses et notations

Soit une population finie U constituée d'individus indicés de 1 à N dans laquelle on tire selon un plan de sondage aléatoire et connu un échantillon s de n individus. Chaque individu a une probabilité π_i strictement positive et connue de se trouver dans l'échantillon, et on lui affecte un

 ISSN: 2102-6238

poids de sondage de Horvitz and Thompson (1952) égal à l'inverse de sa probabilité d'inclusion $d_i = \frac{1}{\pi_i}$. Les poids après calage seront quant à eux notés w_i.

Soit $y_i(t)$ la consommation de l'individu i à l'instant t, t allant de 1 à T. On cherche à reconstituer la courbe de charge de l'ensemble de la population : $(Y_t)_{t=1\ldots T}$ avec $Y_t = \sum_U y_i(t)$, estimé par $\hat{Y}(t) = \sum_s w_i(t) y_i(t)$. [1] Cette estimation sera faite a posteriori, à l'issue d'une période longue (tous les ans par exemple), et on suppose qu'au moment de l'estimation on dispose pour l'ensemble de la population et l'ensemble des instants de la période d'estimation de données de relèves dites "encadrantes", c'est-à-dire de différences entre le dernier index relevé avant l'instant et le premier index relevé après.

Le plan de sondage envisagé dans notre contexte est un plan de sondage stratifié en fonction du contrat du client, de sa consommation de l'année précédente déduite de ses dernières factures, et aussi le cas échéant de son régime d'heures creuses (c'est-à-dire les heures de la journée au cours desquelles le prix de l'électricité est plus bas), cette dernière variable ayant un impact fort sur l'allure des consommations.

Par ailleurs, on supposera que les index de chaque individu de la population sont relevés le même jour chaque mois à minuit (par exemple, tous les 5 du mois), aucun index n'étant relevé le 31. Pour le mois de février, les index théoriquement relevés le 29 ou le 30 le seront au dernier jour du mois. On considèrera que la donnée de relève, qui sera à terme collectée à l'aide de compteurs communicants, est remontée sans erreur ni valeur manquante et que les dates et heures de relevés sont rigoureusement respectées. De plus, on fera l'hypothèse que les dates de relevés des index des clients, définies dans une optique de facturation, résultent d'un processus aléatoire totalement indépendant des consommations, que ce soit en niveau ou en forme, et même des caractéristiques des clients. Le tirage de l'échantillon sera également indépendant des dates de relevés.

A chaque instant, on dispose de la consommation mensuelle totale sur la période de relève en cours, c'est-à-dire entre le relevé précédant immédiatement l'instant et celui qui le suit immédiatement, notée $x_i(t)$. On notera $X(t) = \sum_U x_i(t)$ la somme des consommations issues des données de relèves encadrantes à l'instant t sur l'ensemble de la population.

Dans le paragraphe suivant, nous présenterons la méthode de calage naïve qui consiste à utiliser telle quelle cette information auxiliaire.

2.2. *Le problème classique du calage*

Dans le cadre classique des sondages (étendu à notre problématique où les variables d'intérêt sont issues de la discrétisation d'une courbe), on utilise l'estimateur $\hat{Y}(t) = \sum_s w_i(t) y_i(t)$ avec les poids $w_i(t)$ les plus proches des poids de sondage d_i, au sens d'une métrique prédéfinie et telle que les contraintes : $\sum_s w_i(t) x_i(t) = X(t)$ soient vérifiées pour tout t. D'autres contraintes sur un vecteur $\mathbf{Z} = (Z_1, \ldots Z_p)'$ constitué d'autres variables auxiliaires (par exemple, les puissances souscrites des clients ou leur localisation géographique) peuvent également s'ajouter au problème.

Pour résoudre ce problème, Deville and Särndal (1992) ont appliqué la méthode des multiplicateurs de Lagrange. Ainsi, les poids de calage sont égaux à $w_i = d_i F(x_i' \lambda)$ où λ est le vecteur des multiplicateurs de Lagrange et $F(.)$ la fonction de calage. Le vecteur λ est déterminé à partir des équations de calage.

[1] Alors que le poids de sondage d_i est constant au cours du temps, le poids après calage $w_i(t)$ est amené à fluctuer

 ISSN: 2102-6238

Différentes méthodes de calage ont été définies à partir de différentes distances entre d_i et w_i dont la méthode linéaire, la méthode exponentielle, la méthode logit et la méthode linéaire tronquée (cf. Deville and Särndal, 1992).

Ces méthodes permettent d'obtenir des résultats cohérents avec les données disponibles par ailleurs mais aussi, lorsque l'on utilise des variables de calage fortement corrélées avec la variable d'intérêt, d'améliorer la précision de l'estimateur en incorporant une information auxiliaire pertinente. Cela permet également de réduire d'éventuels biais de couverture et de non réponse.

2.2.1. Biais et variance

L'estimateur du total après calage possède un biais dont l'ordre de grandeur est en $\frac{1}{n}$, ce biais sera donc faible pour des échantillons de grande taille. De plus, on peut approximer la variance de l'estimateur du total après calage ; pour cela on calcule le résidu de la régression de la variable expliquée $y_i(t)$ sur le vecteur des variables explicatives (les données de relevés d'index et d'éventuelles autres variables auxiliaires) $\mathbf{X}_i(t) = (x_i(t), \mathbf{Z}_i')'$:

$$y_i(t) = \beta(t)\mathbf{X}_i(t) + e_i(t), \forall i \in s$$

On estime $e_i(t)$ par $\hat{e}_i(t) = y_i(t) - \hat{\beta}(t)\mathbf{X}_i(t)$ avec $\hat{\beta}(t) = (\sum_s \mathbf{X}_i(t)\mathbf{X}_i'(t)/\pi_i)^{-1}(\sum_s \mathbf{X}_i(t)y_i(t)/\pi_i)$ l'estimateur des moindres carrés. La variance de l'estimateur du total calé peut ensuite être estimée (Ardilly, 2006, voir par exemple) par :

$$\hat{V}(\hat{Y}(t)) = \sum_{i\in s}\sum_{j\in s} A_{ij}\hat{e}_i(t)\hat{e}_j(t)$$

avec $A_{ij} = \frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}$ et π_{ij} la probabilité d'inclusion double des individus i et j.

2.2.2. Application directe : le calage "naïf"

L'application directe de la technique classique de calage sans adaptation à l'asynchronisme des relevés, sera dénommée par la suite **"calage naïf"**. Intuitivement, on peut penser que cette méthode n'est pas optimale dans notre contexte. En effet, puisque les relevés sont asynchrones, l'information $x_i(t)$, ne mesure pas la même réalité pour tous les individus : par exemple, pour le calage du 10 novembre, la consommation utilisée sera celle du 9 novembre au 9 décembre pour certains et celle du 11 octobre au 11 novembre pour d'autres. Il est donc fort possible que, pour deux clients aux consommations similaires, l'information collectée soit différente du fait de conditions météorologiques différentes pour les deux mois entre les deux relevés ou encore de l'impact de particularités calendaires (vacances scolaires, jours fériés,...). Par exemple si le mois de novembre est plus froid que le mois d'octobre, celui des deux clients dont la consommation est relevée le plus tardivement apparaîtra à tort comme consommant plus que l'autre. Cela conduira donc à donner des poids de calage différents à ces deux individus pourtant similaires, ce qui apparaît contre-intuitif. On voit donc bien que la fluctuation des dates de relevés vient diluer l'information et risque de dégrader l'efficacité du calage.

 ISSN: 2102-6238

Une autre façon de voir les choses pourrait être de se dire qu'on souhaite caler la consommation à l'instant t à l'aide de la consommation du mois qui l'entoure (débutant quinze jours avant et finissant quinze jours après). Cette information $x_i^*(t)$, identique pour tous, n'est disponible qu'entachée d'une "erreur de mesure" qui correspond au décalage des dates de relevés : $x_i(t) = x_i^*(t) + \varepsilon(t)$. Ici l'erreur $\varepsilon(t)$ n'est pas d'espérance nulle puisqu'elle correspond en fait aux différence de consommations entre les périodes, liées notamment au climat ou aux particularités calendaires.

Dans la section suivante, nous allons donc proposer des modifications du calage naïf consistant à regrouper les clients dont les index sont relevés aux mêmes dates (paragraphe 2.3.1) ou approximativement aux mêmes dates (paragraphe 2.3.2).

2.3. Regroupement des clients par date de relevé d'index

On cherche à se ramener au cas habituel du calage sur une information auxiliaire identique (ou approximativement identique) pour chaque individu afin de limiter la dégradation du pouvoir explicatif de notre variable de calage due à l'asynchronisme des relevés d'index. Pour cela, l'idée est de séparer les clients en plusieurs groupes en fonction de leur jour de relevé puis de réaliser le calage indépendamment sur chacun des groupes : chacun des sous-échantillons ainsi constitué sera calé sur l'information de la sous-population issue de la population entière correspondant aux mêmes dates de relevé. Ainsi, dans chaque groupe, le calage sera réalisé sur une variable représentant (au moins approximativement) la même réalité pour tous les individus.

On pourra donc espérer que cette technique permette d'utiliser une information auxiliaire plus corrélée avec les variables d'intérêt car moins "bruitée" par le mécanisme des relevés d'index ; le résidu de la régression des variables d'intérêt sur l'information auxiliaire aura donc vraisemblablement une variance plus faible et donc la variance de l'estimateur par calage sera vraisemblablement plus faible également.

On peut voir cette technique comme un calage par domaines, les domaines étant les groupes définis en fonction des jours de relevés. Comme nous avons supposé que les dates de relevés étaient indépendantes de la consommation, les clients des différents groupes n'ont a priori aucune raison d'avoir des consommations ou des propriétés différentes. Ici, les domaines ne sont donc pas délimités de manière à regrouper des clients similaires et donc à inclure une information supplémentaire mais uniquement afin de s'adapter à une contrainte externe (les relevés asynchrones).

Formellement, on peut écrire ce calage de la manière suivante : On notera $g(i)$ le groupe auquel appartient un individu i. Le nombre total de groupes G sera compris entre 1 (le calage naïf présenté plus haut) et 30 (on crée un groupe différent pour chaque jour de relevé).

L'ensemble des clients de la population (respectivement de l'échantillon) appartenant au groupe g sera noté U_g (respectivement s_g). On crée ensuite des variables composées $x_i^{*g}(t) = x_i(t)\mathbf{1}_{g(i)=g}$ valant $x_i(t)$ si le client i appartient au groupe g et 0 sinon. Dans le cas du calage en un seul groupe, on a évidemment $x_i^{*g}(t) = x_i(t)\,\forall i$. Le calage sera donc réalisé sur ces g variables issues des données de relèves que nous regrouperons en un vecteur $\mathbf{X_i}(t) = (x_i^{*1}(t), \ldots x_i^{*G}(t))'$ et on notera $\mathbf{X}(t) = \sum_U \mathbf{X}_i(t)$ Pour un instant t donné, les contraintes de calage s'écriront donc

$$\sum_{i \in s} w_i(t)\mathbf{X_i}(t) = \mathbf{X}(t)$$

 ISSN: 2102-6238

On peut également les écrire de manière alternative :

$$\sum_{i \in s_g} w_i(t) x_i^{*g}(t) = \sum_{i \in U_g} x_i^{*g}(t) \, \forall g \in [1, \dots G]$$

Si on le souhaite, on peut également ajouter au vecteur $\mathbf{X_i}(t)$ des variables de calage additionnelles correspondant à d'autres caractéristiques pertinentes des clients.

Dans les deux paragraphes suivants, nous évoquerons un peu plus en détail les deux cas particuliers de cette méthode que nous avons mis en oeuvre sur nos jeux de données et comparés au calage naïf : le calage en trente groupes et le calage en quatre groupes.

2.3.1. Version en trente groupes

La version "extrême" de notre méthode de calage consiste à séparer la population en trente groupes : un pour chaque jour de relevé possible. Ainsi le calage, effectué indépendamment pour chaque groupe s'il n'y a pas de variables auxiliaires additionnelles, porte sur une information auxiliaire qui concerne exactement la même période pour tous les individus du groupe et n'est donc pas du tout "bruitée" par le mécanisme de relevé d'index.

Cette méthode ne pose pas de problèmes de mise en oeuvre dans le cas où on dispose d'un très grand nombre de courbes de charge (à terme, nous disposerons de dizaines voire de centaines de milliers de clients dans nos panels, cette méthode sera donc pertinente). Cependant, sur des échantillons plus réduits, le fait de diviser en 30 un échantillon de taille déjà modérée risque de provoquer une certaine instabilité numérique et parfois d'éloigner assez fortement les poids de leur valeur d'origine. C'est pourquoi nous allons proposer dans le paragraphe suivant une version qui représente un compromis entre le calage naïf et le calage en trente groupes.

2.3.2. Version en quatre groupes

Comme précédemment, on souhaite réduire l'impact de l'asynchronisme en regroupant les clients dont les index sont relevés à des dates proches. Cependant on propose de ne constituer qu'un nombre restreint de groupes de relevés (quatre groupes par exemple), qui ne sont plus synchrones (c'est-à-dire que les index sont relevés à la même date) cette fois, mais simplement "à peu près synchrones". Cela permet une plus grande stabilité, en conservant des tailles de sous-échantillons plus importantes. Cette méthode sera donc préconisée pour les échantillons de taille plus modérée.

Une manière de constituer les groupes pourrait par exemple être la suivante :
- groupe 1 : clients dont les index sont relevés entre le 1 et le 8 inclus
- groupe 2 : clients dont les index sont relevés entre le 9 et le 16 inclus
- groupe 3 : clients dont les index sont relevés entre le 17 et le 24 inclus
- groupe 4 : clients dont les index sont relevés à partir du 25 inclus

Au sein d'un même groupe, les relevés sont "à peu près synchrones" : par exemple dans le deuxième groupe, on peut dire que le relevé a lieu le 12 à trois jours ou quatre jours près. On fera donc "comme si" l'information auxiliaire collectée pour chacun des membres de ce groupe

 ISSN: 2102-6238

correspondait à la consommation de cette période fixe, avec une légère erreur de mesure due au décalage de date. Cette information auxiliaire servira à caler l'ensemble des instants entre le 12 d'un mois donné et le 12 du mois suivant.

L'impact des dates de relèves asynchrones sera donc réduit par la constitution des groupes (mais moins que pour 30 groupes) et donc on espère gagner sur la variance de l'estimateur calé.

3. Application à des courbes de charges réelles

Nous avons testé les performances des trois méthodes présentées ici (calage naïf, calage en trente groupes et calage en quatre groupes) sur un jeu de données réelles, constitué de 28000 courbes de clients industriels pour une période allant de septembre à novembre. Nous avons aussi utilisé les courbes d'août et de décembre pour constituer l'information auxiliaire dans le cas des relevés asynchrones. Sur ces données, nous avons simulé le tirage de $R = 200$ échantillons de 830 clients par sondage aléatoire simple stratifié en fonction de la durée d'utilisation (caractéristique tarifaire égale au ratio de la consommation totale sur la puissance souscrite, et liée à la forme globale de la courbe) et des consommations de l'année précédente (100 classes dont les seuils ont été optimisés par la méthode de Dalénius – Dalenius and Hodges 1959). Les tailles d'échantillon dans chaque strate sont déterminées par allocation optimale.

Nous avons simulé des relevés synchrones (les index de tous les clients sont relevés le premier du mois à minuit) et asynchrones (l'index de chaque client est relevé tous les mêmes jours du mois à minuit, ce jour étant tiré aléatoirement avec une probabilité uniforme entre 1 et 30). Lorsque les relevés sont synchrones, les techniques de calage en trente ou en quatre groupes ne peuvent pas être mises en oeuvre car tous les clients appartiennent au même groupe. Néanmoins elles n'auraient dans ce cas pas d'intérêt car il n'y aurait pas d'effet de l'asynchronisme à corriger. Dans ce cas on ne testera donc que le calage naïf.

Nous testerons donc trois méthodes de calage pour le cas des relevés asynchrones (nommées respectivement dans la suite "asynchr naïf", "asynchr 30classes" et "asynchr 4classes") et une seule pour les relevés synchrones (nommé "synchrone calé"). De plus, nous présenterons également les performances de l'estimateur sans calage nommé "sans calage" (qui est donc le même que les relevés soient synchrones ou asynchrones).

Afin d'évaluer la précision des estimations, nous utiliserons le coefficient de variation défini comme suit : soit $\hat{Y}_r(t)$ l'estimateur du total à l'instant t (après calage selon la méthode choisie), pour l'échantillon r (r allant de 1 à R). La vraie consommation à l'instant t de la population est connue et sera notée $Y(t)$. Cette mesure intégrera à la fois le biais et la variance de l'estimateur.

Pour chaque méthode, on estimera à chaque instant le coefficient de variation

$$cv(t) = \frac{\sigma_{\hat{Y}_t}}{Y(t)}$$

avec $\sigma_{\hat{Y}_t} = [\frac{1}{R-1}\sum_{r=1}^{R}(\hat{Y}_r(t) - Y(t))^2]^{1/2}$

Ces quantités seront calculées instant par instant, puis, afin de produire un indicateur unique permettant de comparer les méthodes, on propose d'utiliser la moyenne de ce coefficient sur la

 ISSN: 2102-6238

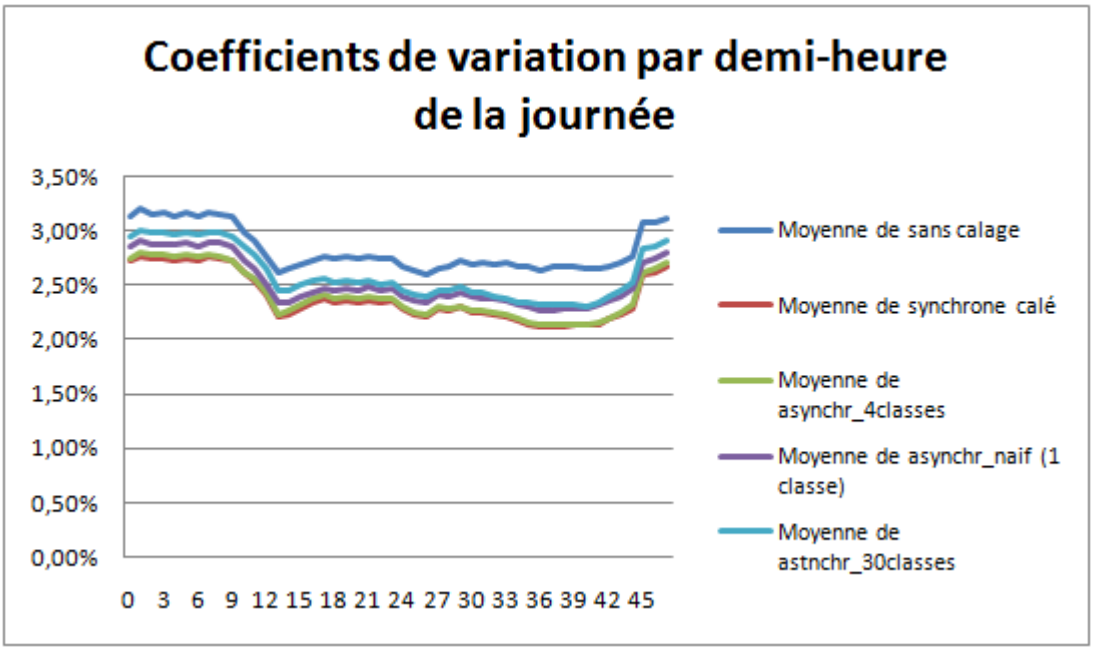

FIGURE 1. *Coefficients de variation moyens des estimateurs par demi-heure obtenus sans calage (en bleu), avec calage pour des relevés synchrones (en rouge), avec un calage en quatre classes pour des relevés asynchrones (en vert)avec un calage en une classe pour des relevés asynchrones (en violet) et avec un calage en trente classes pour des relevés asynchrones (en cyan)*

moyenne sur l'ensemble des instants de la période :

$$cv_{global} = \frac{1}{T}\sum_{t=1}^{T} cv(t)$$

On obtient alors les résultats suivants : dans tous les cas, le calage permet donc une amélioration sensible de la précision de l'estimateur. Ainsi, on passe de 2.82% sans calage à 2.39% lorsque les relevés sont synchrones (cas "idéal" où l'information collectée sur toute la population est la même). La méthode de calage en quatre classes permet d'obtenir des résultats très proches du cas synchrone (2.41%). Cette méthode est meilleure que la méthode naïve consistant à ignorer l'asynchronisme des données (2.53%).

En revanche, les performances de la méthode en trente classes, avec un coefficient de variation de 2.59%, sont très décevantes, car moins bonnes que celles la méthode naïve consistant à ignorer l'asynchronisme. On peut penser que le découpage en un trop grand nombre de classes induit des effectifs trop faibles pour chaque classe, donc une forte distorsion des poids et des biais importants. Cette hypothèse est étayée par l'observation des g-poids (rapports entre les poids calés et les poids de sondage) qui sont nettement plus dispersés pour le calage en trente classes que pour les autres méthodes : ainsi, comme le montre la Figure 2, le coefficient de variation de ces g-poids est en moyenne de 1.3% pour le calage sur relevés synchrones, de 1.9% que pour le calage "naif" , de 24% pour le calage en quatre classes, et de 50% pour le calage en trente classes. Cette dernière méthode semble donc réservée au cas où les volumes de données disponibles sont

 ISSN: 2102-6238

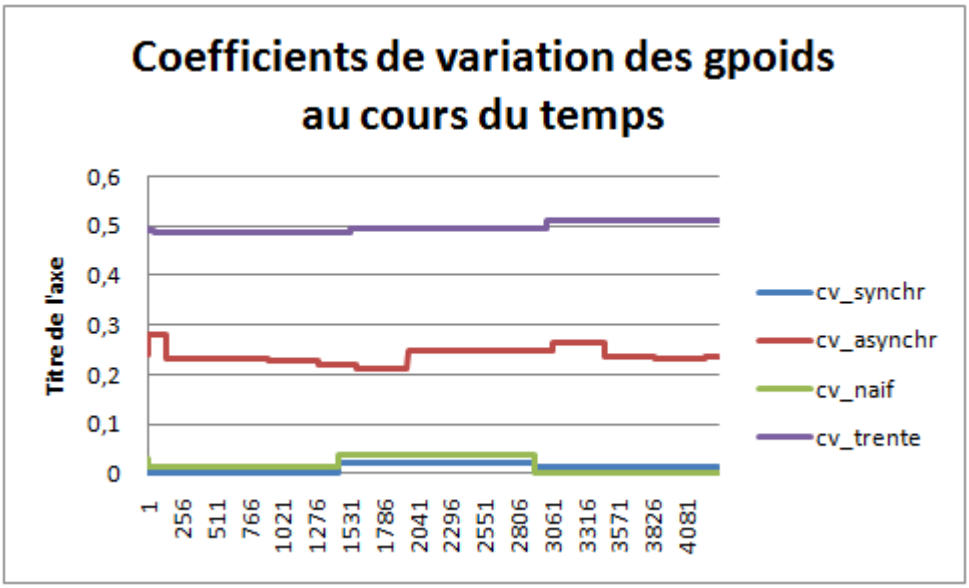

FIGURE 2. *Coefficients de variation des g-poids au cours du temps pour le calage sur relevés synchrones (en bleu, confondu avec le vert), le calage naif sur relevés asynchrones en une classe (en vert), le calage asynchrone en quatre classes (en rouge) et le calage asynchrone en trente classes (en violet)*

extrêmement importants.

Même si les résultats des différentes méthodes de calage sont très proches les uns des autres, la domination de la méthode de calage en quatre groupes sur la méthode naïve et la méthode en trente classes est confirmée par le fait que cette méthode est la meilleure pour 67% des instants (en moyenne sur les échantillons), contre 3% pour la méthode en trente classes et 26% pour la méthode naïve. De même, elle est la meilleure pour 45% des échantillons (en moyenne sur les instants) contre 27% pour la méthode en trente classes et 28% pour la méthode naïve.

Nous avons également détaillé ces résultats par demi-heure de la journée : on constate que l'ordre de classement des méthodes est le même pour chaque instant.

4. Discussion

Il faut en outre garder à l'esprit que les résultats obtenus ici concernent des clients industriels, ils ne peuvent donc pas être extrapolés aux clients résidentiels dont la consommation est plus chahutée et plus sensible au climat et donc à des décalages de périodes de relèves. En particulier, les résultats des méthodes de calage sont ici très proches pour les différentes méthodes et on pourrait s'attendre à des écarts plus importants sur des données résidentielles.

5. Conclusions

Dans cet article, nous avons proposé une adaptation simple de la méthode de calage classique pour le cas où l'information auxiliaire disponible, en l'occurence des consommations électriques issues de relevés d'index, ne correspond pas aux mêmes périodes pour l'ensemble des individus.

 ISSN: 2102-6238

Le principe de base de la méthode est de découper la population en groupes de clients relevés approximativement aux même dates de facon à éviter que l'information contenue dans la consommation déduite des relevés ne soit "parasitée" par des périodes de relèves différentes puis de caler séparément sur chacun de ces groupes. Dans le cas de très gros échantillons, on peut définir trente groupes de dates de relevés afin d' obtenir une information rigoureusement synchrone dans chaque groupe. Lorsque la taille des échantillons est plus réduite, cette méthode présente un risque d'instabilité des poids et donc de forte erreur quadratique moyenne sur des échantillons de taille plus réduite, et il est alors préférable d'utiliser un nombre réduit de groupes de clients dont les dates de relevés sont seulement "à peu près synchrones" (quatre classes dans l'exemple présenté ici). Notre méthode s'intègre aisément dans le cadre habituel du calage sur marges et il est donc aisé de calculer la variance de l'estimateur du total après repondération.

Ces techniques ont été testées sur des données réelles, comparées entre elles et avec le cas où on dispose de relevés synchrones. Le calage asynchrone en quatre classes, meilleur que la méthode naïve, permet de limiter très fortement la perte de précision due au fait de disposer de relevés asynchrones au lieu de relevés synchrones. Sur notre jeu de données de quelques dizaines de milliers de clients, la méthode en trente classes fournit des résultats décevants, mais il est très possible que sa précision s'améliore pour des plus grands échantillons comme ceux qui seront collectés par les futurs compteurs communicants.

Le caractère dynamique de notre problématique (à la fois dans la quantité estimée, l'information auxiliaire, mais aussi potentiellement les échantillons) peut créer des discontinuités dans les poids des individus et donc potentiellement dans la courbe estimée. Afin de limiter l'impact de ces discontinuités, il pourrait être intéressant d'étudier des méthodes de lissage de poids, pouvant s'appliquer au cours des transitions les plus importantes. Cela serait pertinent en particulier pour la phase de montée en puissance progressive de l'échantillon, dont la taille augmentera par vagues successives, les panélistes des différentes vagues pouvant être très différents les uns des autres. De premiers éléments de réponse sur ce sujet ont déjà été évoquée dans Dessertaine (2010) .

Remerciements

L'auteur remercie l'éditeur associé ainsi que les deux rapporteurs anonymes dont les remarques ont permis d'améliorer grandement cet article.

Références

Ardilly, P. (2006). *Techniques de sondages*. Technip.

Dalenius, T. and Hodges, J. (1959). Minimum variance stratification. *Journal of the Americal Statistical Association*, 54 :88–101.

Dessertaine, A. (2010). Sondages et courbes de charge electriques :introduction a la notion de calage dynamique. In *Colloque francophone sur les sondages*.

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

Horvitz, D. and Thompson, M. (1952). A generalization of sampling without replacement from a finite universe. *J. Atrmr. Statist. Assoc.*, 47 :663– 685.

 ISSN: 2102-6238