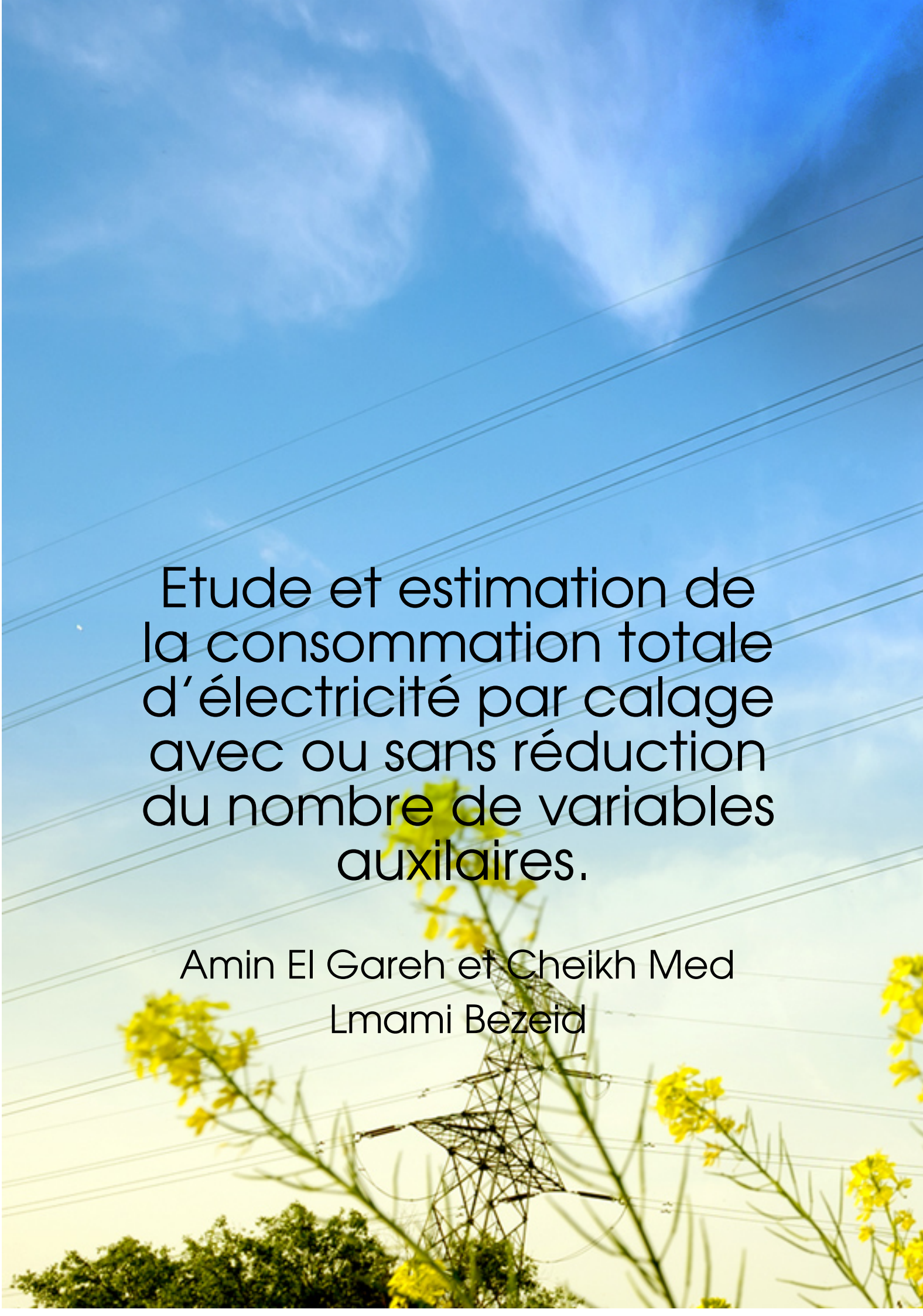


Etude et estimation de la consommation totale d'électricité par calage avec ou sans réduction du nombre de variables auxilaires.

Amin El Gareh et Cheikh Med Lmami Bezeid

Table des Matières

1. Présentation des données d'électricité

1.1 Introduction des données

On s'intéresse à des données d'électricité irlandaise de très grandes dimensions. Il s'agit de la consommation d'électricité enregistrée toutes les 30 minutes pendant 2 semaines (du lundi 5 octobre 2009 à 00:00 au dimanche 18 octobre 2009 à 23h30) pour 6291 individus: résidentiels, petites & moyennes entreprises, et autres. Ces données sont des données fonctionnelles car elles sont constituées d'un certain nombre de valeurs discrètes qui ont été mesurées, enregistrées par le CER (Commission for Energy Regulation), mais dont l'ensemble reflète une variation régulière. Comme en théorie, on pourrait obtenir une grande quantité de points, aussi rapprochés que l'on veut, on voit que l'on obtient une courbe et que l'on peut traiter la donnée comme une fonction.

1.2 Etude descriptive des données

Notre objectif ici est d'isoler le ou les comportements de consommation d'un ensemble d'individus appartenant à une même classe: résidentiels, petites & moyennes entreprises, ou autres.

1.2.1 Consommation moyenne

On considère la consommation moyenne comme étant la consommation totale prise en moyenne sur toute la population et par jour de la semaine. Sur la figure 1.1, on a représenté cette consommation moyenne en fonction du temps en minutes, et les jours de la semaine y sont délimités par des traits verticals rouges. L'analyse de la courbe des individus "résidentiels" et "autres", nous révèle le caractère cyclique de la consommation moyenne sur une période de 24h. La courbe des individus "petites & moyennes entreprises" indique une consommation moyenne qui s'apparente être cyclique entre le lundi et le vendredi, mais qui ne l'est pas le week-end.

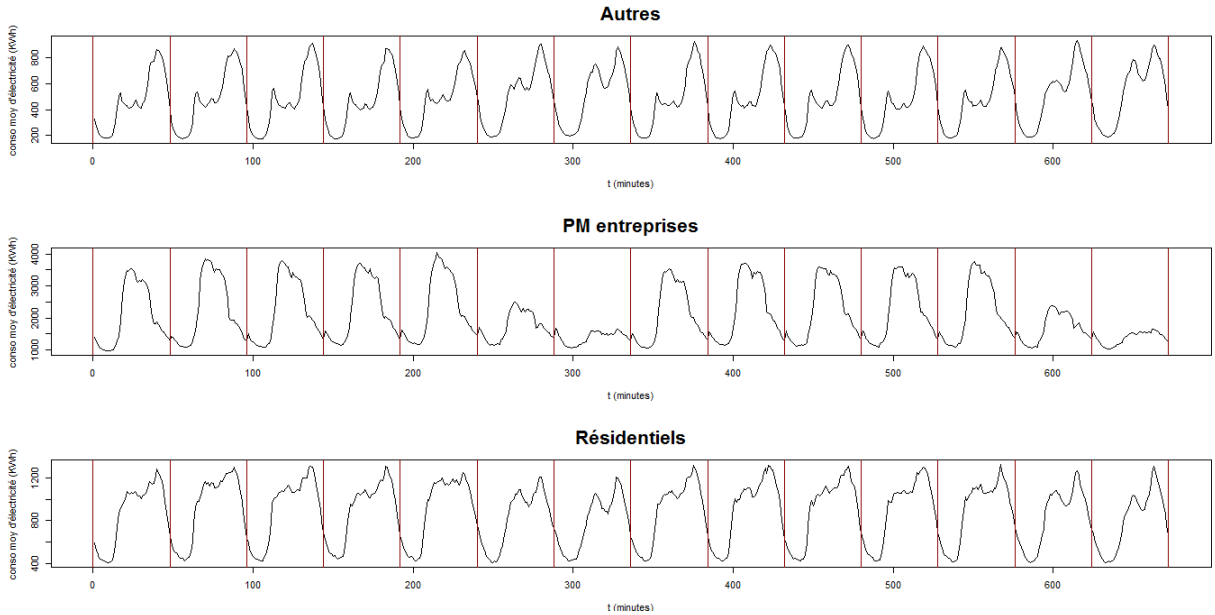


Figure 1.1 - Courbes de la consommation moyenne en fonction du temps (en minutes)

1.2.2 Mesure de la dispersion

L'illustration 1.2 met en relation la variance de la consommation des individus avec la classe d'appartenance. On peut distinguer un écart important entre les variances des petites & moyennes entreprises et les autres classes.

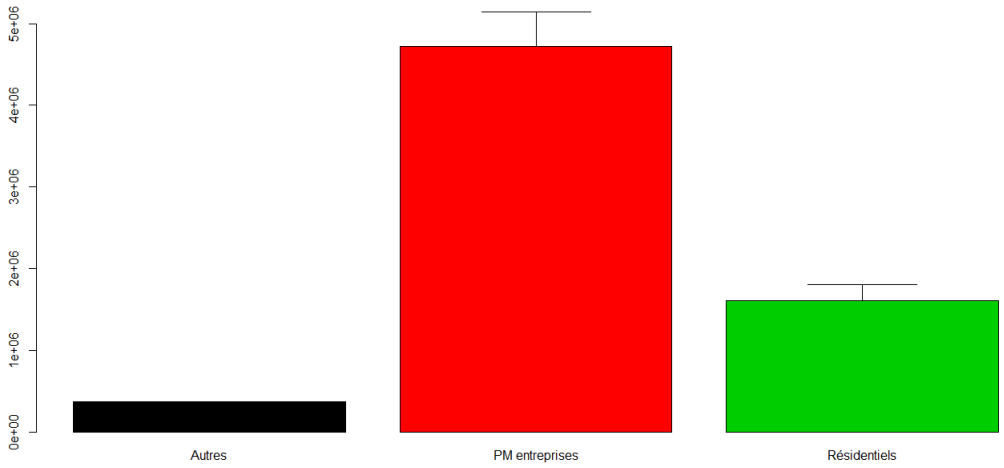


Figure 1.2 - Diagrammes en colonnes : variance de la consommation des individus par classes

Etudier la répartition journalière des boîtes à moustaches (n°1: lundi 5 octobre, ... , n°14: dimanche 18 octobre) permet d'abord de confirmer l'intuition que nous avions concernant la consommation moyenne journalière d'électricité qui s'avère être effectivement régulière. De plus, la figure 1.3 indique la présence d'un nombre important d'individus atypiques, en particulier "résidentiel", et qui par conséquent peuvent être problématique car ils peuvent biaiser les résultats, notamment pour la variance intra-classe.

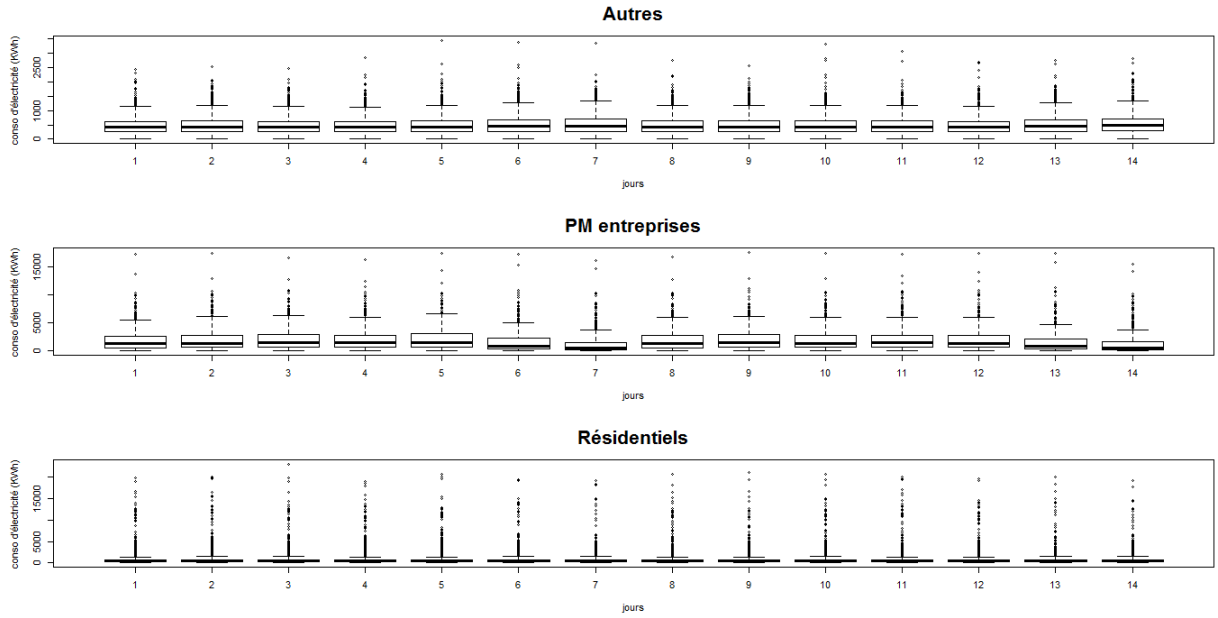


Figure 1.3 - Boîtes à moustaches de la consommation des individus par jour

1.2.3 Analyse en composantes principales

La première étape consiste à sélectionner le nombre d'axes factoriels. En utilisant le critère de Kaiser, nous sélectionnons les 3 premières valeurs propres qui expliquent 72.72% de l'inertie totale du nuage de points. Néanmoins, comme le troisième axe n'est corrélé significativement qu'avec une seule variable, nous ne le considérons pas dans l'interprétation.

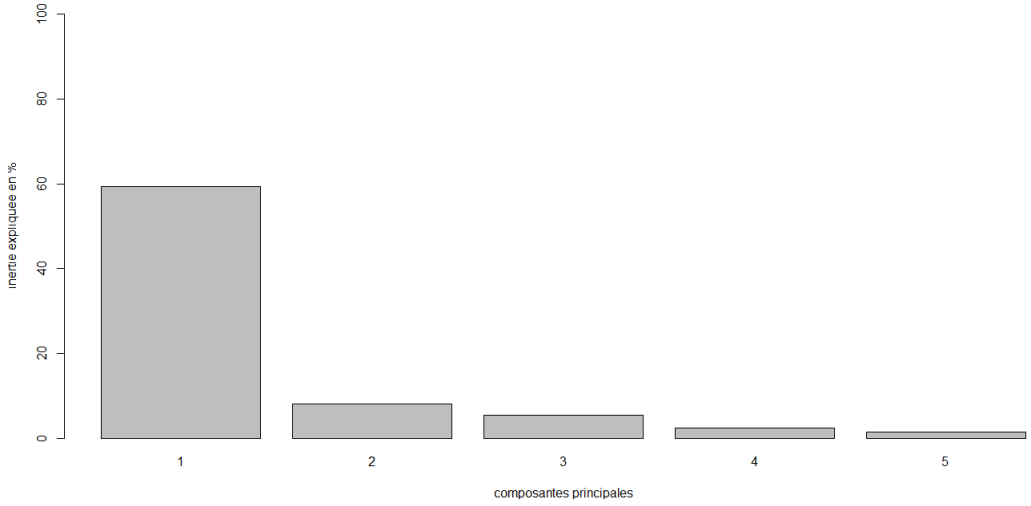


Figure 1.4 - Part de l'inertie

Le cercle des corrélations pour le plan formé des deux premiers axes factoriels est représenté figure 1.5 et 1.6. Toutes les variables sont bien représentées dans ce plan factoriel puisqu'elles sont proches du bord du cerle de corrélation. Sur la figure 1.5 on a choisi de prendre les variables représentatives d'un jour entre le lundi et le samedi, soit le vendredi 9 octobre 2009, et pour une meilleure visibilité nous avons divisé la représentation des variables en deux parties: matin-midi et midi-soir. Tandis que sur la figure 1.6 nous avons pris les variables représentatives du dimanche 11 octobre 2009.

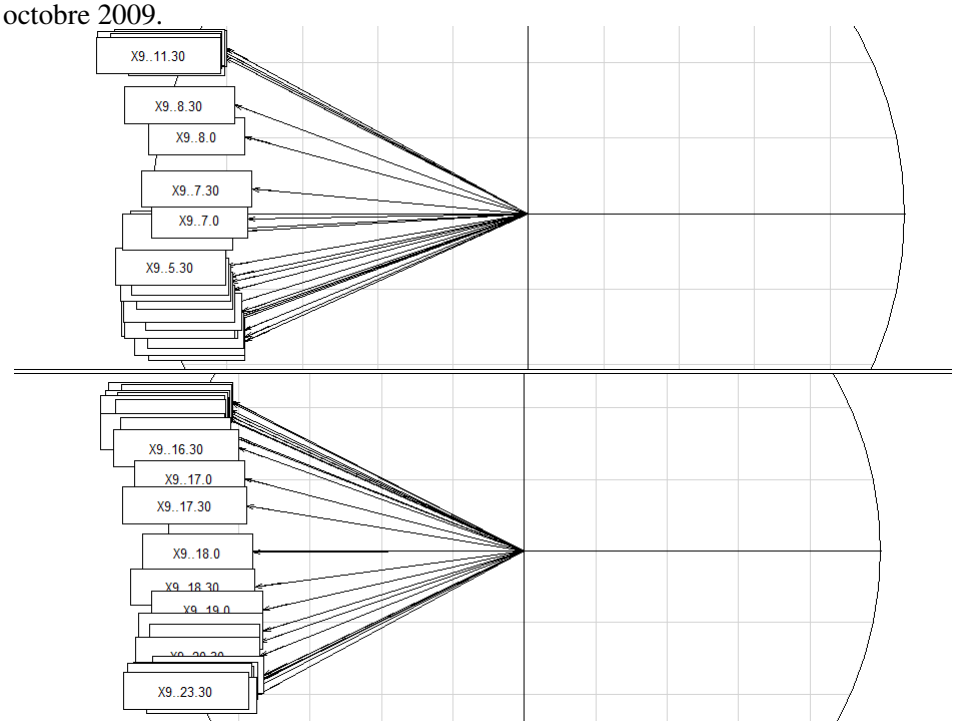


Figure 1.5 - Cercle des corrélations, représentation des variables du vendredi 9 octobre 2009

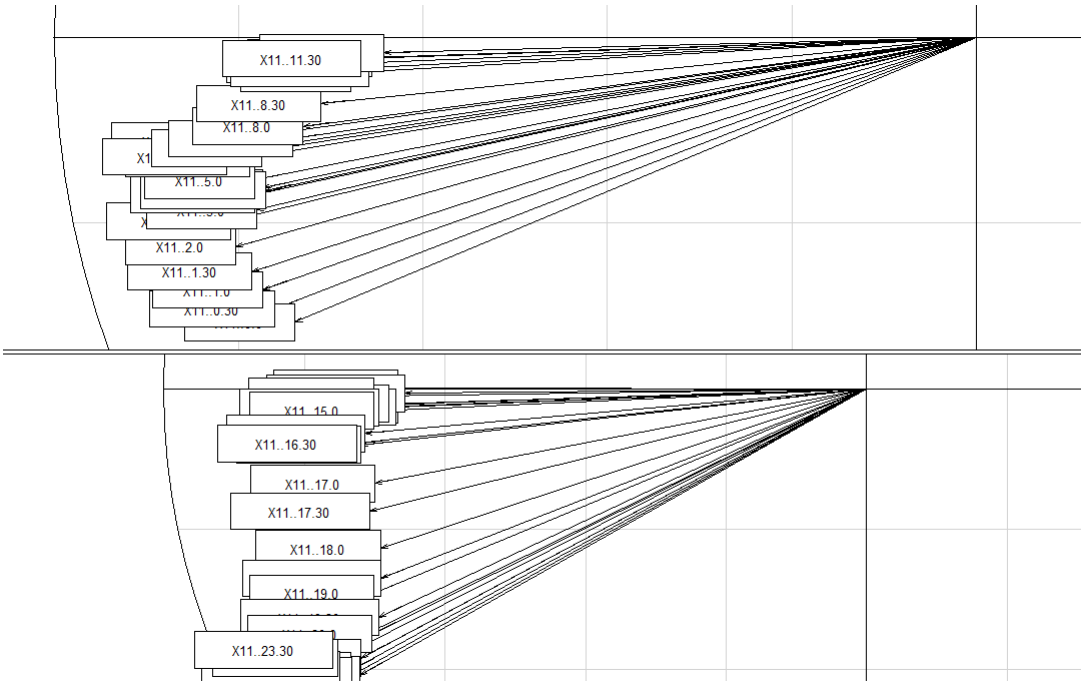


Figure 1.6 - Cercle des corrélations, représentation des variables pour le dimanche 11 octobre 2009

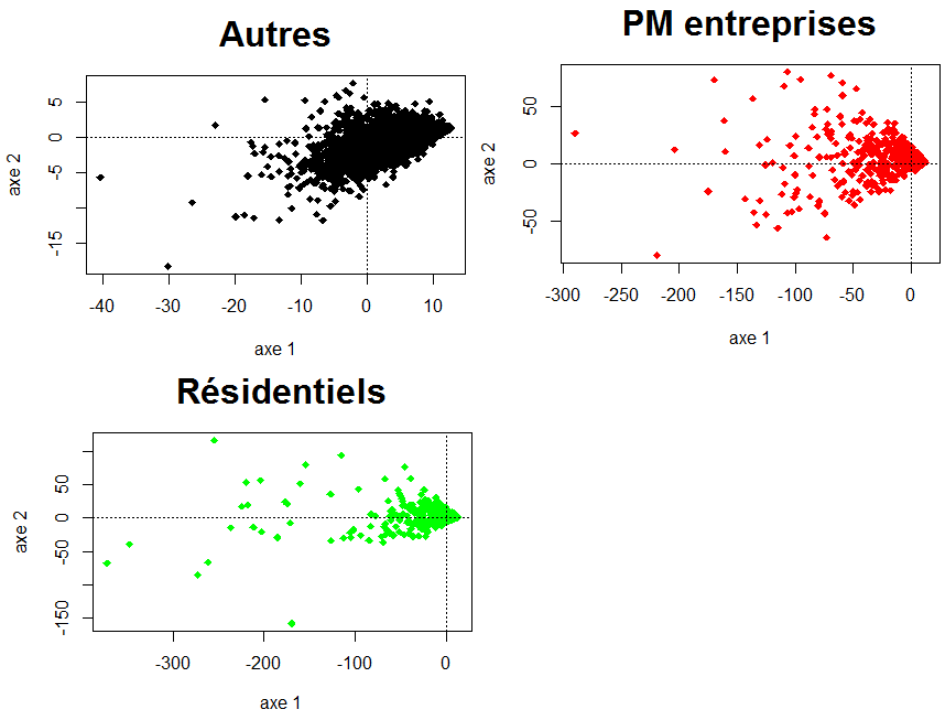


Figure 1.7 - Représentation des individus dans le plan factoriel selon la classe d'appartenance

On constate que les variables entre 8h et 18h sont bien représentées par rapport au premier axe tandis que le reste des variables est moins bien représenté. On peut donc dire que ce dernier explique la consommation d'électricité.

En utilisant une approche simultanée entre les graphes des individus et des variables, on s'aperçoit que tous les individus sont bien représentés par rapport au premier axe. Par conséquent, ces mêmes individus ont des valeurs élevées pour les variables entre 8h et 18h.

Pour le cas particulier du dimanche, on constate que les variables sont toutes négativement corrélés au deuxième axe. Cet axe explique la consommation, qui s'avère être faible.

2. Présentation du calage

En théorie des sondages, nous sommes souvent confrontés à l'estimation du total d'une variable d'intérêt Y

$$t_y = \sum_{k \in U} y_k$$

2.1 Estimation par Horvitz-Thompson

Horvitz et Thompson (1952) ont présenté un estimateur linéaire sans biais d'un total t_y valable pour tout plan de sondage

$$\hat{t_y} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

Cet estimateur est appelé le π-estimateur ou estimateur de Horvitz-Thompson. En effet, les valeurs prises par le caractère y sur les unités de l'échantillon sont dilatées par l'inverse des probabilités d'inclusion.

Theorem 2.1.1 Si $\pi_k > 0$, $k \in$ U, alors $\hat{t_y}$ estime t_y sans biais.

Demonstration.

$$\begin{aligned} \mathbb{E}[\hat{ty}] &= \mathbb{E}\left[\sum_{k \in s} \frac{y_k}{\pi_k}\right] \\ &= \mathbb{E}\left[\sum_{k \in U} \frac{I_k y_k}{\pi_k}\right] \end{aligned}$$

où I_k est une variable indicatrice qui vaut 1 si $k \in s$ et 0 sinon

$$= \sum_{k\in U} \frac{\mathbb{E}[I_k]y_k}{\pi_k}$$
$$= \sum_{k\in U} yk$$
$$= t_y$$

Par conséquent, $Biais(\hat{ty}) = \mathbb{E}[\hat{ty}] - ty = 0$ ■

2.2 Estimation par calage

Les méthodes de calage ont été formalisées par Deville et Särndal (1992). ils donnent un cadre général pour l'estimation par calage et étudient les propriétés de ces estimateurs. L'information auxiliaire utilisée est un vecteur de totaux connus t_x.

Definition 2.2.1 — Estimateur de calage. La méthode de calage consiste à chercher de nouveaux coefficients de pondération à affecter aux variables. L'estimateur par calage du total est un estimateur linéaire homogène qui s'écrit:

$$\hat{t}_{w,y} = \sum_{k\in s} w_{ks} y_k$$

Proposition 2.2.1 Les poids w_{ks} dépendent de l'échantillon aléatoire et sont déterminés de sorte que l'estimateur soit calé sur les totaux des caractères auxiliaires.

$$\hat{t}_{wx} = t_x \quad \text{où} \quad \hat{t}_{wx} = \sum_{k\in s} w_{ks} x_k$$

Notation 2.1. *Par commodité on pose $w_k = w_{ks}$, pour tout $k \in s$*

R Comme il existe une infinité de poids w_k qui satisfait la proposition 2.2.1, on va chercher des poids proches des poids π_k^{-1} du π-estimateur. Notons:

$$d_k = \frac{1}{\pi_k}, \quad k \in s$$

Theorem 2.2.2 Si $\pi_k > 0$, $k \in$ U, alors $\hat{t_{wy}}$ estime asymptotiquement sans biais t_y

Demonstration.

Le choix des poids wk proches de d_k assure la convergence asymptotique de l'estimateur de calage $\hat{t_{wy}}$ vers l'estimateur de Horvitz-Thompson $\hat{t_y}$

$\sum_{k\in s} w_k x_k \to \sum_{k\in s} d_k x_k \Leftrightarrow \hat{t_{wy}} \to \hat{t_y}$

Comme $\hat{t_{wy}}$ est uniformement bornée car c'est une somme finie alors, on a:

$\hat{t_{wy}} \to \hat{t_y} \Rightarrow \mathbb{E}[\hat{t_{wy}}] \to \mathbb{E}[\hat{t_y}]$

Par conséquent, $Biais(\hat{t_{wy}}) \to \mathbb{E}[\hat{t_y}] - t_y = 0$ ■

Pour garantir le critère de proximité entre w_k et d_k, on va utiliser une pseudo-distance de Khi-deux, $F_k(w_k,d_k) = \frac{(wk-dk)^2}{dk}$.
La fonction $F_k(w_k,d_k)$ est supposée positive, dérivable par rapport à w_k et strictement convexe, telle que $F_k(w_k,d_k) = 0$. Les poids w_k, $k \in s$, sont obtenus en résolvant le problème (P) de minimisation

$$(P) := \begin{cases} \sum_{k\in s} F_k(w_k,d_k) = \sum_{k\in s} \dfrac{(w_k-d_k)^2}{d_k} \leftarrow \textbf{min} \\ \\ \hat{t}_{wx_j} = t_{x_j} \qquad\qquad j = 1,...,p \end{cases}$$

On peut écrire la fonction Lagrangienne :

$$\begin{aligned} L(w_k,\lambda_j) &= \sum_{k\in s} \frac{(w_k-d_k)^2}{d_k} - \sum_{j=1}^{p} \lambda_j(\hat{t}_{wx_j} - t_{x_j}) \\ &= \sum_{k\in s} \frac{(w_k-d_k)^2}{d_k} - \sum_{j=1}^{p} \lambda_j \left(\sum_{k\in s} w_k x_{k_j} - \sum_{k\in U} x_{k_j} \right) \end{aligned}$$

où les λ_j sont les multiplicateurs de Lagrange. En annulant les dérivées partielles du Lagrangien par rapport aux w_k, on obtient:

$$\frac{\delta L(w_k,\lambda_j)}{\delta w_k} = 0 \quad \Leftrightarrow \; 2\left(\frac{w_k-d_k}{d_k}\right) - \sum_{j=1}^{p} \lambda_j x_{k_j} = 0$$

$$\Leftrightarrow \; w_k = d_k \left(\frac{1}{2} \sum_{j=1}^{p} \lambda_j x_{k_j} + 1 \right) \tag{2.1}$$

En considérant l'expression des poids w_k et d'après les contraintes de (P), on obtient:

$$\begin{aligned} t_{x_j} = \sum_{k\in s} w_k x_{k_j} \;\Leftrightarrow\; t_{x_j} &= \sum_{k\in s} d_k x_{k_j} \left(\frac{1}{2} \sum_{j=1}^{p} \lambda_j x_{k_j} + 1 \right) \\ \Leftrightarrow\; t_{x_j} &= \sum_{k\in s} d_k x_{k_j} + \frac{1}{2} \sum_{k\in s} d_k x_{k_j} \sum_{j=1}^{p} \lambda_j x_{k_j} \end{aligned}$$

L'expression peut également s'écrire sous forme matricielle

$$t_x \;=\; \hat{t_x} + \frac{1}{2} \sum_{k\in s} d_k x_k^t x_k \lambda$$

Sous réserve que la matrice $\sum_{k\in s} d_k x_k^t x_k$ soit inversible, on peut déterminer λ

$$\lambda \;=\; 2\left(\sum_{k\in s} d_k x_k^t x_k \right)^{-1} (t_x - \hat{t_x})$$

En remplaçant λ dans l'expression (2.1) on obtient les poids w_k

$$w_k = d_k\left(\frac{1}{2}{}^t\lambda x_k + 1\right) = d_k\left({}^t(t_x - \hat{t_x})\left(\sum_{k\in s} d_k x_k^t x_k\right)^{-1} x_k + 1\right) \tag{2.2}$$

Enfin, l'estimateur de calage $\hat{t_{wy}} = \sum_{k\in s} w_k y_k$ peut s'écrire

$$\hat{t_{wy}} = \hat{t_y} + {}^t(t_x - \hat{t_x})\left(\sum_{k\in s} d_k x_k^t x_k\right)^{-1} \sum_{k\in s} x_k d_k y_k \tag{2.3}$$

2.3 Application du calage

On souhaite estimer la consommation totale d'électricité de la deuxième semaine, $t_y = 1485176258$. Il y'a 336 variables auxiliaires, qui représentent les consommations prises toutes les 30 minutes pendant la première semaine. Et sont présentées dans les vecteurs $(t_{x_j})_{j=1,\ldots,336} = \sum_{k\in U} x_{k_j}$.

2.3.1 Calage selon un plan SAS

Soit U la population d'étude composée de N=6291 individus qui consomment de l'électricité sur deux semaines consécutives. La première approche consiste à sélectionner un échantillon s $\subset$ U selon un plan de sondage p(.) aléatoire simple sans remise (SAS) de taille n=600.

Definition 2.3.1 Un plan de taille fixe n est dit simple sans remise (SAS) si et seulement si

$$p(s) := \begin{cases} \binom{N}{n}^{-1} & \text{si card(s) = n} \\ 0 & \text{sinon} \end{cases}$$

Definition 2.3.2 Les probabilités d'inclusion d'ordre un peuvent être déduites du plan de sondage p(.) SAS

$$\pi_k = \sum_{k\in s} p(s) = \sum_{k\in s} \binom{N}{n}^{-1} = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}, \text{ pour tout k} \in \text{ U}$$

R Les probabilités d'inclusion associées à nos données sur l'électricité sont $\pi_k = \frac{600}{6291}$.

On a simulé I=500 échantillons aléatoires simples sans remise de taille n=600, en utilisant la méthode de calage présentée au paragraphe 2.2 on obtient une estimation moyenne de $\bar{\hat{t}}_{wy} = \frac{1}{I}\sum_{i=1}^{I} \hat{t}_{wy}^{(i)} = 1492219885$.

Pour comparer l'estimateur de Horvitz-Thompson avec celui par le calage on utilise le rapport des variances R qui suit:

$$R = \frac{Var(\hat{t_{wy}})}{Var(\hat{t_y})} \approx \frac{\sum_{i=1}^{I}(\hat{t_{wy}}^{(i)} - t_y)^2}{\sum_{i=1}^{I}(\hat{t_y}^{(i)} - t_y)^2} = \frac{1.747882e+17}{3.690269e+18} = 0.047365$$

La même précision est obtenue en utilisant le calage avec un plan SAS de taille n = 600*0.047365 = 29 qu'en prennant la méthode de Horvitz-Thompson avec un plan SAS de taille n = 600.

Le coefficient de variation c_v est une mesure de dispersion relative, qui va être utilisé comme contrôle de qualité pour l'estimateur de calage.

$$c_v = \frac{\sqrt{Var(\hat{t_{wy}})}}{\bar{\hat{t}}_{wy}} = 0.011618 \quad \text{qui peut aussi s'exprimer en pourcentage, soit cv = } 1.1618\%$$

2.3.2 Calage selon un plan Stratifié avec SAS

Notre population d'étude U est divisée dans $H = 3$ strates avec N_h unités dans la strate h.
Les tailles des strates sont connues et valent $N_1 = 4225$ pour la strat "Autres", $N_2 = 485$ pour la strat "P&M entreprises", et $N_3 = 1581$ pour la strat "Résidents". Et vérifient la condition suivantes: $N_1 + N_2 + N_3 = N = 6291$

La deuxième approche pour sélectionner notre échantillon s sera d'appliquer le plan SAS dans chacune des trois strates.

Pour trouver le nombre d'unités échantillonnées n_h dans chaque strate on utilise une allocation proportionnelle: $\frac{n_h}{N_h} = \frac{n}{N}$.

L'estimateur de Horvitz-Thompson selon le plan Stratifié pour le total t_y est donné par:

$$\hat{t_{str}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k.$$

Et l'estimateur de calage selon le plan Stratifié est de:

$$\hat{t_{wstr}} = \sum_{h=1}^{H} w_{k,s_h} \sum_{k \in s_h} y_k.$$

On a simulé $I = 500$ échantillons tirés selon le plan Stratifié avec SAS chacun de taille $n = 600$. On obtient une estimation moyenne de $\hat{\bar{t}}_{wstr} = 1504942728$.

Pour comparer l'estimateur de Horvitz-Thompson avec celui du calage pour ce plan on calcule

$$R = \frac{Var(\hat{t}_{wstr})}{Var(\hat{t}_{str})} \approx \frac{\sum_{i=1}^{I}(\hat{t}_{wstr}^{(i)} - t_y)^2}{\sum_{i=1}^{I}(\hat{t}_{str}^{(i)} - t_y)^2} = 0.231561$$

La même précision est obtenue en utilisant le calage par Stratification avec SAS de taille n = 600* 0.231561 = 139 qu'en prennant la méthode de Horvitz-Thompson avec un plan SAS de taille n = 600.

On calcule le coefficient de variation c_v, qui vaut:

$$c_v = \frac{\sqrt{Var(\hat{t}_{wstr})}}{\hat{\bar{t}}_{wstr}} = 0.02203803 \quad \text{qui peut aussi s'exprimer en pourcentage, soit cv = } 2.203803\%$$

R Le calage qui a été effectué suivant l'un ou l'autre des plans utilise une pseudo-distance de khi-deux. Cette pseudo-distance est à l'origine de l'obtention de poids négatifs.
Pour y remédier on peut utiliser d'autres méthodes comme logistique, raking ratio, linéaire tronquée, ...

3. Présentation du calage sur CP

Le calage en présence de beaucoup de variables auxilaires peut s'avérer instable surtout s'il y'a des collinéarités entre les variables auxiliaires. Une méthode pour palier cet incovénient est de réduire l'information auxiliaire en réalisant une ACP.

3.1 Méthode de calage sur CP

3.1.1 Information auxiliaire sur CP

Soient $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ les valeurs propres de matrice de variance-covariance $\frac{1}{N}X^TX$, où X est la matrice formée des vecteurs $(X_j)_{1,...,p}$ contenant l'information auxiliaire de départ. Et soient $v_1,...,v_p$ les vecteurs propres associés aux $(\lambda_j)_{j=1,...,p}$.

Les composantes principales $Z_1,...,Z_p$ qui sont les combinaisons linéaires des $(X_j)_{1,...,p}$ sont définies par la relation $Z_j = Xv_j, \qquad j = 1,...,p.$

On va sélectionner selon le critère de Kaiser les r premières composantes principales, qui vont formées les nouvelles variables auxiliaires. La matrice d'information auxiliaire est maintenant donnée par:

$$Z = (Z_1,...,Z_r) = (z_k)_{k\in U} \qquad \text{où} \qquad z_k = (z_{k1},...,z_{kr})^T$$

3.1.2 Calage sur CP

Notre objectif est de trouver les poids de calage $(w_k)_{k\in s}$ qui vérifient:

$$(P') := \begin{cases} \sum_{k\in s} F_k(w_k,d_k) = \sum_{k\in s} \dfrac{(w_k-d_k)^2}{d_k} \leftarrow \textbf{min} \\ \\ \hat{t}_{wz_j} = t_{z_j} \qquad\qquad j = 1,...,r \end{cases}$$

R Les équations de calage qui peuvent aussi s'écrire sous la forme $\sum_{k\in s} w_k z_{k_j} = \sum_{k\in s} z_{k_j}$ où $\hat{t}_{wz_j} = \sum_{k\in s} w_k z_{k_j}$ est un estimateur dépendant du paramètre r

- Si $r = 0$ on obtient l'estimateur de Horvitz-Thompson $\hat{t}_y$ qui n'utilise pas l'information auxiliaire.
- Si $r = p$ on obtient l'estimateur calé sur les p variables de départ.

La résolution du problème de minimisation (P') est la même que celle présentée au paragraphe 2.2, les poids w_k sont donc:

$$w_k = d_k\left(\frac{1}{2}{}^t\lambda z_k + 1\right) = d_k\left({}^t(t_z - \hat{t}_z)\left(\sum_{k\in s} d_k z_k^t z_k\right)^{-1} z_k + 1\right)$$

Et l'estimateur de calage $\hat{t_{wy}} = \sum_{k\in s} w_k y_k$ peut s'écrire

$$\hat{t_{wy}} = \hat{t}_y + {}^t(t_z - \hat{t}_z)\left(\sum_{k\in s} d_k z_k^t z_k\right)^{-1}\sum_{k\in s} z_k d_k y_k$$

3.2 Application du calage sur CP

On souhaite toujours estimer la consommation totale d'électricité de la deuxième semaine, $t_y = 1485176258$. La matrice des variables auxiliaires de départ X est de dimension 6291 x 336. On rappelle que la nouvelle matrice des variables auxiliaires $Z = (Z_1, ..., Z_r)$, dépend d'un paramètre r, que l'on peut faire varier entre 1 et $p = 336$. Le choix du nombre r de variables auxiliaires ou autrement dit composantes principales vont permettre d'étudier la qualité de l'estimation par calage.

On a simulé I=500 échantillons aléatoires simples sans remise de taille n=600, en utilisant la méthode de calage sur les r=2 composantes principales on obtient une estimation moyenne de t_y qui est $\bar{\hat{t}}_{wy} = \frac{1}{I}\sum_{i=1}^{I}\hat{t}_{wy}^{(i)} = 1480138029$.

Le rapport des variances R vaut:

$$R = \frac{Var(\hat{t_{wy}})}{Var(\hat{t}_y)} \approx \frac{\sum_{i=1}^{I}(\hat{t_{wy}}^{(i)} - t_y)^2}{\sum_{i=1}^{I}(\hat{t}_y^{(i)} - t_y)^2} = \frac{1.088333e+17}{3.678668e+18} = 0.029585$$

La même précision est obtenue en utilisant le calage sur CP avec un plan SAS de taille n = 600*0.029585 = 18 qu'en prennant la méthode de Horvitz-Thompson avec un plan SAS de taille n = 600.

Le coefficient de variation vaut: $cv = \frac{\sqrt{Var(\hat{t_{wy}})}}{\bar{\hat{t}}_{wy}} = 0.009377838$ qui peut aussi s'exprimer en pourcentage, soit cv = 0.9377838%

On va montrer que l'estimateur du calage sur les deux premières CP est dans notre cas le meilleur estimateur. Pour se faire on a calculer l'estimateur moyen $\bar{\hat{t}}_{wy}$, le rapport R et le coefficient c_v pour les CP compris entre 2 et 336.

D'après la figure 1.8, la valeur de l'estimateur moyen par calage diminue en fonction du nombre de CP retenus. Le trait horizontal rouge représente la valeur exact de la consommation total d'électricité.

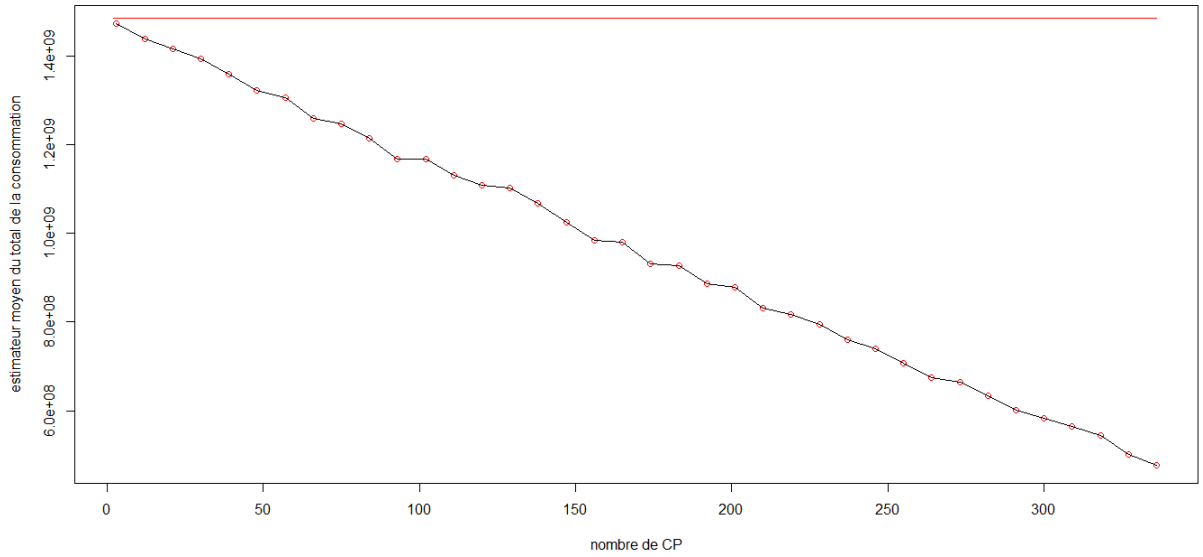


Figure 1.8 - Estimateur moyen $\hat{\bar{t}}_{wy}$ du total de la consommation en fonction du nombre de CP

D'après la figure 1.9, le rapport des variances R évolue progressivement jusqu'à un nombre de CP égal à 200. Une fois ce nombre de CP dépassé le coefficient R devient élevé et instable.

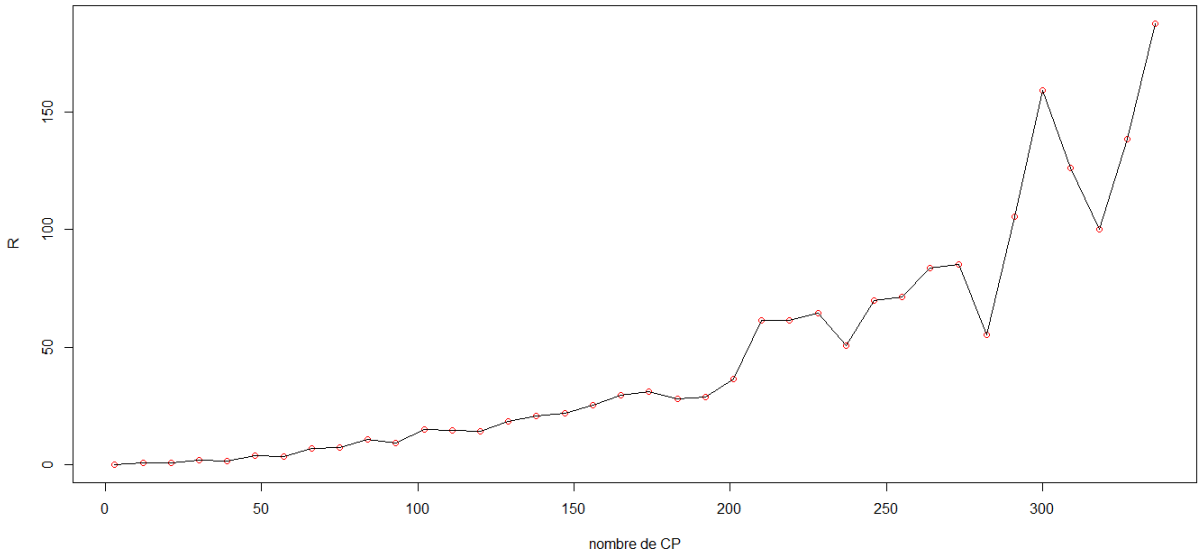


Figure 1.9 - Le rapport R en fonction du nombre de CP

De la figure 1.10 on retiendra que le choix des 2 premières composantes principales implique une faible variation $c_v \approx 1\%$, alors que pour un nombre de CP au déla de $r = 50$ le coefficient cv devient relativement important (supérieur à 5%).

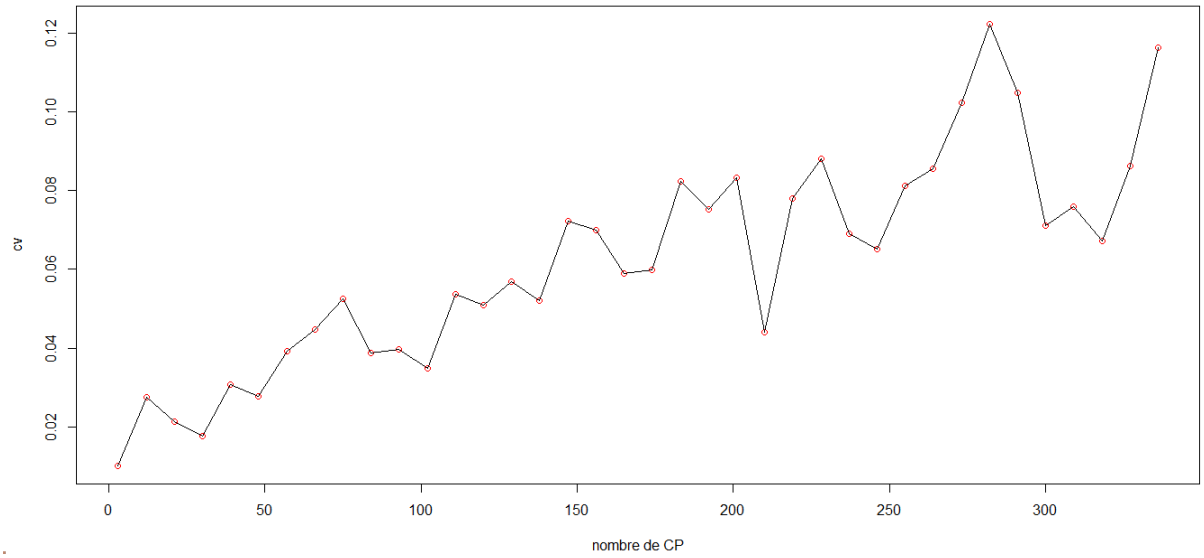


Figure 1.10 - Le coefficient de variation c_v en fonction du nombre de CP

Interprétation de la méthode du calage sur les CP

• **Avantages**

- Rapide lorsque le calage est fait sur un nombre très faible de CP, car la taille de la matrice $\left(\sum_{k\in s} d_k z_k^t z_k\right)$ à inverser est petite.
- Facile à mettre en oeuvre, elle permet d'obtenir de gains importants en termes de variance.
- La précision R de l'estimateur de calage est meilleure par rapport aux deux autres plans, à condition que l'on retienne un nombre faible de CP.
- Le problème de positivité des poids est résolus, ceci dû au fait d'une variance répartit sur les CP.

• **Inconvénient**

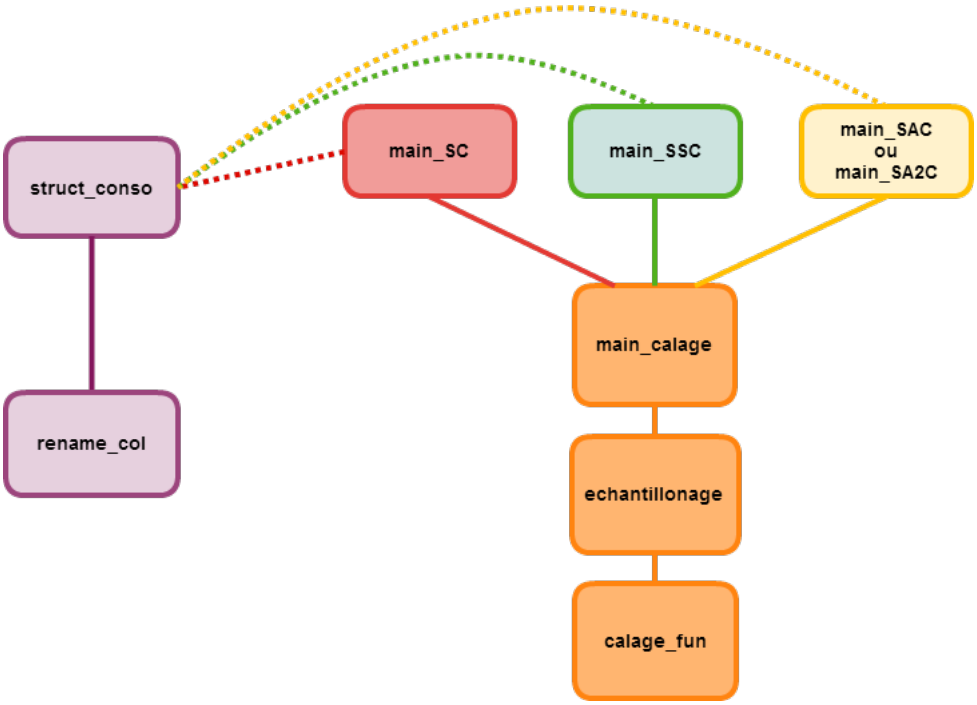
- Instable pour un grand nombre de CP.

4. Présentation des programmes R

Nous avons réalisés l'étude de la méthode de calage selon plusieurs plans:

- SC = Calage selon un plan SAS (voir 2.3.1)
- SSC = Calage selon un plan Stratifié avec SAS (voir 2.3.2)
- SAC = Calage sur Composantes principales avec SAS (voir 3.2).
- SA2C = Calage sur les 2 premières Composantes principales avec SAS (le meilleur plan).

4.1 Organigramme



Pour chacun des plans, on a crée un "main" ou autrement dit un programme principal, dont les dépendances ont été généralisées. On distingue d'un côté une dépendance conditionnelle marquée par des pointillés, c'est-à-dire que struct_conso s'exécute que si la table conso n'existe

pas dans le workspace.
De l'autre côté on effectue un calage étape par étape à partir de main_calage, qui se résume de la facon suivante:

1. échantillonnage selon un plan p(.) SAS avec la fonction echantillonage.
2. calcul des poids de calage avec la fonction calage_fun.
3. calcul des estimateurs de calage et de Horvitz-Thompson.

4.2 Bibliothèque des programmes R

```
# struct_conso.R

# struct_conso permet de structurer la table "smart.278co.csv" avec "contrat.smartco.csv"
#
#=> en sortie :
# ~ conso : table des conso. des ind. classés par contrat
# ~ nconso : table uniquement avec les conso. des ind.

struct_conso=function(){

  source('rename_col.R')

  smart=read.table('smart.278co.csv', header = TRUE, sep=",")
  smart[,1]=1:nrow(smart)

  contrat=read.table('contrat.smartco.csv', header = TRUE, sep=",")

  # renomme la table contrat, 1:Autres, 2:PM entreprises, 3:Résidentiels
  levels=c("Autres","PM entreprises","Résidentiels")
    for( i in 1:3){ contrat[which(contrat[,2]==i),2]=levels[i]; }

  tab.conso=merge(smart,contrat, by="X") # fusion de la table conso. des ind. avec contrat
  tab.conso[,1]=tab.conso[,ncol(tab.conso)]
  tab.conso=tab.conso[,-ncol(tab.conso)]

  # renomme les colonnes de la table conso, nouveau format: JOURS/HEURE
  conso=rename_col(tab.conso,p=(ncol(tab.conso)-1))
  nconso=conso[,-1] # table conso sans var. qualitatives

list(conso=conso,nconso=nconso)
}
```

```
# rename_col.R

# rename_col creer un nouveau format "JOURS/ HEURE"
# pour le nom des variables de la table conso

rename_col=function(conso,p){

v=rep(0,p)
h=0; m=0; j=5
v[1]="5/ 0:0"

for( i in 1:(p-1) ){

  if(i%%48!=0){
      m=m+30
      if(m==60){ m=0; h=h+1;}
      }
   else{ j=j+1; h=0; m=0;}
  v[i+1]=paste(c(as.character(j),"/ ",as.character(h),":",as.character(m)),sep=" ",
  collapse="")

  }

  v=c("code",v);  colnames(conso)=v;

  return(conso)
}
```

```
# main_calage.R

# main_calage permet de calculer les estimateurs de calage et de horvitz-thompson
#
#=> en entrée :
# ~ tab : la matrice contenant toutes les données
# ~ X_U : la matrice contenant l'information auxiliaire dans U
# ~ N : la taille de la population U   # ~ n : la taille de la pop. S
# ~ dk : l'inverse de la probabilité d'inclusion pour un plan p(.) SAS
# ~ sem1.var : les indices des var. de la semaine 1
# ~ sem2.var : les indices des var. de la semaine 2
# ~ nom_plan : SC = SAS+Calage | SSC = Strat+SAS+Calage | SAC = SAS+ACP+Calage

main_calage=function(tab,X_U,N,n,dk,sem1.var,sem2.var,nom_plan){

echt = echantillonage(tab,X_U,N,n,sem1.var,sem2.var,nom_plan) # echantillonage
X_S = echt$X_S; Y_S=echt$Y_S

yk=apply(Y_S,1,sum) # conso. de l'individu k dans S pdt la sem2

calage.conso = calage_fun(dk,X_S,X_U) # fonction de calage

est.ty = calage.conso$wk %*% yk # estimateur de calage
est.hty = sum(dk*yk)  # estimateur de hovitz-thompson

list( est.ty=est.ty, est.hty=est.hty)
}
```

```
# echantillonage.R

# echantillonage renvoit les matrices X_S et Y_S prises aléatoirement dans U
#
#=> en entrée :
# ~ tab : la matrice contenant toutes les données
# ~ X_U : la matrice contenant l'information auxiliaire dans U
# ~ N : la taille de la population U   # ~ n : la taille de la pop. S
# ~ sem1.var : les indices des var. de la semaine 1
# ~ sem2.var : les indices des var. de la semaine 2
# ~ nom_plan : SC = SAS+Calage | SSC = Strat+SAS+Calage | SAC = SAS+ACP+Calage

echantillonage=function(tab,X_U,N,n,sem1.var,sem2.var,nom_plan){

  rand.ind=sample(1:N, n, replace=F) # indice des individus appartenant a S pris alea.

  if(nom_plan=="SC" | nom_plan=="SSC" ){ # plan SC ou SSC
    X_S=as.matrix(tab[rand.ind,sem1.var])  # conso. des ind. de S pour la sem1
  }
  else{ # plan SAC
    X_S=as.matrix(X_U[rand.ind,])  }  # conso. des ind. de S pour la sem1

  Y_S=as.matrix(tab[rand.ind,sem2.var])  # conso. des ind. de S pour la sem2

  list( X_S=X_S, Y_S=Y_S)
}

# calage_fun.R

# calage_fun permet de calculer les poids de calage
#
#=> en entrée :
# ~ dk : l'inverse de la probabilité d'inclusion pour un plan p(.) SAS
# ~ X_S : la matrice contenant l'information auxiliaire dans S
# ~ X_U : la matrice contenant l'information auxiliaire dans U

calage_fun=function(dk,X_S,X_U){

   tx = apply(X_U,2,sum) # vecteur contenant les totaux pour l'info. aux. dans U
   twdx = apply(X_S,2,sum) # vecteur contenant les totaux pour l'info. aux. dans S
   txtwdx = tx - dk*twdx
   mat = ginv(dk * crossprod(X_S,X_S))
   wk = dk + dk*(txtwdx %*% mat) %*% t(X_S) # poids de calage
   list(wk=wk)

  }
```

```
# main_SC.R

#####################################################################

#######               CALAGE AVEC UN PLAN SAS             ########

#####################################################################

setwd("F:/Master 2 MIGS/projet/")  # chemin du dossier courant

# si la table conso n'existe pas dans le workspace, alors
# on creer la table conso et nconso (sans var. qualitative)
if( exists("conso")=="FALSE" ){
  source('struct_conso.R')
  struct_conso=struct_conso()
  conso=struct_conso$conso; nconso=struct_conso$nconso
}

source('echantillonage.R'); source('calage_fun.R'); source('main_calage.R')
library(MASS)

N=nrow(nconso) # U
n=600 # S
P=ncol(nconso) # nombre de variables
p=336 # nombre de variables auxiliaires
sem1.var=1:p # indice des var de la sem1
sem2.var=(p+1):P # indice des var de la sem2
X_U=as.matrix(nconso[,sem1.var]) # conso des ind de U pour la sem1
Y_U=as.matrix(nconso[,sem2.var])  # conso des ind de U pour la sem2
yk=apply(Y_U,1,sum) # conso de l'individu k dans U pdt la sem2
ty=sum(yk) # conso total sem2
dk=N/n # l'inverse de la probabilité d'inclusion pour un plan p(.) SAS

ni=50 # nombre d'echantillons I de taille n=600
i=seq(1,ni,1)
lest=sapply( i,function(x) main_calage(nconso,X_U,N,n,dk,sem1.var,sem2.var,"SC") )

est.ty=unlist(lest[1,]); est.hty=unlist(lest[2,])

tty = rep( ty, ni )
R = sum( (est.ty - tty)^2 ) / sum( (est.hty - tty)^2 )
cv = sqrt( var(est.ty) ) / mean(est.ty)

cat( sprintf( "\n L'estimateur moyen de la conso totale de la semaine 2 est de %d \n
              La qualite R de l'estimateur est de %f \n
              Le coefficient de variation est de %f \n" , round(mean(est.ty)), R, cv) )
```

```
# main_SSC.R

#########################################################################
#
# #######          CALAGE SELON UN PLAN STRAT AVEC SAS          ########
#
# #########################################################################

setwd("F:/Master 2 MIGS/projet/") # chemin du dossier courant

# si la table conso n'existe pas dans le workspace, alors
# on creer la table conso et nconso (sans var. qualitative) avec la fonction struct_conso
if( exists("conso")=="FALSE" ){
  source('struct_conso.R')
  struct_conso=struct_conso()
  conso=struct_conso$conso; nconso=struct_conso$nconso
  n=struct_conso$n; p=struct_conso$p
}

source('echantillonage.R'); source('main_calage.R'); source('calage_fun.R')
library(MASS)

N=nrow(nconso) # U
n=600; # S
P=ncol(nconso) # nombre total de variables
p=336 # nombre de variables auxiliaires
sem1.var=1:p # indice des var de la sem1
sem2.var=(p+1):P # indice des var de la sem2

class.contrat=conso[,1]
fac=as.factor(class.contrat)
levels=c("Autres","PM entreprises","Résidentiels")

sem2.var=(p+1):P # indice des var de la sem2
Y_U=as.matrix(nconso[,sem2.var])
yk=apply(Y_U,1,sum) # conso de l'individu k dans U pdt la sem2
ty=sum(yk) # conso total sem2

lest.ty=list(); lest.hty=list()

  for(j in 1:10){
    twy=list(); htwy=list(); k=1

      for(i in levels){
        aux = factor( (fac%in%i)*1 )
        Nh = length( which(aux==1) ) # Uh
        nh = round(n*Nh/N) # Sh
```

```
        dk = Nh/nh # l'inverse de la proba. d'inclusion p(.) SAS pour la k-ieme strat

        mat = as.matrix(nconso[aux%in%1,]) # matrice de données pour la k-ieme strat
        X_U = mat[,sem1.var]; Y_U=mat[,sem2.var]

        lest = main_calage(mat,X_U,Nh,nh,dk,sem1.var,sem2.var,"SSC")

        twy[[k]] = lest$est.ty # estimation par calage pour la k-ieme strat
        htwy[[k]] = lest$est.hty # estimation par horvitz-thompson pour la k-ieme strat
        k=k+1
      }

    lest.ty[[j]] = sum( unlist(twy) ) # j-ieme estimation par calage
    lest.hty[[j]] = sum( unlist(htwy) ) # j-ieme estimation par horvitz-thompson
  }

est.ty=unlist(lest.ty); est.hty=unlist(lest.hty)
tty = rep( ty, length(est.ty) )
R = sum( (est.ty - tty)^2 ) / sum( (est.hty - tty)^2 )

cat( sprintf( "\n L'estimateur moyen de la conso totale de la semaine 2 est de %d \n
              La qualité R de l'estimateur est de %f" , round(mean(est.ty)), R) )



# main_SA2C.R

######################################################################

#######       CALAGE SUR LES 2 CP AVEC UN PLAN SAS        ########

######################################################################

setwd("F:/Master 2 MIGS/projet/") # chemin du dossier courant

# si la table conso n'existe pas dans le workspace, alors
# on creer la table conso et nconso (sans var. qualitatives) avec la fonction struct_conso
if( exists("conso")=="FALSE" ){
  source('struct_conso.R')
  struct_conso=struct_conso()
  conso=struct_conso$conso; nconso=struct_conso$nconso
}

source('echantillonage.R'); source('calage_fun.R'); source('main_calage.R')
# install.packages("ade4")
library(ade4)
library(MASS)

N=nrow(nconso) # U
n=600 # S
```

```
P=ncol(nconso) # nombre de variable
p=336 # nombre de variable auxiliaire de d?part
sem1.var=1:p # indice des var de la sem1
sem2.var=(p+1):P # indice des var de la sem2
X_U=as.matrix(nconso[,sem1.var]) # conso des ind de U pour la sem1
Y_U=as.matrix(nconso[,sem2.var])  # conso des ind de U pour la sem2
yk=apply(Y_U,1,sum) # conso de l'individu k dans U pdt la sem2
ty=sum(yk) # conso total sem2
dk=N/n # l'inverse de la probabilite d'inclusion pour un plan p(.) SAS


# ACP centree reduite sur les variables auxiliaires
#--------------------------------------------------
conso.acp=dudi.pca(X_U,center=TRUE,scale=TRUE,scannf=FALSE,nf=p)

# Estimation optimale par calage sur les CP 1 et 2
#--------------------------------------------------
X_U=-conso.acp$li[,1:2] # matrice des var. auxiliaires sur les CP
sem1.var=1:2  # indice des var de la sem1
ni=50 # nombre d'echantillons I de taille n=600
i=seq(1,ni,1)
lest=sapply( i,function(x) main_calage(nconso,X_U,N,n,dk,sem1.var,sem2.var,"SAC") )
est.ty=unlist(lest[1,]) # ni estimation par calage
est.hty=unlist(lest[2,]) # ni estimation par horvitz-thompson

tty = rep( ty, ni )
R = sum( (est.ty - tty)^2 ) / sum( (est.hty - tty)^2 )
cv = sqrt( var(est.ty) ) / mean(est.ty)

cat( sprintf( "\n L'estimateur moyen de la conso totale de la semaine 2 est de %d \n
              La qualite R de l'estimateur est de %f \n
              Le coefficient de variation est de %f \n" , round(mean(est.ty)), R, cv) )
```

```
# main_SAC.R

##############################################################################

#######      CALAGE SUR CP ENTRE 2 et 336 AVEC UN PLAN SAS          ########

##############################################################################

setwd("F:/Master 2 MIGS/projet/") # chemin du dossier courant

# si la table conso n'existe pas dans le workspace, alors
# on creer la table conso et nconso (sans var. qualitative) avec la fonction struct_conso
if( exists("conso")=="FALSE" ){
  source('struct_conso.R')
  struct_conso=struct_conso()
  conso=struct_conso$conso; nconso=struct_conso$nconso
}

source('echantillonage.R')
source('calage_fun.R')
source('main_calage.R')

# install.packages("ade4")
library(ade4)
library(MASS)

N=nrow(nconso) # U
n=600 # S
P=ncol(nconso) # nombre de variable
p=336 # nombre de variable auxiliaire de depart
sem1.var=1:p # indice des var de la sem1
sem2.var=(p+1):P # indice des var de la sem2
X_U=as.matrix(nconso[,sem1.var]) # conso des ind de U pour la sem1
Y_U=as.matrix(nconso[,sem2.var])  # conso des ind de U pour la sem2
yk=apply(Y_U,1,sum) # conso de l'individu k dans U pdt la sem2
ty=sum(yk) # conso total sem2
dk=N/n # l'inverse de la probabilite d'inclusion pour un plan p(.) SAS


# ACP centree reduite sur les variables auxiliaires
#--------------------------------------------------
conso.acp=dudi.pca(X_U,center=TRUE,scale=TRUE,scannf=FALSE,nf=p)

#
lnb.cp=seq(2,p,by=10) # le nombre de CP a garder, compris entre 3 et 336
lest=list()
k=1

  for( pj in lnb.cp ){
```

```
    X_U=-conso.acp$li[,1:pj] # matrice des var. auxiliaires sur les CP
    sem1.var=1:pj  # indice des var de la sem1
    ni=10 # nombre d'echantillons I de taille n=600
    i=seq(1,ni,1)
    lest[[k]]=sapply( i,function(x) main_calage(nconso,X_U,N,n,dk,sem1.var,sem2.var,"SAC")
    k=k+1
    }

i=seq(1,(k-1),1)
# ni estimation par calage pour la ieme CP
est.ty=sapply( i,function(x) unlist(lest[[x]][1,]) )
# ni estimation par HT pour la ieme CP
est.hty=sapply( i,function(x) unlist(lest[[x]][2,]) )
mest.ty=sapply( i,function(x) mean(unlist(lest[[x]][1,])) )

tty = rep( ty, ni )
R=sapply( i,function(x) sum((est.ty[,x] - tty)^2) / sum((est.hty[,x] - tty)^2) )
cv=sapply( i, function(x) sqrt( var(est.ty[,x]) ) / mest.ty[x] )


# representation de la qualite de l'estimateur en fonction du nombre de CP
#-------------------------------------------------------------------------

# affichage du coefficient R en fonction du nb de CP
plot(lnb.cp,R, type="p", col="red",xlab="nombre de CP")
lines(lnb.cp,R,col="black")

# affichage du coefficient de variation cv en fonction du nb de CP
plot(lnb.cp,cv, type="p", col="red",xlab="nombre de CP",ylab="cv")
lines(lnb.cp,cv,col="black")

# affichage de l'estimateur moyen en fonction du nb de CP
plot(lnb.cp,mest.ty, type="p", col="red",xlab="nombre de CP",
ylab="estimateur moyen du total de la consommation")
lines(lnb.cp,mest.ty,col="black")
lines(2:p,rep(ty,(p-1)),col="red")
```

5. Conclusion

Ce travail visait à développer plusieurs plans permettant d'estimer la consommation totale d'électricité. Introduire des méthodes issues de la théorie des sondages, nous a permis de comprendre les principes et les enjeux de cette discipline. La théorie des sondages suscite des travaux d'approfondissement récents et innovants en tout point. Du fait qu'elle porte sur des populations finies, d'existence bien concrète, elle ne peut ignorer les contraintes du monde réel. C'est dans ce sens que nous portons un intérêt pour son étude et pour son application.

Bibliographie

1. Tillé, Y. (2001), *Théorie des sondages*, chez Dunod.
2. http://sondages2012.ensai.fr/wp-content/uploads/2011/01/expose_rennes_gogaShehzad_20121.pdf