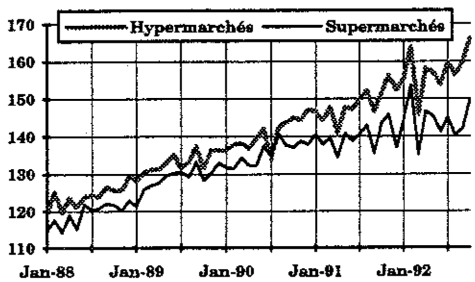
UN CADRE DE RÉFÉRENCE POUR CORRIGER LES EFFETS DE COURT TERME : L'EXEMPLE DES CJO

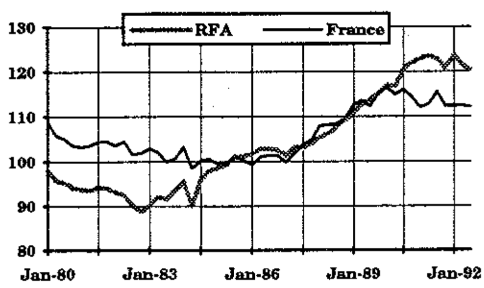
Vincent MAILLARD,
Département de la conjoncture

Beaucoup dé séries économiques présentent de fortes irrégularités à très court terme. Ces aléas se retrouvent la plupart du temps avec le même signe et la même intensité sur d'autres séries statistiques voisines. Ainsi, la comparaison des séries de chiffre d'affaires des supermarchés et des hypermarchés révèle exactement les mêmes mouvements au mois le mois alors que les évolutions de moyen-long terme sont divergentes. De même, l'examen des indices *trimestriels* de production industrielle en biens manufacturés de la France et de la RFA montre des fluctuations parallèles à court terme. En l'absence de liens directs de cause à effet entre les séries étudiées, ces exemples suggèrent l'existence d'une explication commune aux aléas de court terme qui n'ait pas d'incidence à long terme. En règle générale, les effets de structure journalière du mois -ou effets de "CJO"- constituent de bons candidats pour expliquer et estimer ces aléas.

Dans le cas des hypermarchés et des supermarchés, le simple examen du calendrier permet en effet de comprendre les fortes fluctuations de début 1992. Le pic de février 1992 correspond ainsi à un mois de février de 29 jours qui comprenait 5 samedis, jours forts pour les ventes de ces secteurs. Une configuration aussi exceptionnelle ne se rencontre que tous les 30 ans!



Source : INSEE, CA3 en valeur



Source : INSEE, OCDE

Plus simplement, la présence d'un dimanche en plus ou en moins peut affecter très sensiblement la plupart des séries économiques (tout du moins les séries de *flux*) puisque le dimanche n'est pas ouvré et qu'il correspond à un point bas pour la plupart des activités. L'ampleur du phénomène peut être estimé naïvement à 4 % pour les séries mensuelles -resp. à plus de 1% pour les séries trimestrielles- ce qui correspond à l'impact d'un jour d'activité en plus ou en mois sur 25 jours ouvrables -resp. 75 jours-[1]. Même sur les données annuelles, l'effet de jours ouvrables ne peut être négligé. En effet, même si son impact est d'autant plus faible que la périodicité est longue, l'exigence de l'économiste sur la fiabilité et la pertinence des données est aussi d'autant plus forte.

Plus généralement, la structure hebdomadaire du mois c'est à dire le nombre de lundi, mardi etc. influe sur le niveau mensuel. Pour la consommation des ménages par exemple, c'est souvent le nombre de

[1] Cet impact dépend en réalité de la série concernée, il est en général inférieur à 4% sur les séries mensuelles. Sur quelques séries l'ordre de grandeur de 4% est cependant bien vérifié.

samedis, jour fort des achats, qui explique une grande partie des fluctuations des séries CVS. Il faut donc souvent accorder un poids différent à chacun des jours de la semaine pour effectuer une correction efficace. Pour des raisons pratiques, ces corrections seront aussi dénommées "CJO" même si elles sont **beaucoup plus générales** que ce que suggère une appellation qui semble désigner une simple règle de trois sur le nombre de jours ouvrables du mois.

Les aléas climatiques peuvent aussi perturber nombre de séries économiques. L'exemple de la consommation d'énergie domestique et notamment d'électricité -énergie non stockable- vient naturellement à l'esprit ; mais les aléas climatiques affectent aussi la consommation en textile-habillement, en produits alimentaires à cause des boissons. Certains secteurs productifs comme le bâtiment ont une activité sensible au nombre de gelées... La correction des aléas de température permettrait ainsi de mieux comprendre les évolutions à court-terme de nombreux indicateurs.

Première partie : la correction pour jours ouvrables

Le problème des CJO peut-être abordé dans deux optiques différentes :
- soit dans le but d'obtenir une bonne estimation de l'effet *global CVS et CJO* sans chercher à décomposer précisément entre ces deux termes.
- soit dans le but d'obtenir une bonne estimation de l'effet CJO seul

La deuxième optique ne va pas sans poser des problèmes conceptuels. Sur le principe il n'est pas en effet totalement évident de savoir si *les effets moyens de calendrier* -longueur des mois, nombre moyen de jours fériés du mois par exemple- doivent être intégrés dans la saisonnalité ou dans les effets de jours ouvrables. Le bon sens qui semble indiquer qu'il faut a priori plutôt ranger ces effets dans les CJO, est contredit par tous les praticiens des séries temporelles. En effet, toutes les séries CVS (non-CJO) intègrent des effets moyens de calendrier sans que cela n'ait jamais troublé personne. Notons simplement que les deux méthodes qui sont exposées ci-dessous divergent de ce point de vue : celle exposée en 1.a suppose implicitement que les effets de calendrier moyens[2] font partie de la saisonnalité alors que le choix inverse est retenu en 1.b.

Dans la méthode proposée, les effets moyens de jours ouvrables sont de facto intégrés dans la saisonnalité. Mais le but est ici d'obtenir un bon indicateur *global CVS et CJO.* La décomposition proposée entre partie saisonnière et partie CJO non saisonnière se justifie parce qu'elle est plus simple et permet d'obtenir un bon indicateur global.

[2] Plus exactement : les effets de calendrier de l'année de base (ex : le nombre de jours fériés de mai 1985 pour un indice base 1985) font partie de la saisonnalité de la série.

I.1 Les méthodes traditionnelles

I.1.a les méthodes non-économétriques

Je n'évoque que pour mémoire les méthodes traditionnelles de CJO qui partent en général d'enquêtes spécifiques sur le profil infra-hebdomadaire de production. Ces enquêtes attribuent des poids à chacun des jours de la semaine (poids sont la somme doit valoir 7, un poids spécifique est affecté aux jours fériés) et le coefficient de CJO d'un mois m est alors égal à la somme des poids des jours divisé par la somme des poids des jours du même mois de l'année de base. Ce coefficient correctif semble en général sur-correcteur et n'est pris en compte que partiellement, à 70% par exemple. Ce n'est qu'après que les séries sont éventuellement désaisonnalisées.

Ces méthodes on fait leurs preuves, elles sont néanmoins très coûteuses puisqu'elles nécessitent la réalisation d'enquêtes spécifiques sur le profil infra-hebdomadaire de la variable. Pratiquement, seules quelques séries de production sont ainsi traitées.

I.1.b Les méthodes économétriques traditionnelles

Il existe des méthodes d'estimation des CJO purement économétriques[3]. En général, elles estiment les coefficients journaliers en effectuant une régression directe de l'irrégulier de la série sur 7 régresseurs correspondant à chacun des jours de la semaine. L'article de DAGUM, QUENEVILLE, SUTRADHAR décrit le modèle le plus couramment employé (et même, à vrai dire, le seul que je connaisse à quelques variantes près). Il se résume ainsi :

1- Pour chaque mois t (on pourrait évidemment aussi prendre le trimestre) la variable de flux y_t peut se décomposer comme la somme des flux journaliers selon la formule :

$$y_t = \sum_{i=1}^{7} N_{it} \times \bar{y}_{it} \qquad (1)$$

où N_{it} désigne le nombre de Dimanches (i=1) Lundis (i=2) ... Samedis (i=7) et $\bar{y}_{it}$ désigne la valeur moyenne de la variable sur les différents jours du mois

2- Dans ces conditions, on montre facilement que :

$$y_t = \sum_{i=2}^{7} (N_{it} - N_{1t})(\bar{y}_{it} - \bar{y}_{0t}) + N_t \bar{y}_{0t} \qquad (2)$$

Avec: $\bar{y}_{0t} = \frac{1}{7}\sum_{i=1}^{7} \bar{y}_{it}$ (3)

Ce qu'on note: $y_t = D_t + N_t \bar{y}_{0t}$ (4)

N_t représente la moyenne du niveau espéré de la variable sur le mois

N_t peut se décomposer en $N_t \bar{y}_{0t} = T_t + S_t + U_t$ (5)

$T_t = Tendance,\ S_t = Saisonnalité,\ U_t = Irrégulier$

[3] L'option TDREG= de l'instruction MONTHLY de la PROC X11 utilise cette méthode. La procédure ne permet pas de corriger les séries trimestrielles.

Si $N_t\bar{y}_{0t}$ peut être une variable intégrée[4] ou non, il est supposé que D_t est une variable stationnaire au moins à l'ordre 1 (souvent même de moyenne nulle). D_t représente quant à lui l'effet de structure de jours ouvrables du mois.

Si on note $\delta_{it} = (\bar{y}_{it} - \bar{y}_{0t})$ et $x_{it} = (N_{it} - N_{1t})$ on obtient :

$$D_t = \sum_{i=2}^{7} x_{it}\delta_{it} = {}^t\underline{x_t}\,\underline{\delta_t} \qquad (6)$$

$\underline{x_t}$ est une variable "extensive" qui mesure l'écart à une norme de référence (en l'occurrence l'écart entre le nombre de lundis, mardis, etc. et le nombre de dimanches du mois). $\underline{\delta_t}$ est une variable "intensive" qui décrit l'impact de cet écart sur la variable totale, c'est le vecteur des différences entre la moyenne de chacun des jours de la semaine et la moyenne hebdomadaire divisée par 7. Par exemple, si la production moyenne sur un lundi est de 3, que la production moyenne d'une semaine est de 14 on aura $\underline{\delta_{2t}} = 3 - 14/7 = 1$ et, pour un mois avec 5 lundis et 4 dimanches on aura $\underline{x_{2t}} = 5 - 4 = 1$ (contribution du lundi). Pour obtenir D_t, il faut prendre en compte la somme de ces contributions pour l'ensemble des jours de la semaine *qui ne sont pas indépendantes,* on montre en effet que :

$$\sum_{i=1}^{7} \delta_{it} = 0 \quad (7)$$

I.1.c Les problèmes de cette méthode

L'estimation des $\underline{\delta_t}$ se fait généralement par des MCO, en régressant alors la série "détendancée" et éventuellement corrigée des variations saisonnières (c'est à dire, selon la terminologie habituelle l'irrégulier de désaisonnalisation) sur $\underline{x_t}$. En ce cas, $\underline{\delta_t}$ ne dépend pas de t. Quelques améliorations permettent d'utiliser des coefficients variables, c'est à dire que $\underline{\delta_t}$ dépend alors de t selon un modèle ARIMA d'ordre peu élevé, l'estimation utilise alors le filtrage de Kalman. Enfin, cette méthode est bien entendu utilisable avec des modèles multiplicatifs. Il faut néanmoins préciser deux hypothèses implicites très fortes de cette méthode.

- Il n'est jamais supposé que les δ_{it} peuvent être saisonniers : l'impact d'un lundi supplémentaire est par exemple supposé être le même quelque soit le mois considéré. Cette hypothèse peut sembler naturelle. Elle pourrait être contredite dans quelques cas : certains commerces ont le droit d'être ouverts les dimanches du seul mois de décembre et l'effet d'un dimanche supplémentaire peut en ce cas être différent en décembre.

- La méthode proposée effectue une correction selon trois critères : la tendance, la saisonnalité, **puis** les jours ouvrables. Or cette méthode n'est valable que si les critères de déflation sont orthogonaux entre eux[5]. En d'autres termes, cet algorithme n'est valable que si la saisonnalité[6] ne prend pas déjà en compte une partie de l'effet jours ouvrables. Il est vrai qu'en

[4] Une variable intégrée à l'ordre n est une variable temporelle dont la différence n-1 ème est instationnaire et dont les différences d'ordre supérieur à n sont stationnaires.

[5] Ce résultat est connu dans le cas paramétrique par le théorème de Frisch-Waugh mais il est clair que le même problème se pose ici même si la désaisonnalisation n'est pas une régression paramétrique.

[6] En toute rigueur, le même problème se pose pour le partage entre tendances et jours ouvrables, mais il paraît raisonnable de supposer que ces deux composantes sont orthogonales.

dehors du problème des longueurs inégales des mois - qu'on peut facilement traiter- les phénomènes de calendrier sont orthogonaux aux phénomènes saisonniers : il n'y a pas de mois qui aient en moyenne plus de samedis ou de dimanches par exemple. Malheureusement cette hypothèse n'est plus valable dès que l'on veut traiter de manière symétrique les phénomènes de jours fériés : tout français sait que certains mois ont structurellement plus de jours fériés[7]. Ce problème est à mon sens à l'origine de beaucoup des déboires des utilisateurs de l'option CJO de X11.

I.2 La méthode proposée.

I.2.a Aperçu théorique.

En reprenant les mêmes notations qu'au paragraphe précédent, on ajoute aux 7 jours du modèle deux types de jours particuliers : les jours de semaine fériés et les samedis fériés. Les dimanches fériés sont assimilés à des dimanches normaux. Le nombre N_t de jours de la période (mois ou trimestre) t se décompose ainsi en :

$$n_t = \sum_{i=1}^{9} n_{i,t} = n_{.,t} \quad (8)$$

Pour une date t, l'année est notée n et la période (mois ou trimestre) est notée p. t décrit le temps et peut varier sur toute la longueur de la série (t=1 à T). per désigne la fonction qui à t renvoie la période correspondante **per**(t)=p.

En s'inspirant du théorème de Frisch-Waugh, il faut modifier la méthode précédente. L'irrégulier de désaisonnalisation, par nature, ne prend pas en compte la structure moyenne en jours fériés de la période considérée : le fait qu'il y a structurellement plus de jours fériés en mai est incorporé dans le coefficient de CVS du mois et *non dans l'irrégulier*. Il ne faut donc pas régresser l'irrégulier de désaisonnalisation directement sur la structure du mois en jours fériés (i.e. les n_{it}) qui contient aussi la structure moyenne mais plutôt sur la différence entre la structure du mois et la structure moyenne du mois qui lui est orthogonale. Si **np**(t) est le nombre de fois que la période (mois ou trimestre) considérée apparaît dans la série (approximativement la longueur de la séries en années).:

$$\forall i \in \{1,\ldots,9\}$$

$$\bar{n}_{i,t} = \frac{1}{\mathrm{np}(t)} \sum_{\substack{\tau\in\{1,\ldots,T\} \\ \mathrm{per}(\tau)=\mathrm{per}(t)}} n_{i,\tau} \quad (9) \qquad d_{i,t} = n_{i,t} - \bar{n}_{i,t} \quad (10)$$

$$\text{On a généralement : } \sum_{i=1}^{9} d_{i,t} = 0 \quad (11)$$

$\bar{n}_{i,t}$ est le nombre moyen de jours de type i sur la période (nombre moyen de dimanches en février par exemple...). $d_{i,t}$ la différence à cette moyenne. La relation (11) est vraie pour presque toutes les périodes -sauf les mois de février et les premiers trimestres- puisque :

[7] En revanche, ce problème peut n'avoir que peu d'impact pour la désaisonnalisation des séries américaines ou le problème des jours fériés est moins crucial.

$$\sum_{i=1}^{9} n_{i,t} = \sum_{i=1}^{9} \overline{n}_{i,t} = n_t \quad (12)$$

Le modèle correct à coefficients constants consiste alors à régresser l'irrégulier de désaisonnalisation (ou le log de l'irrégulier multiplicatif) sur les variables $d_{i,t}$ selon le modèle :

$$I_t = \sum_{i=2}^{9} \delta_i d_{i,t} + \varepsilon_t \quad (14)$$

Ce modèle est estimé par une procédure de moindres carrés ordinaires. Les coefficients dominicaux ne sont pas estimés puisque le modèle est "presque" mal spécifié à cause de la relation (13)[8]. L'interprétation des coefficients calculés est d'autre part délicate et sera discutée dans des cas particuliers. En première analyse les δ_i s'interprètent comme ceux du paragraphe précédent, ils mesurent alors la différence entre le niveau moyen du jour considéré et le niveau moyen normal sur l'ensemble des jours. Plus il est élevé, plus le niveau de la variable étudiée est important sur le jour considéré, le niveau arbitraire de référence étant 0 pour les dimanches. Il faut néanmoins garder à l'esprit que les régresseurs sont naturellement corrélés : le nombre de lundis est plutôt corrélé au nombre de mardis et anticorrélé au nombre de jeudis. Dans ces conditions, il faut se méfier d'une interprétation trop rapide des coefficients estimés (ce qui ne doit pas remettre en cause la pertinence de la régression) et ceci d'autant plus que la régression s'appuie sur peu de points, c'est à dire que la série est courte.

I.2.b L'algorithme choisi

Comme d'une part il est préférable d'opérer une désaisonnalisation sur des séries préalablement corrigées des jours ouvrables, et que d'autre part il est difficile d'estimer les coefficients de CJO avant d'avoir désaisonnalisé, une procédure itérative a été retenue.

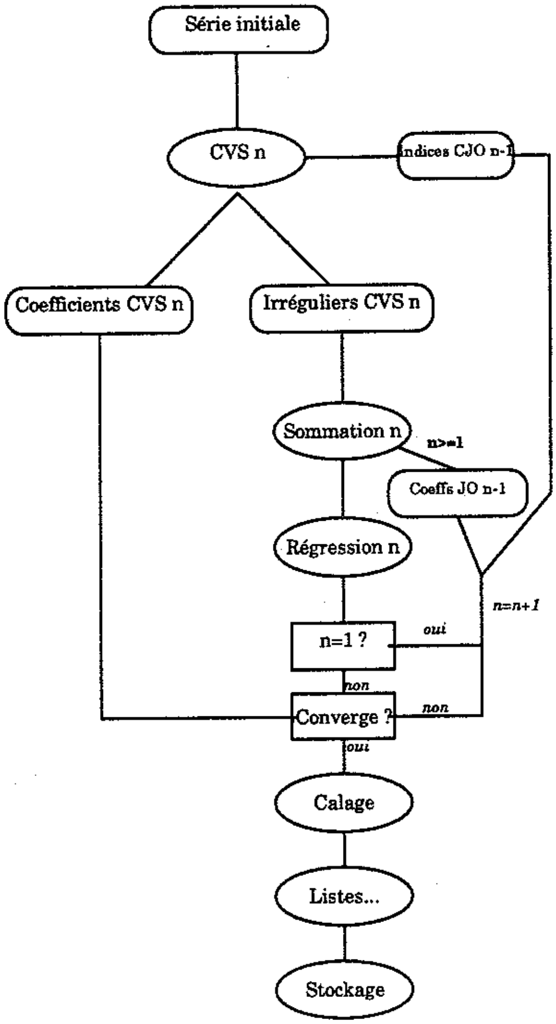
Le principe consiste à désaisonnaliser par la PROC X11 la série (éventuellement) corrigée des jours ouvrables, on récupère alors un log-irrégulier de désaisonnalisation. Cet irrégulier comprend encore un effet de CJO résiduel, c'est pourquoi on l'ajoute aux coefficients de CJO estimés lors de l'étape précédente. Cette somme est régressée par MCO sur les d_i, on obtient ainsi un nouvel estimateur de correction JO et un nouvel indice CJO. La boucle peut alors recommencer. Au départ, les coefficients de CJO sont initialisés à 0 (pas d'effet de CJO).

Comme on s'intéresse principalement à obtenir de bonnes estimations des coefficients de CVS-CJO, le critère de convergence retenu est la révision quadratique moyenne de ces coefficients. Ce critère peut être interprété comme une estimation de la précision finale des coefficients de CVS-CJO et donc de la série corrigée. En deçà d'un seuil paramétrable (fixé par exemple à 0,001 ce qui correspond à une précision de 0,1%) l'algorithme s'arrête[9]. Ce critère se révèle, *a posteriori*, pertinent et robuste : il décroît quand le nombre de boucles augmente, il permet d'arrêter à un seuil raisonnable l'algorithme de CJO, et ce pour la quasi-totalité des séries qui ont été testées (environ une quarantaine).

---

[8] Les coefficients dominicaux sont contraints à 0. Même si la relation (13) n'est pas vérifiée en février elle est proche de l'être les années non bissextiles (la somme vaut -0,25 au lieu de 0). De plus, comme le dimanche est en général un jour creux, il est clair que le coefficient dominical est proche de 0. Sur les séries étudiées, les multiplicateurs associés aux contraintes sur le coefficient dominical n'ont jamais été significativement non nuls.

L'algorithme complet se schématise ainsi :



9. Il s'arrête également au delà d'un certain nombre (paramétrable lui aussi) d'itérations.

I.2.c L'algorithme de calage[10]

Bien que les phénomènes de CJO puissent influer sur les niveaux annuels, il est parfois souhaité que les variables CVS-CJO soient calées aux données brutes sur l'année civile, c'est à dire qu'elles aient, en moyenne, la même valeur sur chaque année. Il s'agit alors de redresser les données CVS-CJO de manière à ce que la somme des valeurs CVS-CJO et la somme des valeurs brutes soient égales sur chaque année civile. Il n'est pas du tout évident (et même presque sûrement faux) que cette contrainte assure également que les données CJO (non-CVS) et brutes seront calées ce qui constituerait pourtant une autre contrainte "naturelle". En général toutefois, le calage annuel des données CVS-CJO et brutes permet de réduire la différence entre les données brutes et les données CJO.

Cette contrainte fixée, il faut trouver un critère qui permette de sélectionner un redressement parmi tous ceux qui satisfont les contraintes imposées. Deux types de critères paraissent naturels : ils consistent à minimiser soit :
- la révision des indices CVS-CJO.
- la révision des évolutions des indices CVS-CJO

En pratique, on se limite aux critères quadratiques qui permettent un formalisme mathématique plus simple et pour lesquels les problèmes de minimisation sont résolubles analytiquement.

La variable est notée x, la donnée brute x^b, la donnée CVS-CJO avant calage est notée x^j, la donnée CVS-CJO après calage x^c avec $x^c = \alpha x^j$. La date est notée avec le double indice *(n,p)* *n* étant l'année et *p* la période (mois ou trimestre). p(n) désigne le nombre de périodes connues de l'année n, en général $n=12$ ou 4 sauf pour la première et la dernière année. Pour des raisons de commodité la première année est supposée complète p(1)=P (4 ou 12), sinon l'année sera retirée ce qui est peu gênant pour les séries assez longues. La dernière année N pose un problème plus délicat quand elle est incomplète puisque le calage est alors impossible. Il est également impossible de caler partiellement les données CVS-CJO sur les n(N) premières données brutes puisque ces dernières étant saisonnières n'ont aucune raison de fournir une bonne estimation de la moyenne. Dans ce cas les données CVS fournissent un meilleur estimateur de la somme, l'avantage étant que ces données sont partiellement calées par le logiciel X11. De plus, la contrainte sur la dernière année est trop forte. Si la dernière année connue se termine en janvier par exemple la contrainte impose que les données CVS-CJO calées soient égales aux données CVS!. C'est pourquoi, il semble meilleur de caler les données CVS-CJO non sur les données CVS, mais sur une interpolation linéaire entre les données CVS-CJO non modifiées et les données CVS. Cette interpolation est de la forme :

$$x^i_{N,p} = \left(1 - \mathrm{f}\left(\tfrac{\mathrm{p}(N)}{P}\right)\right) x^{cvs}_{N,p} + \mathrm{f}\left(\tfrac{\mathrm{p}(N)}{P}\right) x^j_{N,p} \quad P \in \{4, 12\}$$

$$0 \le \mathrm{f}\left(\tfrac{\mathrm{p}(N)}{P}\right) \le 1 \ \text{ t.q } \ \mathrm{f}(1) = 0 \text{ et } \mathrm{f}(0) = 1$$

f est doit normalement être décroissante : plus la dernière année est avancée, plus il faut se caler sur les données CVS. La forme de la fonction f(x) peut être choisie dans la classe des fonctions puissance :

$$\mathrm{f}(x) = \left(1 - x^\beta\right) \text{ avec : } \beta > 0$$

[10] Le calage est optionnel, il fait appel au module SAS/IML®

L'interpolation linéaire ($\beta = 1$) est utilisée par défaut.

Pour simplifier les écritures, nous noterons pour la dernière année $x^b_{N,p}$ pour $x^i_{N,p}$.

Premier critère (moindre révision des niveaux)

On résout :

$$\begin{cases} \text{Min} \sum_{n=1}^{N} \sum_{p=1}^{p(n)} (1-\alpha_{n,p})^2 \\ \forall n, \sum_{p=1}^{p(n)} \alpha_{n,p} x^j_{n,p} = \sum_{p=1}^{p(n)} x^b_{n,p} \end{cases}$$

On en déduit (cf. annexe 2) :

$$\Rightarrow \forall n, p, \alpha_{n,p} = 1 + \left(\frac{\sum_{p=1}^{p(n)} x^b_{n,p} - \sum_{p=1}^{p(n)} x^j_{n,p}}{\sum_{p=1}^{p(n)} (x^j_{n,p})^2} \right) x^j_{n,p}$$

Comme en première approximation les $x^j_{n,p}$ sont égaux sur une année, cette formule montre que les $\alpha_{n,p}$ sont approximativement égaux sur une année. Autrement dit, cette méthode conduit à modifier uniformément tous les mois[11] d'une même année selon un même multiplicateur. Dès lors c'est au changement d'année, entre Décembre et Janvier dans le cas mensuel, que la rupture liée au calage va apparaître puisque ces deux périodes auront des multiplicateurs différents. Cette méthode paraît donc peu satisfaisante. C'est pourquoi un critère un peu plus symétrique qui minimise par exemple les révisions d'évolutions a été préféré.

Deuxième critère (moindre révision des évolutions)

On résout :

$$\begin{cases} \text{Min } (1-\alpha_{1,1})^2 + \sum_{n=1}^{N} \sum_{p=1}^{p(n)} (\alpha_{(n,p)+1} - \alpha_{n,p})^2 \\ \forall n, \sum_{p=1}^{p(n)} \alpha_{n,p} x^j_{n,p} = \sum_{p=1}^{p(n)} x^b_{n,p} \end{cases}$$

$(n,p)+1$ représente la date suivant (n,p)

Par convention : $\alpha_{(N,p(N))+1} = 1$, le terme $(1-\alpha_{1,1})^2$ s'explique par symétrie.

Les calculs complets figurent en annexe 2. Cette méthode donne de bons résultats. Elle est relativement rapide et on constate à l'usage que les profils sont peu modifiés. En revanche, le paramètre β n'a pas été optimisé, il se peut que sa valeur par défaut (1) ne convienne pas notamment pour les séries trimestrielles quand seul le premier trimestre de la dernière année est connu.

[11] Une fois pour toutes, ce qui est valable en mensuel est adaptable au cas trimestriel

I.2.d La robustesse des résultats

La régression principale qui estime les coefficients journaliers utilise les MCO, qui ne sont pas robustes. Pour "robustifier" cette régression, les points aberrants tels que détectés par X11 ont été éliminés de la régression. Ceci a l'avantage d'être un critère transparent et automatique pour l'utilisateur. Il est vrai que la détection des points aberrants par X11 peut-être trop sévère, elle élimine un peu plus de 5% des données. Néanmoins comme les points éliminés à tort sont distribués aléatoirement les estimateurs ne devraient pas être biaisés par cette méthode. Les paramètres par défaut de X11 se révèlent par ailleurs satisfaisants dans cette optique.

I.2.e La stabilité des résultats

La stabilité des équations peut être établie de deux manières :

- d'une part en s'assurant que le critère de convergence est bien rempli et qu'il a bien diminué au fil des itérations. Ce critère empirique teste en fait la stabilité de la convergence de l'algorithme.

- d'autre part en testant la stabilité des régressions de CJO, ceci est fait par un test de Cusum en appliquant la méthode de Ploberger-Kramer. De fait, les équations sont en général relativement stables au moins depuis 1970. Ceci est lié au fait que la plus grande partie de la variance est liée au nombre de jours ouvrés (tous types confondus) et que les évolutions ont été faibles depuis cette époque.

I.3 Améliorations et perspectives

I.3.a La correction pour jour de Pâques

Pâques peut être fêtée soit en Mars (premier trimestre), soit en avril (deuxième trimestre). L'impact de cette fête n'est pas négligeable pour les séries de consommation alimentaire en général et plus particulièrement pour la consommation en produits de la meunerie et en épicerie sèche.

Si les achats liés à la fête Pascale sont de CP, ces achats peuvent se répartir sur deux périodes Mars et Avril (resp. : 1er et 2ème trimestre). La clé de répartition de ces achats dépend de la position relative de la date de Pâques par rapport au 31 Mars. Un modèle simple paramétrable définit une classe naturelle de clés de répartition (voir *annexe 2*). Une fois ces clés mensuelles définies selon:

$$\forall m \neq 3,4 \quad k_m = 0$$
$$k_3 = k \quad k_4 = 1 - k$$

L'effet de Pâques vaut $k_m CP$, il est estimé par MCO, les déflateurs étant les k_m.

I.3.b Les coefficients variables

Les coefficients journaliers sont supposés constants tout au long de la série. En général, ces coefficients semblent stables tout au long de la série. Il peut être néanmoins nécessaire de calculer des coefficients variables. L'équation (13) doit alors être réécrite ainsi (avec les notations classiques)

$$I_t = \sum_{i=2}^{9} \delta_{i,t} d_{i,t} + \varepsilon_t \quad \varepsilon_t \text{ iid}$$

$\forall i, \exists \eta_{i,t}$ bruit blanc indépdt des ε_t tq : $\Phi(L)\delta_{i,t} = \Theta(L)\eta_{i,t}$

Dans le cas markovien simple : $\delta_{i,t} = \delta_{i,t-1} + \eta_{i,t}$

Ce modèle ne peut être estimé par des procédures standard, il faut utiliser un filtrage de Kalman[12]. Faute de disposer d'un tel outil, ce développement n'a pu être mise en oeuvre. Ceci n'est pas gênant pour l'instant, mais il se peut que l'extension progressive du travail dominical modifie quelque peu les coefficients journaliers dans un avenir proche, il sera alors utile de disposer d'une telle amélioration.

Deuxième partie : la correction des aléas de température

II.1 Une formulation mathématique simple

Le problème de correction des effets de température est assez proche de celui de la CJO. Il s'agit d'un phénomène fortement saisonnier (d'ailleurs bien plus saisonnier que l'effet de jours ouvrables) qui doit donc être traité en même temps que la désaisonnalisation. De même que pour les CJO, l'impact de la température *moyenne* -dite normale saisonnière- sera intégré dans la saisonnalité *moyenne* de la série à corriger. Seuls les *écarts* à la normale saisonnière interviendront pour expliquer les *irrégularités* de la série. Les aléas climatiques ont en revanche un impact variable selon la période de l'année : un degré en moins par rapport à la normale saisonnière en décembre n'est pas équivalent à un degré en moins au mois en juillet.

Une idée simple consiste à poser que

$$\begin{aligned} & I_t = k_{p(t)}\left(T_t - \overline{T}_{p(t)}\right) + \varepsilon_t \\ & p(t) \ : \ \textit{période (mois ou trimestre)} \\ & T_t \ : \ \textit{température}, \quad \overline{T}_{p(t)} \ : \ \textit{normale saisonnière} \\ & k_{p(t)} \ : \ \textit{impact d'un degré en plus par rapport à la normale} \end{aligned} \qquad (15)$$

Ce modèle conduit à estimer 12 coefficients (k_1 à k_{12}) pour les séries mensuelles et 4 coefficients dans le cas trimestriel. Pour réduire le nombre de coefficients estimés il est possible de décomposer la série des douze coefficients en série de Fourier sous la forme :

$$k_j = \mu + \alpha_0 \sin(j\pi/6) + \alpha_1 \sin(j\pi/3) + \beta_0 \cos(j\pi/6) + \beta_1 \cos(j\pi/3) \quad (16)$$

[12]Qui n'est malheureusement pas disponible sur SAS®

Les équations (15) et (16) sont alors résolues concomitamment par MCO. On peut encore restreindre le nombre de coefficients estimés en imposant $\beta_0=\beta_1=0$.

II.2 Quelques remarques

La correction pour aléas climatiques est d'un grand intérêt pour analyser nombre de séries économiques, notamment les séries de consommation énergétique ou même de consommation en textiles habillement. Contrairement aux CJO -et le lecteur me pardonnera cette vérité première- elle ne permet en aucun cas de fournir une prévision de ces séries sauf à attendre que la météorologie fournisse des prévisions meilleures et plus lointaines que l'économiste. Il n'y de plus aucune raison de caler les données détempératurisées sur une moyenne annuelle.

Dans la pratique, on calcule directement les écarts à la normale saisonnière comme résidu de la désaisonnalisation de la série des températures brutes. Cette méthode est meilleure puisqu'elle utilise des filtrages proches pour estimer les aléas climatiques et les aléas de la série à corriger.

Dans le cas des séries trimestrielles l'équation (16) est mal spécifiée puisqu'il existe une relation linéaire entre les déflateurs, on impose donc en ce cas :$\beta_1=0$.

L'algorithme est exactement le même que celui décrit pour les CJO, notons que pour l'instant le programme ne permet pas d'effectuer la détempératurisation en dehors de la CJO, il est toutefois possible de séparer *in fine* les effets de CJO des effets climatiques.

Troisième partie : la mise en oeuvre informatique

La procédure décrite est programmée sous forme d'une macro **SAS®**. Sa syntaxe est la suivante :

Paramètres obligatoires

Paramètres facultatifs
Valeurs par défaut en italique

```
%cjp(table,var,
        maxiter=5, seuil=0.001
        sortie=_cjo_
        calage=o, beta=1
        fullw=1.5,zerow=2.5,
        printout=standard, charts=none,
        csem=0, cfer=0, force=n,
        alppaq=0, betpaq=0, paques=n
        detemp=n,nh=1,
)
```

Avec:

table : table SAS où se trouve les données, **cette table doit obligatoirement contenir une variable *de nom date* aux normes SAS de dates.**
var : nom de la variable à corriger (une seule variable admise)
maxiter : nombre maximal d'itérations (en général 5 ou 10 suffisent amplement).
seuil : seuil de convergence, on prend 0.001 ou 0.0005 ce qui signifie que les données corrigées sont fiables à cet ordre de grandeur relatif près.
sortie : nom de la table en sortie
calage : (o/n) calage sur les données brutes de l'année civile
beta : coefficient β de l'algorithme de calage
fullw : paramètre fullweight des demandes X11
zerow : paramètre zeroweight des demandes X11
printout,charts : paramètres de sorties de la demande X11
csem : (0/3/4) contraintes sur les coefficients hebdomadaires
 0 : aucune contrainte
 3 : Lundis, Mardis, Mercredis et Jeudis égaux
 4 : Lundis, Mardis, Mercredis, Jeudis et Vendredis égaux
cfer : (0/1) contraintes sur les coefficients de jours fériés
 0 : le samedi férié est distingué des jours de semaine fériés
 1 : le samedi férié n'est pas distingué des jours de semaine fériés
force : (o/n) o si on veut que l'algorithme de CJO soit lancé même si le test de Fisher rejette l'hypothèse d'existence d'effets de CJO (lors du premier tour). n sinon.
alppaq : (entier positif) paramètre α de correction du jour de Pâques
betpaq : (entier positif) paramètre β de correction du jour de Pâques
paques : (o/n) o si on veut corriger des effets de Pâques
dtemp : (o/n) o si on veut corriger des effets de température, n sinon
nh : (1/2) nombre d'harmoniques pour le calculs des coefficients climatiques
 1 : $\beta_0=\beta_1=0$, 2 sinon

Le contenu de la table en sortie :
La table en sortie contient la variable date, la variable initiale plus les variables suivantes :

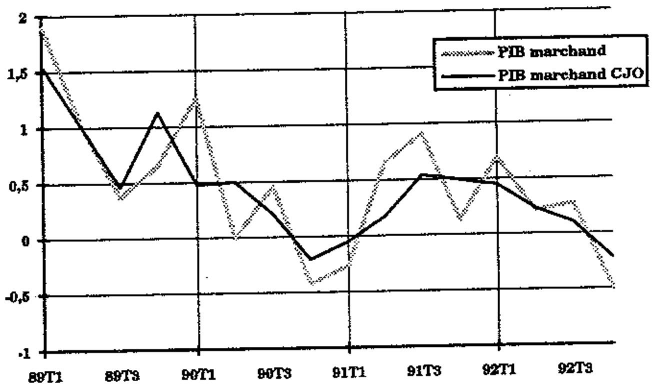
kcjo	Coefficient de CJO
kcvs	Coefficient de CVS
kcvsjo	Coefficient de CVS et CJO (sauf si détempératurisation)
icvsjo	Variable CVS et CJO (sauf si détempératurisation)
icvs1	Variable CVS au premier tour (avant CJO)
calage	Coefficient de calage sur données annuelles (des données CVS-CJO)
_kt_	Coefficient de correction de la température
kcvsjt	Coefficient de correction de CVS-JO et température
icvsjt	Variable CVS-CJO et détempératurisée.

Quelques exemples de correction des jours ouvrables et de détempératurisation

Le PIB trimestriel

Le Produit Intérieur Brut marchand trimestriel (CVS) est affecté d'assez fortes fluctuations à court terme. L'application de la procédure de CJO est ici possible même si elle ne part pas de données brutes, en effet la régression se faisant sur des données "orthogonales" à la saisonnalité, les estimateurs ne seront pas biaisés.

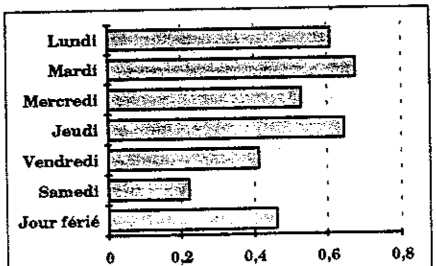
Comparaison du PIB marchand CVS et CVS-CJO
Glissements trimestriels



Une grande partie des fluctuations du PIB marchand est expliquée par les effets de CJO.

Les coefficients journaliers

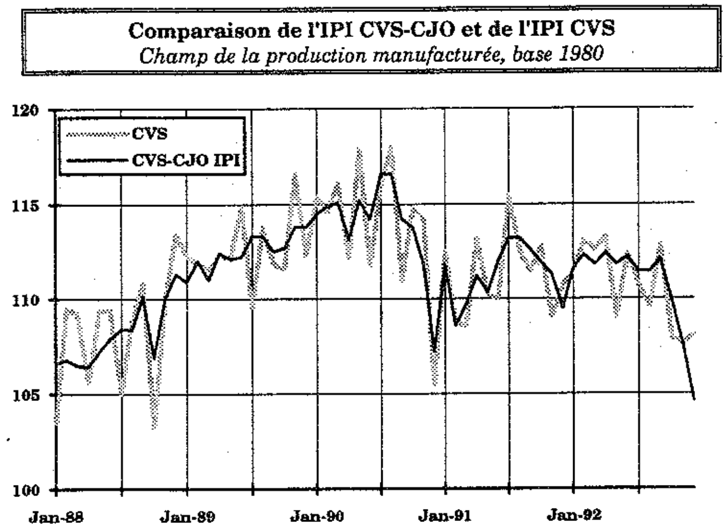
montrent que le PIB marchand est maximal sur les quatre premiers jours de la semaine, il baisse sensiblement dès le vendredi et est fortement réduit le samedi. L'impact des jours fériés est positif : ceci s'explique par un transfert sur les autres jours du mois de la "production" qui ne sera pas effectuée lors de la fermeture. Ce transfert n'existant naturellement pas pour les dimanches, le coefficient des jours fériés est supérieur à celui des dimanches (qui vaut 0).

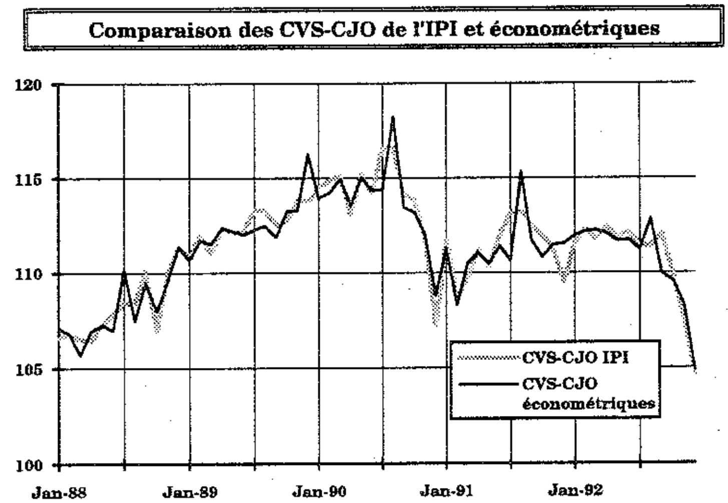


Lecture *: la substitution d'un vendredi à un samedi induit une augmentation relative de 0,2% du PIB marchand : Coef(Vendredi)-Coef(Samedi)=0,4-0,2.*

L'Indice de la Production Industrielle Manufacturière

Cet indice est déjà corrigé des jours ouvrables aux niveaux fins selon des méthodes non-économétriques. La correction telle qu'elle est faite permet de réduire fortement la variabilité de la série mensuelle agrégée:

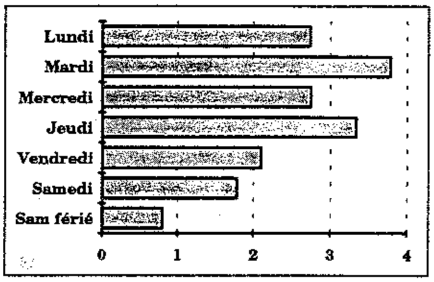




En comparant cet indice CVS-CJO avec celui calculé par l'algorithme précédent, on constate que la méthode économétrique fournit une correction assez proche. Globalement, elle apparaît équivalente, la plupart des divergences provenant des mois d'été sur lesquels la section responsable de l'IPI opère un traitement particulier (les Indices CVS-CJO de Juillet et Août sont égaux, les indices bruts de ces mois font l'objet d'un traitement largement conventionnel).

Les coefficients journaliers

ressemblent à ceux trouvés pour le PIB trimestriel sauf pour les jours fériés : la production est forte du mardi au jeudi, un peu moins élevée à la fois en début et en fin de semaine. Le samedi est relativement plus faible que pour le PIB ce qui s'explique partiellement par le fait que le champ ne comprend pas certains secteurs qui produisent le samedi (énergie, commerce).

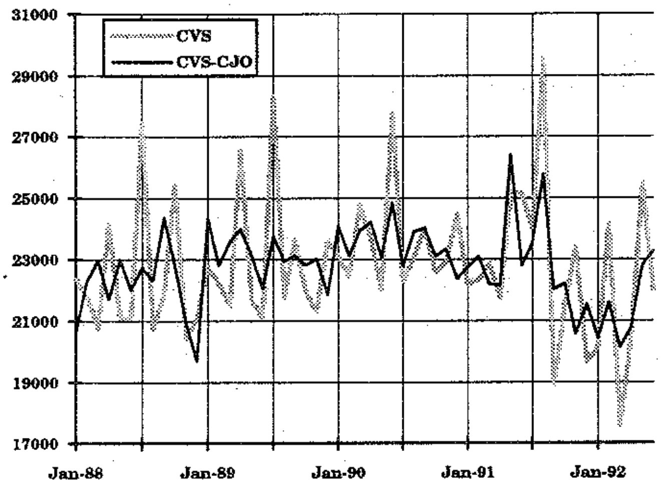


NB : Les jours fériés ont un impact comparable aux dimanches.

Les mariages

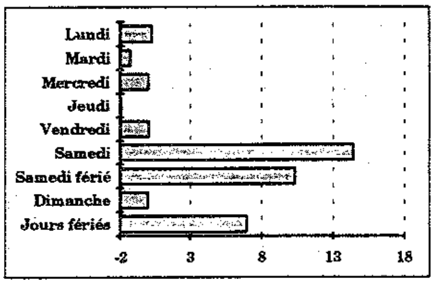
La série mensuelle du nombre de mariage est, on s'en doute, extrêmement sensible au nombre de samedis du mois. A titre anecdotique, la procédure de CJO a été testée sur cette série. Elle donne le résultat suivant :

CJO des séries de mariages célébrés
Séries mensuelles en niveau



Les coefficients journaliers

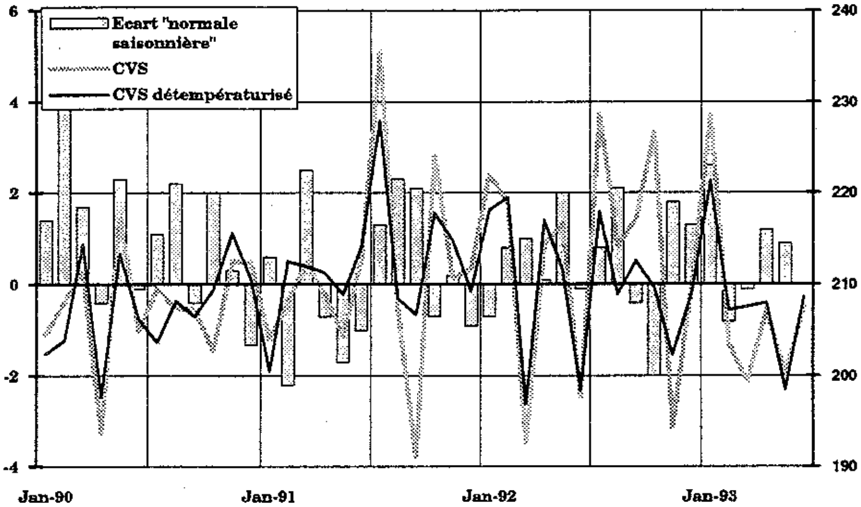
montrent qu'effectivement les mariages ont souvent lieu le samedi (qu'il soit férié ou non). Les jours fériés ont aussi un coefficient particulièrement fort, plus élevé que celui des dimanches. Il faut sans doute interpréter ceci par le fait que les jours fériés sont souvent accompagnés de "ponts" qui sont propices aux mariages. En ce cas, il n'est pas évident que les mariages aient lieu exactement le jour férié mais plutôt un jour proche (et peut-être un samedi).



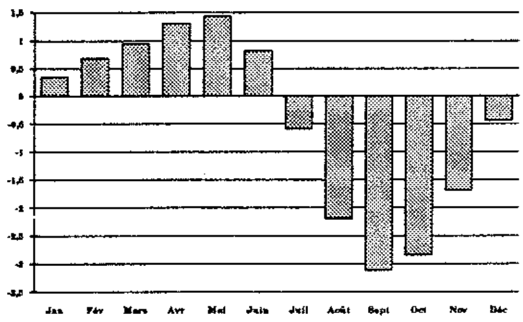
Consommation en textile-habillement et aléas climatiques

La consommation en habillement est marquée par un rythme saisonnier caractéristique que les phénomènes de mode entretiennent. Le renouvellement des collections se fait essentiellement selon une périodicité bisannuelle printemps-été et automne-hiver. En revanche, les dates d'achat de ces collections dépendent fortement du climat. Un printemps précoce fait avec les premiers beaux jours d'Avril chanter les tiroirs-caisses des commerçants. Si l'automne est précoce, les premières rigueurs d'Octobre précipitent le chaland dans les magasins. La procédure de détempératurisation a ainsi été testée sur la consommation textile.

Consommation textile et aléas climatiques

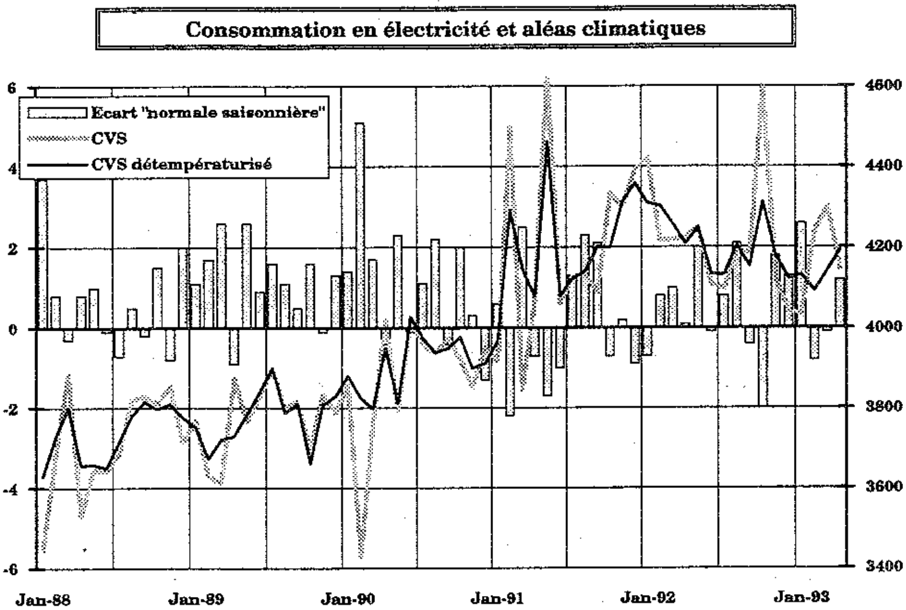


Le profil des coefficients climatiques montre d'ailleurs qu'une température supérieure à la normale saisonnière est favorable au printemps et défavorable à l'automne.

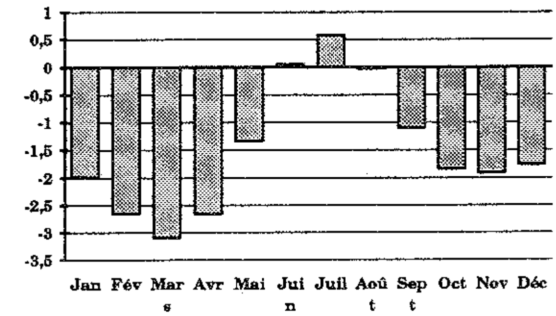


Impact en % d'un degré en plus par rapport à la normale saisonnière, selon le mois.

Consommation en électricité et aléas climatiques



Le profil des coefficients climatiques montre d'ailleurs qu'une température inférieure à la normale saisonnière augmente presque toujours la consommation électrique, notamment au printemps.



Impact en % d'un degré en plus par rapport à la normale saisonnière, selon le mois.

Bibliographie

BELL, W.R. & HILLMER, S.C., (1983),
Modelling Times Series With Calendar Variations,
Journal of the American Statistical Association,
Volume 78, N° 383, 526-535

CLEVELAND, W.P. & GRUPE, M.R., (1983),
Modelling Times Series When Calendar Effects Are Present,
Applied Times Series Analysis of Economic Data, ed. A. Zellner,
Washington D.C.: U.S. Dept. of Commerce, Bureau of Census, 57-73

DAGUM, E.B., (1980), The X11-ARIMA Seasonnal Adjustment Method,
Statistics Canada, Catalogue N°. 12-564E

DAGUM, E.B., QUENNEVILLE, B. & SUTRADHAR, B., Trading-day Variations Multiple Regression Models with Random Parameters,
International Statistical Review (1992), 60

LAROQUE, GUY, (1977), Analyse d'une méthode de désaisonnalisation : le programme X11, version trimestrielle, *Annales de l'INSEE*, N°28

Note INSEE / Service de la Conjoncture N° JPD 343/902 du 16 mars 1979
Le programme Census X11 et la correction des variations saisonnières

Annexe 1 : les algorithmes de calage

Premier critère (moindre révision des niveaux)

On résout :

$$\begin{cases} \mathrm{Min} \sum_{n=1}^{N} \sum_{p=1}^{p(n)} (1-\alpha_{n,p})^2 \\ \forall n, \sum_{p=1}^{p(n)} \alpha_{n,p} x^j_{n,p} = \sum_{p=1}^{p(n)} x^b_{n,p} \end{cases}$$

En dérivant le Lagrangien, on écrit les conditions du premier ordre (qui sont visiblement suffisantes) :

$$\frac{\partial \ell}{\partial \alpha_{n,p}} = 2(\alpha_{n,p} - 1) - 2\lambda_n x^j_{n,p} = 0$$

$$\Rightarrow \alpha_{n,p} = 1 + \lambda_n x^j_{n,p}$$

$$\Rightarrow \forall n, \sum_{p=1}^{p(n)} (1+\lambda_n x^j_{n,p}) x^j_{n,p} = \sum_{p=1}^{p(n)} x^b_{n,p}$$

$$\Rightarrow \forall n, \lambda_n = \frac{\sum_{p=1}^{p(n)} x^b_{n,p} - \sum_{p=1}^{p(n)} x^j_{n,p}}{\sum_{p=1}^{p(n)} (x^j_{n,p})^2}$$

$$\Rightarrow \forall n, p, \alpha_{n,p} = 1 + \left(\frac{\sum_{p=1}^{p(n)} x^b_{n,p} - \sum_{p=1}^{p(n)} x^j_{n,p}}{\sum_{p=1}^{p(n)} (x^j_{n,p})^2} \right) x^j_{n,p}$$

Deuxième critère (moindre révision des évolutions)

On résout :

$$\begin{cases} \mathrm{Min}\ (1-\alpha_{11})^2 + \sum_{n=1}^{N} \sum_{p=1}^{p(n)} (\alpha_{(n,p)+1} - \alpha_{n,p})^2 \\ \forall n, \sum_{p=1}^{p(n)} \alpha_{n,p} x^j_{n,p} = \sum_{p=1}^{p(n)} x^b_{n,p} \end{cases}$$

$(n,p)+1$ représente la date suivant (n,p)

Par convention : $\alpha_{(N,p(N))+1} = 1$, le terme $(1-\alpha_{11})^2$ s'explique par symétrie.

On calcule de même les conditions de premier ordre sur le Lagrangien :

$$\forall (n,p) \notin \{(1,1),(N,\mathrm{p}(N))\}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha_{n,p}} &= \frac{\partial}{\partial \alpha_{n,p}}\left[\left(\alpha_{(n,p)+1}-\alpha_{n,p}\right)^2+\left(\alpha_{n,p}-\alpha_{(n,p)-1}\right)^2\right]-2\lambda_n x^j_{n,p} \\
&= 2(-\alpha_{(n,p)+1}+2\alpha_{n,p}-\alpha_{(n,p)-1}-\lambda_n x^j_{n,p}) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha_{1,1}} &= \frac{\partial}{\partial \alpha_{1,1}}\left[\left(\alpha_{1,2}-\alpha_{1,1}\right)^2+\left(\alpha_{1,1}-1\right)^2\right]-2\lambda_1 x^j_{1,1} \\
&= 2(2\alpha_{1,1}-\alpha_{1,2}-1-\lambda_1 x^j_{1,1}) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha_{N,p(N)}} &= \frac{\partial}{\partial \alpha_{N,p(N)}}\left[\left(\alpha_{N,\mathrm{p}(N)}-\alpha_{(N,\mathrm{p}(N))-1}\right)^2+\left(\alpha_{N,\mathrm{p}(N)}-1\right)^2\right]-2\lambda_N x^j_{1,1} \\
&= 2(2\alpha_{N,\mathrm{p}(N)}-\alpha_{(N,\mathrm{p}(N))-1}-1-\lambda_N x^j_{N,\mathrm{p}(N)}) \\
&= 0
\end{aligned}$$

Ce qui peut-être noté matriciellement :

$$R=\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \qquad M=\begin{pmatrix} x^j_{1,1} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ x^j_{1,4} & 0 & & \vdots \\ 0 & x^j_{2,1} & & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \vdots & x^j_{2,4} & & x^j_{N,1} \\ \vdots & 0 & & \vdots \\ 0 & \cdots & 0 & x^j_{N,\mathrm{p}(N)} \end{pmatrix}$$

$$\alpha=\begin{pmatrix} \alpha_{1,1} \\ \vdots \\ \vdots \\ \alpha_{N,\mathrm{p}(N)} \end{pmatrix} \qquad \lambda=\begin{pmatrix} \lambda_1 \\ \vdots \\ \vdots \\ \lambda_N \end{pmatrix} \qquad b=\begin{pmatrix} x^b_{1,1} \\ \vdots \\ \vdots \\ x^b_{N,\mathrm{p}(N)} \end{pmatrix} \qquad l=\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}_{(T,1)}$$

Ona alors : $R\alpha = M\lambda + l$ et ${}^tM\alpha = b$

Ces équations peuvent se résoudre en :

$$\alpha = R^{-1}M({}^tMR^{-1}M)^{-1}(b-{}^tMR^{-1}l)+R^{-1}l$$

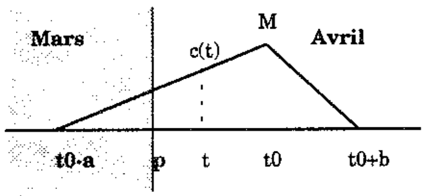
Calcul qui peut être simplifié en remarquant que la matrice R (de taille T) est inversible analytiquement.

Annexe 2 : l'effet "Pâques"

On associe à chaque mois[13] une variable indicatrice k_m du "poids" de l'effet Pâques. Par définition :

$$\forall a, \sum_{m=1}^{12} k_m(a) = 1, a \text{ représente l'année}$$
$$\forall m \notin \{3,4\}, k_m(a) = 0$$

Une modélisation simple de l'effet de Pâques *journalier* consiste à supposer que cet effet est nul jusqu'à un certain temps avant Pâques, qu'il augmente linéairement ensuite jusqu'à atteindre une valeur maximale le jour de Pâques, puis qu'il diminue linéairement. L'effet *mensuel* est calculé par sommation.



t0 : jour de Pâques, p=31 Mars
a=longueur de l'effet Pâques avant
b=longueur de l'effet Pâques après
c(t)=effet journalier
M=effet journalier maximal

Par définition: $\sum_{t=t_0-a}^{t_0+b} c(t) = 1$

D'où : $M\left(1+\sum_{t=t_0-a}^{t_0-1}\left(\frac{t-(t_0-a)}{a}\right)+\sum_{t=t_0+1}^{t_0+b}\left(\frac{(t_0+b)-t}{b}\right)\right)=1$

Soit: $M=\frac{2}{a+b}$

Les poids rétrospectifs de Mars (π_{Mars}) et d'Avril (π_{Avril}) sont définis par:

$\pi_{Mars} = \sum_{t=t_0-a}^{p} c(t), \pi_{Avril} = 1-\pi_{Mars}$

Si $p \le t_0 - a$ alors $\pi_{Mars} = 0$

Sinon si $p < t_0$ alors $\pi_{Mars} = \frac{1}{a(a+b)}(p+a-t_0)(p+a+1-t_0)$

Sinon si $p = t_0$ alors $\pi_{Mars} = \frac{a+1}{a+b}$

Sinon si $p \le t_0 + b$ alors $\pi_{Mars} = 1-\frac{1}{b(a+b)}(t_0+b-p)(t_0+b-p-1)$

Sinon $\pi_{Mars} = 1$

Les paramètres a et b ont été recherchés par balayage à partir de certaines séries économiques sensibles telles que le chiffre d'affaires de l'ensemble du commerce de détail et de boulangerie-pâtisserie, en cherchant à améliorer à la fois le R^2 et les students. a=8 et b=4 constitue un bon compromis. Les k_m sont ensuite ajoutés aux déflateurs de l'équation (13).

[13] Pour des raisons de simplification , seul le cas mensuel est traité ici.