

# REGRESSEURS POUR EFFETS DE CALENDRIER : COMMENT LES CONSTRUIRE, COMMENT LES CHOISIR ?

Ketty ATTAL-TOUBERT (\*)

(\*) Insee, Département des statistiques de court terme

## Introduction

La correction des effets de jours ouvrables est devenue, au même titre que la correction des variations saisonnières, une étape presque systématique du traitement des séries économiques, afin de clarifier le message conjoncturel qu'elles contiennent. Les principaux logiciels de désaisonnalisation, Tramo-Seats et X12-ARIMA, permettent de réaliser ce traitement, à l'aide de régresseurs prédéfinis, dans le cadre d'un modèle Reg-ARIMA. Cependant, les régresseurs prédéfinis ne sont pas adaptés à toutes les séries. En particulier, les régresseurs par défaut ne prennent pas en compte les spécificités des calendriers nationaux. Certes, des outils sont proposés, dans les logiciels eux-mêmes ou dans les interfaces permettant de les utiliser, pour intégrer des éléments sur le calendrier national (liste des jours fériés), mais la façon dont ils sont pris en compte dans les régresseurs manque de transparence et de nombreux problèmes ont été rencontrés par les utilisateurs.

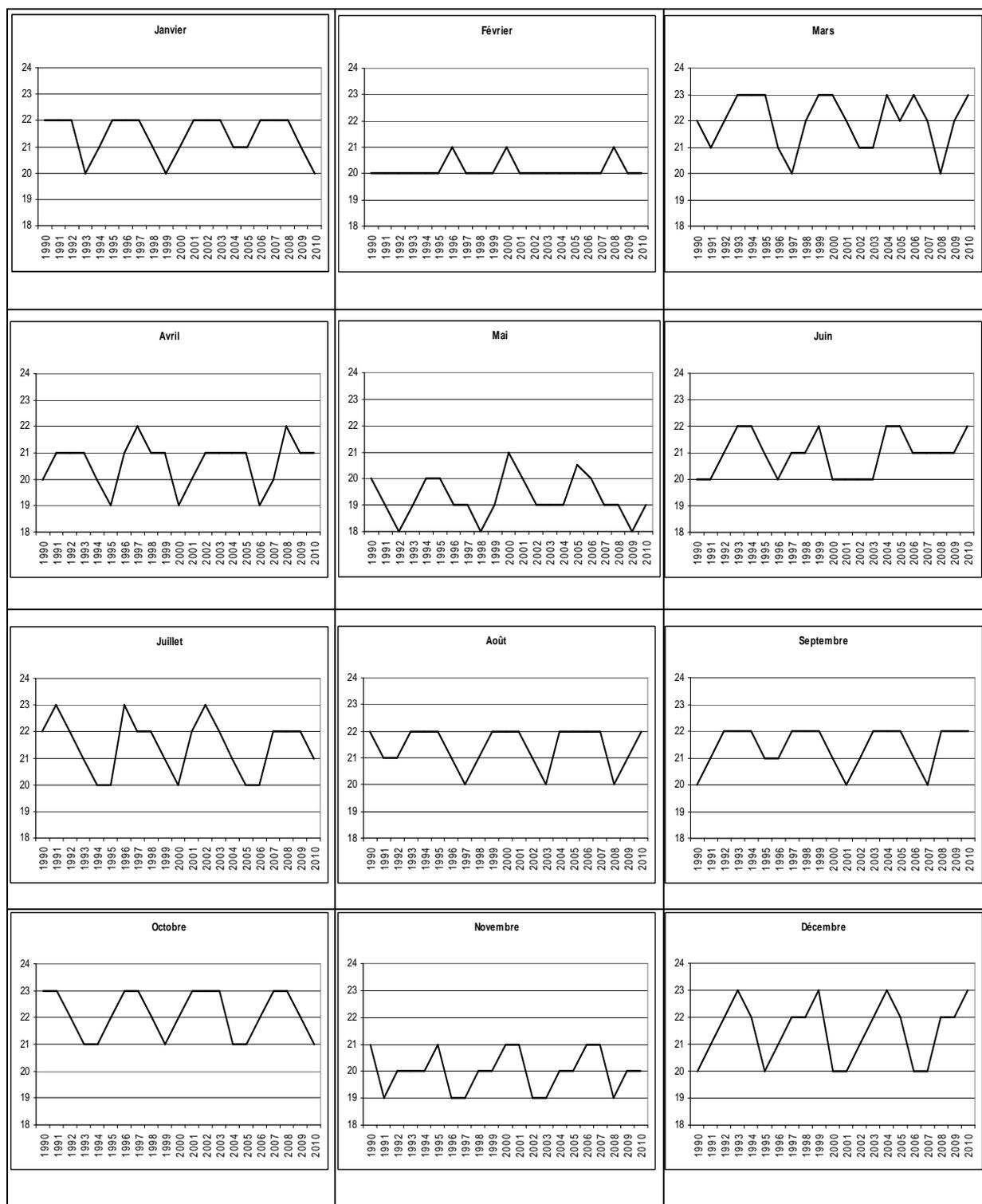
Ainsi la plupart des organismes produisant des séries corrigées des effets de jours ouvrables créent leurs propres régresseurs. Toutefois, il n'existe pas vraiment de concertation entre eux sur la méthode à employer ni sur la façon de construire les régresseurs. Il paraît pourtant souhaitable d'harmoniser les pratiques. Cet article a pour objectif de proposer une description pas à pas de la construction de régresseurs pour effets de calendrier à partir du modèle de base jusqu'à une écriture plus générale. Les régresseurs ainsi construits tiennent compte des spécificités du calendrier national, ainsi que du secteur d'activité étudié. L'article donne par ailleurs un exemple de méthode permettant de sélectionner le jeu de régresseurs le mieux adapté à chaque série et les résultats de l'application de cette méthode aux séries d'indices de chiffres d'affaires de l'Insee.

## 1. Construction de régresseurs pour les effets de calendrier

### 1.1. Les effets de calendrier

L'analyse de la conjoncture est fondée sur l'utilisation de données statistiques de fréquence mensuelle ou trimestrielle. Celles-ci reflètent le plus souvent une activité journalière, dont on n'a une mesure qu'une fois par mois ou par trimestre. Dans certaines séries, cette mesure correspond au total d'une activité au cours du mois (séries de *flux*), dans d'autres, il s'agit de la valeur d'une grandeur économique à une date donnée, par exemple en fin de mois (séries de *stock*). Dans les séries de flux, le nombre et le type de jours qui varient d'un mois ou d'un trimestre à l'autre peuvent expliquer les fluctuations de court terme. Plus précisément, les séries de flux sont influencées par le nombre et le type de jours « travaillés » (jours d'activité). Ainsi, si l'on fait l'hypothèse simpliste que l'activité est uniforme sur tous les jours travaillés, un mois comptant 22 jours travaillés aura 4,8 % d'activité de plus qu'un mois à 21 jours travaillés, et un trimestre à 64 jours travaillés 1,6 % de plus qu'un trimestre n'en comptant que 63. Dans le commerce de détail en France, cette hypothèse n'est généralement pas vérifiée, puisque les samedis sont caractérisés par une activité plus intense que les autres jours. On observe alors des fluctuations liées aux types de jours de la semaine : un mois comptant 5 samedis enregistre une activité plus importante qu'un mois n'en comportant que 4. Les séries de stocks peuvent également présenter des fluctuations liées au calendrier avec toutefois certaines spécificités qui ne seront pas développées dans le présent article.

Graphique 1 : Nombre de jours ouvrables en France sur la période 1990-2010

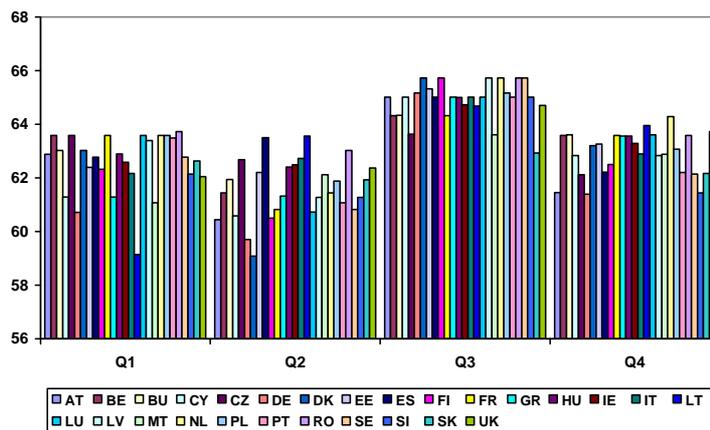


On distingue deux types de jours travaillés : les jours *ouvrables*, qui sont les jours censés être travaillés, et les jours *ouvrés*, qui sont les jours effectivement travaillés. Dans de nombreuses activités économiques, les jours ouvrables sont les lundis, mardis, mercredis, jeudis, vendredis non fériés. Le *graphique 1* représente le nombre de jours ouvrables en France (lundis, mardis, ..., vendredis non

fériés<sup>1</sup> sur la période 1990-2010). Celui-ci varie selon les mois entre 18 et 23. A certaines dates la configuration du calendrier est inhabituelle : les mois de mars de 1997 et 2008 sont caractérisés par un nombre de jours ouvrables plus faible que de coutume (20 au lieu de 21 à 23), à l'inverse les mois d'avril 1997 et 2008 comportent 22 jours ouvrables (au lieu de 19 à 21 habituellement) et le mois de mai 2000 comporte 21 jours ouvrables (au lieu de 18 à 20 habituellement)<sup>2</sup>.

Compte tenu des différences de calendrier des jours fériés, le nombre de jours ouvrables peut varier fortement d'un pays à l'autre, comme le montre le *graphique 2* qui représente le nombre de jours ouvrables par trimestre entre 1940 et 2010 dans les pays européens. Cette diversité des calendriers conduit Eurostat à demander à chaque état-membre d'effectuer sa propre correction des effets de calendrier.

Graphique 2 : Nombre de jours ouvrables par trimestre entre 1940 et 2010 dans les pays européens



Pour améliorer la lisibilité et la comparabilité des indicateurs conjoncturels, il est donc souhaitable d'éliminer ces fluctuations déterministes. Notons que, en général, ce que l'on cherche à gommer, ce sont les fluctuations liées au nombre et au type de jours *ouvrables*, et non de jours *ouverts* : si certains jours sont exceptionnellement ouverts alors qu'ils ne sont habituellement pas des jours travaillés, c'est une information conjoncturelle que l'on veut conserver, signe par exemple d'une forte demande.

Le processus de correction des variations saisonnières résout partiellement le problème des fluctuations liées au calendrier. En effet, le mois de mars par exemple a presque toujours plus de jours ouvrables que le mois de février. Par ailleurs, la plupart des jours fériés tombent toujours le même mois. En revanche, le nombre de jours ouvrables supplémentaires du mois de mars par rapport au mois de février n'est pas toujours le même d'une année sur l'autre. De plus, un jour férié, même s'il tombe toujours le même mois, n'influencera pas de la même façon l'activité s'il tombe pendant le week-end ou pendant la semaine. Enfin, il existe des jours fériés mobiles : Pâques, l'Ascension, la Pentecôte, qui ne tombent pas toujours le même mois. Leur impact sur l'activité (ex. augmentation des ventes de chocolats ou d'agneau dans les jours qui précèdent Pâques) ne se produit pas chaque année à la même période. Il est donc nécessaire d'effectuer un traitement spécifique pour les effets de calendrier qui ne sont pas saisonniers.

En résumé, on peut définir trois principaux types d'effets de calendrier :

- les effets de « jours ouvrables », liés au nombre de jours ouvrables à chaque période
- les effets « type de jour de la semaine », liés au nombre de jours de chaque type (lundi, mardi...) à chaque période
- les effets « fête mobile », liés à la variation d'une année sur l'autre de la date de certaines fêtes, comme Pâques.

<sup>1</sup> Les jours fériés étant : le 1<sup>er</sup> janvier, le 1<sup>er</sup> mai, le 8 mai, le 14 juillet, le 15 août, le 1<sup>er</sup> novembre, le 11 novembre, le 25 décembre, le lundi de Pâques, l'Ascension, la Pentecôte.

<sup>2</sup> Notons qu'en mai 2005, la Pentecôte a été comptée comme un demi jour férié, car bien que cette année-là elle n'était pas officiellement un jour férié, de nombreuses entreprises l'ont accordé comme tel à leur salariés.

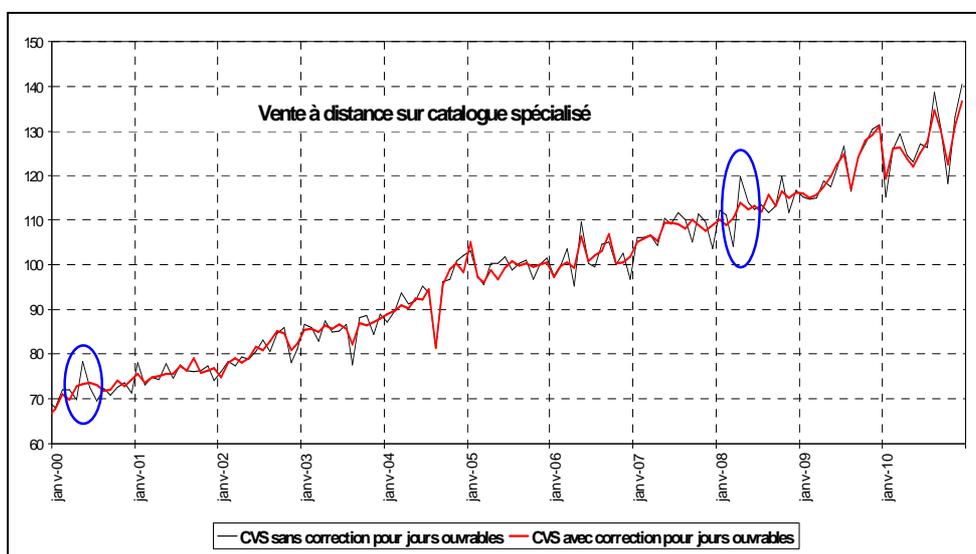
Seuls les deux premiers types d'effets de calendrier seront traités dans cet article. Pour plus d'informations sur la modélisation des effets des fêtes mobiles, voir Bessa et *alii* [5] ou Ladiray et Quenneville [16].

Afin de mieux visualiser l'ampleur que peuvent avoir les fluctuations liées aux effets de calendrier, on a représenté sur un même graphique (cf. *graphique 3*) l'indice de chiffre d'affaires de la vente à distance sur catalogue spécialisé :

- d'une part corrigé des variations saisonnières (CVS)
- d'autre part corrigé des variations saisonnières et des effets de jours ouvrables<sup>3</sup>.

Ce graphique fait apparaître des pics d'ampleur non négligeable liés au nombre de jours ouvrables : en mai 2000 par exemple où le nombre de jours ouvrables était particulièrement élevé (voir *graphique 1*), ou encore en mars et en avril 2008 (effet de ciseau). La correction des effets de jours ouvrables permet de les gommer.

*Graphique 3*: Indice de chiffre d'affaires dans la vente à distance sur catalogue spécialisé : comparaison de la série corrigée des variations saisonnières (CVS) avec la série CVS et corrigée des effets de jours ouvrables.



De même, pour évaluer l'effet spécifique du type de jour de la semaine, on compare sur un même graphique (cf. *graphique 4*) la série des chiffres d'affaires dans les hypermarchés :

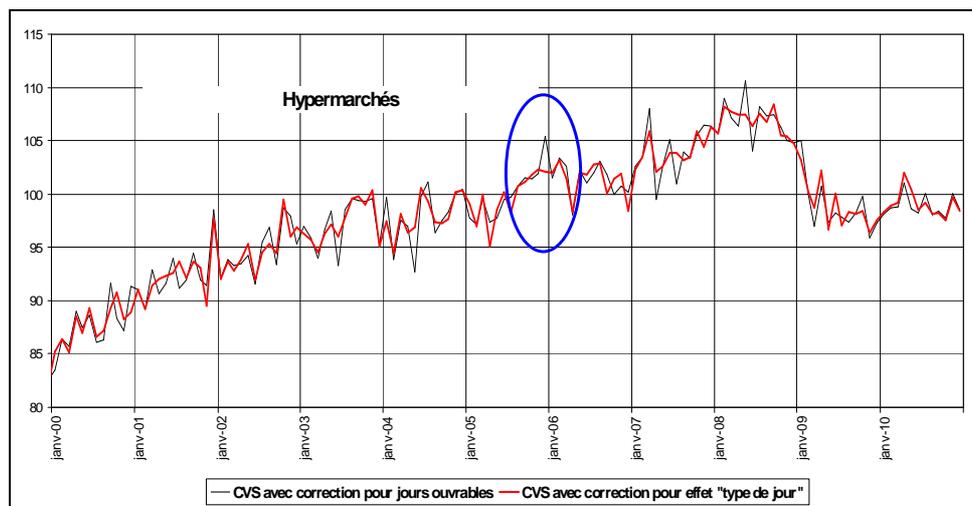
- d'une part corrigée des variations saisonnières et des effets de jours ouvrables<sup>4</sup>.
- d'autre part corrigée des variations saisonnières et de l'effet type de jour de la semaine.

Dans le secteur des hypermarchés, les jours ne sont pas équivalents : l'impact du type de jour de la semaine sur l'activité est fort. Effectuer comme précédemment une correction des effets de jours ouvrables, en considérant les jours de la semaine comme équivalents par opposition aux week-ends, n'est pas efficace puisqu'elle repose sur une hypothèse qui n'est pas valide. La série ainsi corrigée présente donc encore des fluctuations qui peuvent être éliminées par une correction de l'effet type de jour de la semaine. C'est ce que l'on observe par exemple en décembre 2005 (cercle en bleu), où l'on comptait 5 samedis, et où Noël tombait un dimanche.

<sup>3</sup> Ici, ont été considérés comme jours ouvrables les jours de la semaine du lundi au vendredi non fériés

<sup>4</sup> Avec la même définition des jours ouvrables que dans le graphique précédent, à savoir les lundis aux vendredis non fériés.

Graphique 4 : Indice de chiffre d'affaires dans les hypermarchés : comparaison de la série CVS et corrigée des effets de jours ouvrables avec la série CVS et corrigée de l'effet type de jour jour.



## 1.2. Bref historique de la correction des effets de calendrier

Pendant longtemps, les statisticiens ont utilisé la « règle de trois » pour corriger les effets de jours ouvrables : on appliquait à la donnée brute un coefficient égal au rapport entre un nombre de jours ouvrables de référence (par exemple le nombre de jours ouvrables d'un mois moyen de l'année de base pour un indice, ou le nombre de jours ouvrables d'un mois moyen sur la période d'observation) et le nombre de jours ouvrables du mois sous revue. En particulier pour l'indice de production industrielle de l'Insee, des enquêtes sectorielles sur les rythmes de travail permettaient de déterminer les jours habituellement travaillés. Des poids différents étaient le cas échéant attribués aux différents types de jours. Cette méthode conduisait souvent à surestimer les effets de jours ouvrables : elle suppose en effet que la production est proportionnelle au nombre de jours ouvrables, alors que les chefs d'entreprises peuvent ajuster leurs rythmes pour éviter qu'elle ne soit trop affectée par les fluctuations calendaires.

La correction par méthode économétrique a été introduite dans le logiciel de désaisonnalisation X11 en 1980 (Dagum [10]). Elle présente l'avantage de pouvoir tester les effets de calendrier et donc d'éviter une surcorrection. Maillard [18] a repris le modèle de X11 en proposant comme amélioration importante la prise en compte du calendrier national des jours fériés, ainsi que le traitement des séries trimestrielles, qui étaient absents dans X11. Sa méthode de correction itérative a été utilisée au département de la conjoncture de l'Insee et pour les séries de l'indice de la production industrielle jusqu'à la fin des années 1990. Aujourd'hui les principaux logiciels de désaisonnalisation, Tramo-Seats et X12-ARIMA, permettent de corriger les effets de jours ouvrables par la méthode économétrique au moyen de régresseurs prédéfinis, et le logiciel Demetra+, outil permettant d'utiliser ces deux logiciels, offre la possibilité de prendre en compte le calendrier national. Cependant, comme évoqué plus haut, la transparence n'est pas totale sur la façon dont les régresseurs sont construits. De plus, comme les logiciels sous-jacents, Demetra+ ne propose que deux types de modélisation :

- celle où l'on considère comme jours ouvrables les jours de la semaine (du lundi au vendredi, sans les distinguer) par opposition aux week-ends et jours fériés ;
- celle où l'on distingue l'effet de chaque type de jour ouvrable du lundi au samedi, par opposition aux dimanches et jours fériés.

Ces modélisations ne sont pas adaptées à tous les secteurs d'activité : par exemple dans le commerce de détail d'autres configurations pourraient être envisagées : isoler l'effet du samedi, jour de forte activité, du lundi, jour de faible activité, mais regrouper les autres jours (du mardi au vendredi). De plus, dans certains pays, les jours « non ouvrables » ne sont pas les dimanches mais les vendredis ou les samedis. Enfin, certaines séries ne mesurent pas l'activité d'un mois donné mais l'évolution entre deux périodes : c'est le cas des soldes d'opinions des enquêtes de conjoncture sur

l'évolution de l'activité par rapport au mois précédent : à la place de régresseurs « coïncidents », il faut définir des régresseurs mesurant la différence de structure du calendrier entre les deux périodes comparées (différence première entre deux mois par exemple).

Dans la pratique il est donc préférable de reconstruire les régresseurs pour la correction des effets de jours ouvrables. La suite de l'article propose une description de cette démarche.

### 1.3. Les modèles de base

#### 1.3.1. Le modèle de base

Le modèle de base défini dans la littérature (Young [24]) consiste à décomposer la valeur de la variable  $X$  à la date  $t$  selon les jours de la semaine, par l'écriture suivante :

$$X_t = \sum_{i=1}^7 \alpha_i N_{it} + \varepsilon_t$$

où :

- $X_t$  est la série sur laquelle on veut tester et corriger le cas échéant les effets de calendrier,
- les  $N_{it}$  sont les nombres de lundis ( $i=1$ ), mardis ( $i=2$ ), ..., dimanches ( $i=7$ ) à la date  $t$ ,  $t$  étant un mois (ou trimestre) d'une année donnée et  $\varepsilon_t$  est le résidu.

Les coefficients  $\alpha_i$  s'interprètent comme l'effet moyen d'un jour de type  $i$  sur la variable  $X$ .

Notons qu'un tel modèle repose sur l'hypothèse que l'effet de chaque type de jour de la semaine est constant pendant toute la durée d'observation de la série<sup>5</sup>.

Ce modèle présente un problème de colinéarité, évoqué notamment dans Bell et Hillmer [3]. On peut en effet constater que les estimateurs des coefficients  $\alpha_i$  sont très corrélés. Une réécriture du modèle sous une autre forme permet de résoudre ce problème.

Soit  $\bar{\alpha} = \sum_{i=1}^7 \alpha_i$  la moyenne des coefficients.

Pour  $i=1, 2, \dots, 7$  on définit  $\beta_i = \alpha_i - \bar{\alpha}$

On a alors :

$$X_t = \sum_{i=1}^7 \beta_i N_{it} + \bar{\alpha} \sum_{i=1}^7 N_{it} + \varepsilon_t \quad (1)$$

Le terme  $\sum_{i=1}^7 N_{it}$ , que l'on notera dans la suite  $N_t$  est la longueur du mois  $t$  (ou du trimestre  $t$ ).

Cette écriture permet une interprétation intéressante du modèle :

- $\bar{\alpha}$  s'interprète comme l'effet moyen d'un jour quelconque.
- $\beta_i$  s'interprète comme la différence entre l'effet moyen d'un jour de type  $i$  et l'effet moyen d'un jour quelconque, donc comme l'effet spécifique d'un jour de type  $i$ .

Ainsi, comparons deux mois : l'un des deux a un jour de plus que l'autre, et ce jour est un lundi. L'effet de ce jour supplémentaire sur la variable  $X$  est égal à  $\bar{\alpha} + \beta_1$ , c'est-à-dire à la somme de l'effet dû à la présence d'un jour supplémentaire (effet longueur du mois) et de l'effet spécifique dû au fait que ce jour supplémentaire est un lundi.

<sup>5</sup> Il existe des modèles à coefficients variables, permettant de prendre en compte l'évolution des effets de calendrier dans le temps : par exemple des modèles où les coefficients suivent des processus ARMA, ou des modèles où les coefficients sont estimés sur des fenêtres glissantes. On se limite dans cet article aux modèles à coefficients fixes.

On a par construction  $\sum_{i=1}^7 \beta_i = 0$  donc les effets spécifiques des différents types de jours s'annulent sur la semaine.

Cette relation implique par exemple que  $\beta_7 = -\sum_{i=1}^6 \beta_i$ . En remplaçant  $\beta_7$  par cette valeur dans l'expression (1), on obtient :

$$\begin{aligned} X_t &= \sum_{i=1}^6 \beta_i N_{it} + \beta_7 N_{7t} + \bar{\alpha} N_t + \varepsilon_t \\ &= \sum_{i=1}^6 \beta_i N_{it} - \sum_{i=1}^6 \beta_i N_{7t} + \bar{\alpha} N_t + \varepsilon_t \end{aligned}$$

$$X_t = \sum_{i=1}^6 \beta_i (N_{it} - N_{7t}) + \bar{\alpha} N_t + \varepsilon_t \quad (2)$$

Le modèle (2) contient donc 7 régresseurs :

- Pour  $i=1$  à  $6$  :  $(N_{it} - N_{7t})$  : contrastes des nombres de jours de type  $i$  par rapport au nombre de dimanches.
- et  $N_t$  : longueur du mois.

Cette écriture du modèle est plus avantageuse que la formulation initiale car les estimateurs des coefficients  $\beta_i$  sont beaucoup moins corrélés que ceux des coefficients  $\alpha_i$  (Bell et Hillmer [3]). Le problème de colinéarité n'est toutefois pas complètement écarté puisque le nombre de lundis est corrélé au nombre de mardis et anti-corrélé au nombre de jeudis, mais cette colinéarité est en général négligée.

### 1.3.2. Le modèle de correction des effets de jours ouvrables

Dans de nombreuses activités économiques, il n'est pas justifié de différencier tous les jours de la semaine. Mieux vaut dans ce cas définir un modèle plus parcimonieux, dont les coefficients seront mieux estimés. Gomez et Maravall [15] ont proposé une simplification reposant sur l'hypothèse que les jours de la semaine du lundi au vendredi sont similaires et que les samedis sont similaires aux dimanches. On effectue alors simplement une correction de l'effet de « jours ouvrables », en l'occurrence l'effet des jours de la semaine par opposition aux week-ends. Cela revient à écrire les hypothèses suivantes sur les coefficients :  $\beta_1 = \beta_2 = \dots = \beta_5$  et  $\beta_6 = \beta_7$ . Ces hypothèses peuvent être vérifiées par des tests de Fisher d'égalité des coefficients.

La contrainte  $\sum_{i=1}^7 \beta_i = 0$  peut se réécrire :  $\sum_{i=1}^5 \beta_i + \sum_{i=6}^7 \beta_i = 5\beta_1 + 2\beta_6 = 0$  d'où  $\beta_6 = \beta_7 = -\frac{5}{2}\beta_1$ .

Ainsi l'expression (1) devient :

$$X_t = \beta_1 \sum_{i=1}^5 N_{it} - \frac{5}{2} \beta_1 (N_{6t} + N_{7t}) + \bar{\alpha} N_t + \varepsilon_t$$

$$X_t = \beta_1 \left( \sum_{i=1}^5 N_{it} - \frac{5}{2} (N_{6t} + N_{7t}) \right) + \bar{\alpha} N_t + \varepsilon_t \quad (3)$$

Le modèle contient alors 2 régresseurs :

- $(\# \text{lundis} + \# \text{mardis} + \# \text{mercredis} + \# \text{jeudis} + \# \text{vendredis}) - \frac{5}{2} * (\# \text{week-ends})$  : contraste semaine / weekend.
- et  $N_i$  : longueur du mois.

Ce contraste des jours de la semaine par rapport au week-end fait partie des jeux de régresseurs prédéfinis dans Tramo-Seats et X12-ARIMA.

### 1.3.3. Saisonnalité et effets de calendrier

La dissociation entre les effets de calendrier et les effets saisonniers n'est pas toujours évidente. Selon les recommandations de bonnes pratiques d'Eurostat, il est préférable d'isoler les effets de calendrier et de traiter la saisonnalité par le processus de désaisonnalisation. Il faut donc que les régresseurs définis pour la correction des effets de calendrier soient « purs », c'est-à-dire non saisonniers.

Bell [1] fournit une décomposition saisonnière des régresseurs pour effets de calendrier. Pour tout régresseur  $Z_{it}$  du modèle (2), on peut écrire :

$$Z_{it} = (Z_{it} - \bar{Z}_{iT}) + (\bar{Z}_{iT} - \bar{\bar{Z}}_i) + \bar{\bar{Z}}_i$$

où  $\bar{\bar{Z}}_i$  est la moyenne de long terme de  $Z_{it}$  tous mois confondus et  $\bar{Z}_{iT}$  la moyenne de long terme de  $Z_{it}$  calculée sur les mois (ou trimestres) correspondant à la date  $t^6$ . Notons ici que la configuration du calendrier grégorien (sans distinction des jours fériés) est périodique de période 400 ans<sup>7</sup>. La longue période est donc de préférence un multiple de 400 ans<sup>8</sup>.

Dans cette décomposition, la partie saisonnière de  $Z_{it}$  est  $(\bar{Z}_{iT} - \bar{\bar{Z}}_i)$  et le régresseur désaisonnalisé est  $Z_{it}^{cvs} = (Z_{it} - \bar{Z}_{iT}) + \bar{\bar{Z}}_i$ .

Si l'intervalle de temps est assez long (multiple de 400 ans), la composante saisonnière des  $N_{it}$  est la même pour tout  $i$ : dans cet intervalle de temps, il n'y a pas plus de lundis que de mardis en moyenne, dans un type de mois donné, comme le montre le tableau 1. Ainsi  $\bar{N}_{iT}$  est identique pour tout  $i$ , et, de même,  $\bar{\bar{N}}_i$  (moyenne sur longue période du nombre de jours de type  $i$ ) est identique pour tout  $i$ .

Tableau 1 : Nombre moyen par mois de jours de chaque type sur 400 ans

Mois	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
1	4.4300	4.4250	4.4300	4.4275	4.4300	4.4300	4.4275
2	4.0325	4.0375	4.0325	4.0375	4.0325	4.0350	4.0350
3	4.4300	4.4275	4.4300	4.4250	4.4300	4.4275	4.4300
4	4.2850	4.2875	4.2850	4.2875	4.2850	4.2850	4.2850
5	4.4300	4.4250	4.4300	4.4275	4.4300	4.4300	4.4275
6	4.2850	4.2875	4.2850	4.2850	4.2850	4.2850	4.2875
7	4.4300	4.4275	4.4300	4.4300	4.4275	4.4300	4.4250
8	4.4275	4.4300	4.4250	4.4300	4.4275	4.4300	4.4300
9	4.2875	4.2850	4.2875	4.2850	4.2850	4.2850	4.2850
10	4.4250	4.4300	4.4275	4.4300	4.4300	4.4275	4.4300
11	4.2875	4.2850	4.2850	4.2850	4.2850	4.2875	4.2850
12	4.4275	4.4300	4.4300	4.4275	4.4300	4.4250	4.4300

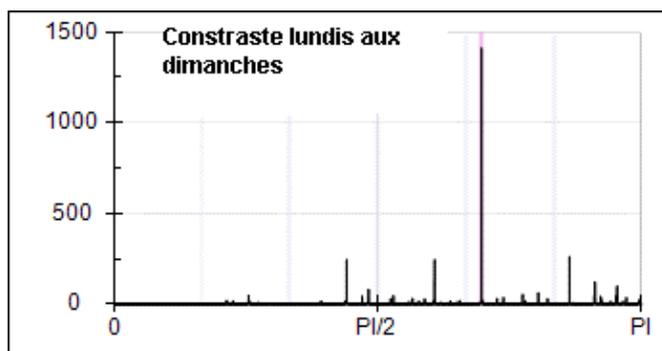
<sup>6</sup> Si la date  $t$  est le mois de mars 1996,  $\bar{Z}_{iT}$  est la moyenne des  $Z_{it}$  de tous les mois de mars de l'intervalle de long terme.

<sup>7</sup> Cette valeur provient de la périodicité hebdomadaire de 7 jours, combinée à la périodicité des années bissextiles (multiples de 4 mais pas de 100, ou multiples de 400).

<sup>8</sup> On trouve ou on construit aisément un calendrier « perpétuel » permettant ce calcul, sinon, prendre un multiple de 28 ans peut constituer une approximation correcte.

Dans le modèle (2), les régresseurs  $(N_{it} - N_{7t})$  ne sont donc pas saisonniers puisque les composantes saisonnières s'annulent par différence. Ceci peut être vérifié empiriquement sur le spectre de ces contrastes, qui ne présente pas de pic à la fréquence saisonnière, à savoir  $\frac{\pi}{6}$  pour les séries mensuelles (cf. par exemple le graphique 5). Un pic à la fréquence matérialisée par le trait rose est en revanche bien visible. Cette fréquence est l'une de celles correspondant aux effets de calendrier. Par défaut, dans Demetra+, elle vaut 2,188 gradients pour les séries mensuelles.

Graphique 5 : Contraste du nombre de lundis au nombre de dimanches



Finalement, seule la variable  $N_t$ , longueur du mois, est saisonnière, et sa valeur désaisonnalisée (CVS) est :

$$N_t^{cvs} = (N_t - \bar{N}_T) + \bar{N}$$

Pour les séries mensuelles,  $N_t - \bar{N}_T$  prend les valeurs suivantes :

$$\begin{cases} 0,7575 & \text{pour le mois de février des années bissextiles} \\ -0,2425 & \text{pour les autres mois de février} \\ 0 & \text{pour tous les autres mois} \end{cases}$$

Cette variable est appelée « **Leap Year** » et est notée  $LY_t$ .

On a ainsi :

$$N_t^{cvs} = LY_t + \bar{N}$$

$\bar{N}$  est le nombre moyen, sur longue période, de jours dans un mois (resp. un trimestre), soit  $\frac{365,2425}{12} = 30,4369$  (resp.  $\frac{365,2425}{4} = 91,3106$ ).

Dans la pratique, la constante  $\bar{N}$  n'est pas introduite dans le modèle.

Après désaisonnalisation, le modèle (2) contient alors les régresseurs suivants :

- Pour  $i=1$  à 6 :  $(N_{it} - N_{7t})$  : contrastes des nombres de jours de type  $i$  par rapport au nombre de dimanches.
- et  $LY_t$  : Leap Year.

Et le modèle (3) contient les régresseurs suivants :

- $(\# \text{ lundis} + \# \text{ mardis} + \# \text{ mercredis} + \# \text{ jeudis} + \# \text{ vendredis}) - \frac{5}{2} * (\# \text{ week-ends})$  : contraste semaine / weekend.
- et  $LY_t$  : Leap Year.

Pour les séries décomposées selon un schéma multiplicatif, il est possible, dans X12-ARIMA (voir Findley et alii [14]), de remplacer l'introduction de la variable Leap Year dans le modèle, et donc l'estimation de son coefficient, par un recalibrage des valeurs des mois de février de la série brute,

avant transformation par le logarithme, en appliquant le coefficient  $\frac{\bar{N}^{Fév}}{N_t^{Fév}}$  où :

$N_t^{Fév}$  est la longueur du mois de février à la date t.

$\bar{N}^{Fév}$  est la longueur moyenne des mois de février ( $\bar{N}^{Fév} = 28,2425$ ).

On peut trouver dans Bell et Martin [4] une justification de cette approche : l'estimation du coefficient de la variable longueur du mois conduirait dans certains cas à des valeurs aberrantes.

#### 1.4. La correction des effets de calendrier

Les régresseurs définis plus haut n'ont pas de tendance, et, si on choisit d'estimer les effets de calendrier « purs » (hors effets saisonniers), ils sont désaisonnalisés. Or la série brute, elle, a en général une tendance et une saisonnalité. On ne peut donc pas estimer les effets de calendrier par les moindres carrés ordinaires en régressant la série brute sur ces variables, le résidu n'étant pas un bruit blanc. Deux solutions peuvent être envisagées. La première consiste à estimer les effets de calendrier non pas sur la série brute mais sur la composante irrégulière issue de la décomposition en tendance, saisonnalité et irrégulier. Cette méthode est proposée dans X11, ainsi que par Maillard ([18]). Dans le cas de séries dont le schéma de composition est multiplicatif, c'est la série I-1 (où I est la composante irrégulière) qui est modélisée pour la correction des effets de calendrier. Cependant, dans cette méthode, la décomposition en tendance, saisonnalité, irrégulier de la série brute peut être perturbée par la présence des fluctuations dues aux effets de calendrier. De nombreux travaux, notamment dans le cadre de groupes de travail Eurostat au milieu des années 1990, ont montré que dans ce cas la décomposition était de moins bonne qualité que si elle s'effectuait sur une série déjà « nettoyée » des effets de calendrier.

Dans Tramo-Seats, comme dans X12-Arima, la solution consiste à utiliser un modèle Reg-ARIMA, c'est-à-dire un modèle de régression avec des résidus suivant un modèle ARIMA<sup>9</sup> :

$$X_t = \sum_i \xi_i Z_{it} + \varepsilon_t$$

Avec :

- $\varepsilon_t \rightarrow ARIMA(p, d, q)(P, D, Q)$
- les  $Z_{it}$  sont les régresseurs et les  $\xi_i$  leurs coefficients

Soit :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D \left( X_t - \sum_i \xi_i Z_{it} \right) = \theta(B)\Theta(B^s)a_t$$

Où :

<sup>9</sup> On utilise ici une formulation générale des régresseurs.

- S est la périodicité de la série (S=12 pour les séries mensuelles, S=4 pour les séries trimestrielles)
- d et D sont les ordres de différenciation simple et saisonnière du modèle ARIMA
- En notant B l'opérateur retard tel que  $BX_t = X_{t-1}$ ,

$$\begin{aligned}\phi(B) &= 1 - \varphi_1 B - \dots - \varphi_p B^p \\ \Phi(B^S) &= 1 - \Psi_1 B^S - \dots - \Psi_p (B^S)^p\end{aligned}$$

Ce sont les polynômes autorégressifs du modèle ARIMA.

$$\begin{aligned}\theta(B) &= 1 - \vartheta_1 B - \dots - \vartheta_q B^q \\ \Theta(B^S) &= 1 - \Xi_1 B^S - \dots - \Xi_Q (B^S)^Q\end{aligned}$$

Ce sont les polynômes moyennes mobiles du modèle ARIMA.

Tramo-Seats et X12-ARIMA permettent d'estimer ce type de modèles.

Dans le cas d'un schéma additif, la série corrigée des effets de calendrier, que l'on notera « CJO », est obtenue de la façon suivante :

$$X_t^{CJO} = X_t^{brute} - \sum \hat{\xi}_i Z_{it}$$

Dans le cas d'un modèle multiplicatif, on effectue préalablement une transformation de la série par le logarithme :

$$\phi(B)\Phi(B^S)(1-B)^d(1-B^S)^D \left( \text{Log}X_t - \sum_i \xi_i Z_{it} \right) = \theta(B)\Theta(B^S)a_t$$

Le choix d'une telle modélisation conduit à formuler deux remarques. La première concerne le nombre de paramètres à estimer dans ce type de modèles. En effet, dans les logiciels Tramo-Seats et X12-ARIMA, la partie « Reg » du modèle Reg-ARIMA est utilisée pour modéliser non seulement les effets de calendrier mais aussi les outliers (valeurs extrêmes, ruptures), voire encore d'autres effets définis par l'utilisateur. Si l'on ajoute les paramètres du modèle ARIMA proprement dit, on arrive parfois à devoir estimer un nombre important de paramètres. Ceci peut justifier le souci de définir, autant que possible, des jeux de régresseurs pour effets de calendrier parcimonieux.

La deuxième remarque concerne la désaisonnalisation des régresseurs. Dans le modèle Reg-ARIMA, la différenciation (facteur  $(1-B)^d(1-B^S)^D$ ) est appliquée non seulement à la série brute mais aussi à la partie « Régresseurs »  $(\sum_i \xi_i Z_{it})$ . On peut alors se demander s'il est nécessaire de désaisonnaliser les régresseurs dès lors que l'on utilise un modèle Reg-ARIMA. Par exemple, dans le cas d'une simple différenciation à l'ordre 12, on obtient :

$$\begin{aligned}(1-B^{12}) \left( \text{Log}X_t - \sum_i \xi_i Z_{it} \right) &= a_t \\ \text{Log}(X_t - X_{t-12}) &= \sum_i \xi_i (Z_{it} - Z_{i(t-12)}) + a_t\end{aligned}$$

Dans ce modèle, si les variables  $Z_{it}$  sont les variables de calendrier ( $N_{it}$ ,  $N_t$ ), les termes  $(Z_{it} - Z_{i(t-12)})$  ne sont pas saisonniers, leur saisonnalité étant stable.

Cependant, il est préférable de construire d'emblée des jeux de régresseurs non saisonniers : on peut en effet être amené à estimer des effets de jours ouvrables sur des séries non saisonnières. Même pour certaines séries saisonnières, l'identification automatique du modèle ARIMA par les logiciels Tramo-Seats ou X12-ARIMA ne conduit pas toujours à un modèle contenant une différenciation à l'ordre 12, mais seulement à des paramètres P et/ou Q saisonniers non nuls. Dans ce cas, la saisonnalité des régresseurs, qui est a priori stable, peut n'être pas correctement éliminée.

## 1.5. Introduction d'un calendrier national

### 1.5.1. Le modèle à 15 variables

L'effet d'un jour de type  $i$  sur l'activité n'est généralement pas le même si le jour est férié ou s'il ne l'est pas. On redéfinit alors le modèle initial avec 14 types de jours au lieu de 7 :

$$X_t = \sum_{i=1}^{14} \alpha_i N_{it} + \varepsilon_t$$

Où les  $N_{it}$  sont les nombres de lundis *non fériés* ( $i=1$ ), mardis *non fériés* ( $i=2$ ), ..., dimanches *non fériés* ( $i=7$ ), lundis *fériés* ( $i=8$ ), mardis *fériés* ( $i=9$ ), ..., dimanches *fériés* ( $i=14$ ) à la date  $t$ .

On reprend l'expression (1) adaptée au cas présent :

$$X_t = \sum_{i=1}^{14} \beta_i N_{it} + \bar{\alpha} N_t + \varepsilon_t \quad (3)$$

où  $\bar{\alpha} = \frac{1}{14} \sum_{i=1}^{14} \alpha_i$  et  $\beta_i = \alpha_i - \bar{\alpha}$

En effectuant les mêmes transformations que dans le modèle à 7 types de jours, avec des contrastes définis en différences par rapport au nombre de dimanches fériés, on obtient :

$$X_t = \sum_{i=1}^{13} \beta_i (N_{it} - N_{14t}) + \bar{\alpha} N_t \quad (4)$$

Dans le cas de la plupart des séries économiques produites à l'Insee, c'est le modèle à 15 variables (4) qui est a priori la base la plus pertinente pour la construction des régresseurs. Cependant, les séries étant relativement courtes, il est préférable de réduire le nombre de paramètres à estimer en imposant des contraintes sur les coefficients. En particulier, comme les jours fériés sont peu nombreux, il est nécessaire d'imposer davantage de contraintes sur les coefficients des jours fériés. On donnera donc des exemples d'hypothèses que l'on peut formuler, et on en déduira les régresseurs qui en découlent. Comme pour le modèle à 7 variables, les hypothèses doivent être vérifiées par des tests de Fisher d'égalité des coefficients.

Hypothèse 1 : Les dimanches fériés ont le même effet sur la variable  $X$  que les dimanches non fériés :  $\beta_7 = \beta_{14}$ .

$$\begin{aligned} \sum_{i=1}^{14} \beta_i &= 0 \Rightarrow \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i + \beta_7 + \beta_{14} = 0 \\ &\Rightarrow \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i + 2\beta_7 = 0 \\ &\Rightarrow \beta_7 = \beta_{14} = -\frac{1}{2} \left( \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i \right) \end{aligned}$$

Dans l'expression (2), on remplace  $\beta_7$  et  $\beta_{14}$  par leur valeur :

$$\begin{aligned}
X_t &= \sum_{i=1}^6 \beta_i N_{it} + \sum_{i=8}^{13} \beta_i N_{it} - \frac{1}{2} \left( \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i \right) N_{7t} - \frac{1}{2} \left( \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i \right) N_{14t} + \bar{\alpha} N_t + \varepsilon_t \\
&= \sum_{i=1}^6 \beta_i N_{it} + \sum_{i=8}^{13} \beta_i N_{it} - \frac{1}{2} \left( \sum_{i=1}^6 \beta_i + \sum_{i=8}^{13} \beta_i \right) (N_{7t} + N_{14t}) + \bar{\alpha} N_t + \varepsilon_t \\
&= \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{2} (N_{7t} + N_{14t}) \right) + \sum_{i=8}^{13} \beta_i \left( N_{it} - \frac{1}{2} (N_{7t} + N_{14t}) \right) + \bar{\alpha} N_t + \varepsilon_t
\end{aligned}$$

$$\boxed{X_t = \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{2} D_t \right) + \sum_{i=8}^{13} \beta_i \left( N_{it} - \frac{1}{2} D_t \right) + \bar{\alpha} N_t + \varepsilon_t}$$

où  $D_t$  est le nombre de dimanches (fériés ou non fériés) à la date  $t$ .

On propose en annexe d'autres hypothèses a priori pertinentes pour les séries économiques et les régresseurs associés. On ne développera donc dans ce paragraphe qu'un autre exemple d'hypothèse.

Hypothèse 2 : Les dimanches non fériés et tous les jours fériés ont le même effet sur  $X$  :  $\beta_7 = \beta_8 = \beta_9 = \dots = \beta_{13} = \beta_{14}$

$$\sum_{i=1}^{14} \beta_i = 0 \Rightarrow \sum_{i=1}^6 \beta_i + 8\beta_7 = 0$$

$$\Rightarrow \beta_7 = \dots = \beta_{14} = -\frac{1}{8} \sum_{i=1}^6 \beta_i$$

$$\begin{aligned}
X_t &= \sum_{i=1}^{14} \beta_i N_{it} + \bar{\alpha} N_t + \varepsilon_t \\
&= \sum_{i=1}^6 \beta_i N_{it} + \sum_{i=7}^{14} \left( -\frac{1}{8} \sum_{i=1}^6 \beta_i \right) N_{it} + \bar{\alpha} N_t + \varepsilon_t
\end{aligned}$$

$$\boxed{X_t = \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{8} \sum_{i=7}^{14} N_{it} \right) + \bar{\alpha} N_t + \varepsilon_t}$$

$\sum_{i=7}^{14} N_{it}$  est la somme du nombre de dimanches et du nombre de jours fériés du lundi au samedi.

Dans Tramo, les régresseurs correspondant à cette hypothèse sont formulés de la façon suivante :

$\left( N_{it} - \sum_{i=7}^{14} N_{it} \right)$ , c'est-à-dire de la même façon mais sans le coefficient 1/8. Cette écriture conduit à

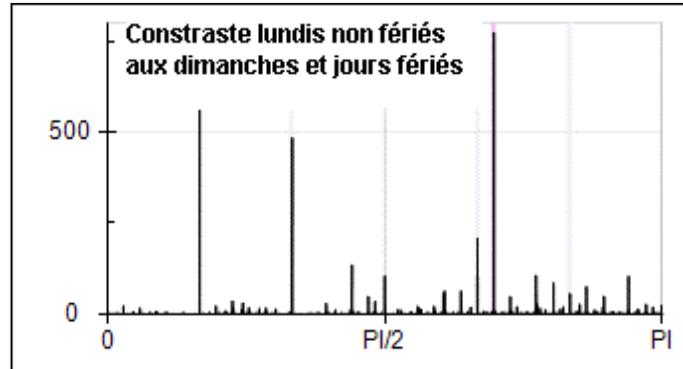
des valeurs différentes des coefficients  $\beta_i$  et  $\bar{\alpha}$ , mais le jeu de régresseurs constitué par ces variables et la variable  $N_t$  engendre le même espace vectoriel que le jeu de régresseurs du modèle ci-dessus.

### 1.5.2. Saisonnalité et effets de calendrier

On a vu dans le paragraphe 1.3.3 que dans le modèle (2) à sept variables, seule la variable  $N_t$  était saisonnière. En revanche dans les modèles qui distinguent les jours fériés des jours non fériés (modèles dérivés du modèle à 14 variables), il subsiste dans les contrastes de type  $(N_{it} - N_{14t})$  une saisonnalité due au fait que certains jours fériés tombent toujours des lundis (Pâques, Pentecôte) ou des jeudis (Ascension) : en moyenne, il y a moins de lundis non fériés en mai qu'en janvier. Cette saisonnalité peut être mise évidence par des pics à la fréquence saisonnière dans le spectre des

régresseurs, par exemple le spectre du nombre de lundis non fériés en contraste au nombre de dimanches et jours fériés (cf. Graphique 6). Il faut donc ôter la composante saisonnière des régresseurs en remplaçant chaque régresseur  $Z_t$  par  $Z_t - \bar{Z}_T + \bar{Z}$ .

Graphique 6 : Contraste du nombre de lundis non fériés au nombre de dimanches et jours fériés



Contrairement au cas du modèle (2), la structure du calendrier n'est plus nécessairement périodique de période 400 ans, du fait de la présence des jours fériés. Du coup, le choix de l'intervalle de long terme n'est pas neutre. C'est une des causes de divergence des régresseurs construits par les différents services producteurs. En théorie, l'intervalle de long terme doit correspondre à la période du calendrier en prenant en compte les jours fériés. Dans le cas où tous les jours fériés sont fixes, cette période est la même que celle du calendrier sans prise en compte des jours fériés, à savoir 400 ans. Si l'on prend en compte les jours fériés mobiles, elle peut être très élevée : 5 700 000 ans pour les jours fériés liés à la fête de Pâques (Pâques, Ascension et Pentecôte) ! Dans la pratique, on se limite à 400 ans mais des travaux en cours au département des statistiques de court terme de l'Insee permettront un calcul des « vraies » moyennes de long terme, c'est-à-dire calculées sur la période complète.

En réalité, le remplacement de  $Z_t$  par  $Z_t - \bar{Z}_T + \bar{Z}$  n'ôte complètement la composante saisonnière que si les jours fériés ne sont pas modifiés au cours de la période d'observation. Si un jour férié n'existe qu'à partir d'une certaine date, le nombre de jours fériés du mois concerné, en écart à sa moyenne de long terme, sera positif chaque année après cette date et négatif chaque année avant cette date. Une solution consiste à scinder l'intervalle de long terme en deux (avant et après cette date), mais cela devient très complexe si plusieurs jours fériés sont concernés, ou si, comme dans certains pays, les jours fériés changent souvent.

### 1.6. Généralisation à une partition quelconque des jours du mois ou du trimestre

On peut généraliser le modèle de base en écrivant :

$$X_t = \sum_{i=1}^p \alpha_i N_{it} + \varepsilon_t$$

où les  $N_{it}$  ( $i=1, \dots, p$ ) forment une partition quelconque du nombre de jours du mois.

En posant  $\bar{\alpha} = \frac{1}{p} \sum_{i=1}^p \alpha_i$  et  $\beta_i = \alpha_i - \bar{\alpha}$ , on obtient

$$X_t = \sum_{i=1}^p \beta_i N_{it} + \bar{\alpha} N_t + \varepsilon_t \quad (3)$$

Pour j quelconque, on a  $\beta_j = -\sum_{\substack{i=1 \\ i \neq j}}^p \beta_i$ .

Ce qui donne :

$$X_t = \sum_{\substack{i=1 \\ i \neq j}}^p \beta_i (N_{it} - N_{jt}) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{Mp})$$

$X_{jt}$  est ici appelée variable de *contraste*.

Comme précédemment, on peut émettre des hypothèses d'égalité des coefficients pour simplifier le modèle. On distinguera deux types d'hypothèses :

- hypothèses d'égalité des coefficients de variables de contraste
- hypothèses d'égalité des coefficients d'autres variables

Soit l'ensemble C des indices des variables de contraste :

$$C = \{i_1, i_2, \dots, i_c\} \subset \{1, 2, \dots, p\} \text{ tel que } \beta_{i_1} = \beta_{i_2} = \dots = \beta_{i_c} \quad (\text{Card } C = c).$$

Soient par ailleurs les ensembles  $E_1, E_2, \dots, E_m$  d'indices de variables autres que des variables de contrastes à regrouper :

$$E_1 = \{j_1, j_2, \dots, j_{e_1}\} \subset \{1, 2, \dots, p\} \setminus C \text{ tel que } \beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_{e_1}} \quad (\text{Card } E_1 = e_1).$$

...

$$E_m = \{k_1, k_2, \dots, k_{e_m}\} \subset \{1, 2, \dots, p\} \setminus \{C \cup E_1 \cup E_2 \cup \dots \cup E_{m-1}\} \text{ tel que } \beta_{k_1} = \beta_{k_2} = \dots = \beta_{k_{e_m}}$$

(Card  $E_m = e_m$ )

$$\sum_{i=1}^p \beta_i = 0 \Rightarrow \sum_{i \in E_1} \beta_i + \dots + \sum_{i \in E_m} \beta_i + \sum_{i \in C} \beta_i + \sum_{i \notin \{C \cup E_1 \cup \dots \cup E_m\}} \beta_i = 0$$

$$\Rightarrow e_1 \beta_{j_1} + \dots + e_m \beta_{k_1} + c \beta_{i_1} + \sum_{i \notin \{C \cup E_1 \cup \dots \cup E_m\}} \beta_i = 0$$

$$\text{puis } \beta_{i_1} = \dots = \beta_{i_c} = -\frac{1}{c} \left( e_1 \beta_{j_1} + \dots + e_m \beta_{k_1} + \sum_{i \notin \{C \cup E_1 \cup \dots \cup E_m\}} \beta_i \right)$$

En intégrant ces contraintes dans l'expression (3) on obtient :

$$X_t = \beta_{j_1} \left( \sum_{i \in E_1} N_{it} - \frac{e_1}{c} \sum_{i \in C} N_{it} \right) + \dots + \beta_{k_1} \left( \sum_{i \in E_m} N_{it} - \frac{e_m}{c} \sum_{i \in C} N_{it} \right) + \sum_{i \notin \{C \cup E_1 \cup \dots \cup E_m\}} \beta_i \left( X_{it} - \frac{1}{c} \sum_{i \in C} N_{it} \right) + \bar{\alpha} N_t + \varepsilon$$

La désaisonnalisation des régresseurs dans ce cas général doit être effectuée.

## 1.7. Compléments

Dans la pratique, quelques compléments aux modèles de base sont parfois apportés pour affiner la correction des effets de calendrier, comme par exemple pour le traitement des séries des comptes nationaux trimestriels.

Le premier consiste à tester l'effet retardé ou avancé de la composition du calendrier : est-ce qu'il existe des effets de rattrapage ou d'anticipation des effets de jours ouvrables ? Ainsi, dans le modèle, sont introduits les régresseurs retardés ou avancés d'un mois (ou trimestre).

L'autre consiste à différencier les effets de calendrier selon les périodes de l'année : l'effet d'un lundi est-il le même en été ou en hiver ? D'une façon plus générale, cela revient à tester l'hypothèse selon laquelle les coefficients des régresseurs sont saisonniers. Les comptes trimestriels testent si les effets de calendrier des mois de juillet et août sont différents de ceux des autres mois<sup>10</sup>. Sachant que dans le commerce de détail les magasins sont généralement ouverts les dimanches de décembre, on pourrait estimer séparément les effets de calendrier du mois de décembre des autres mois.

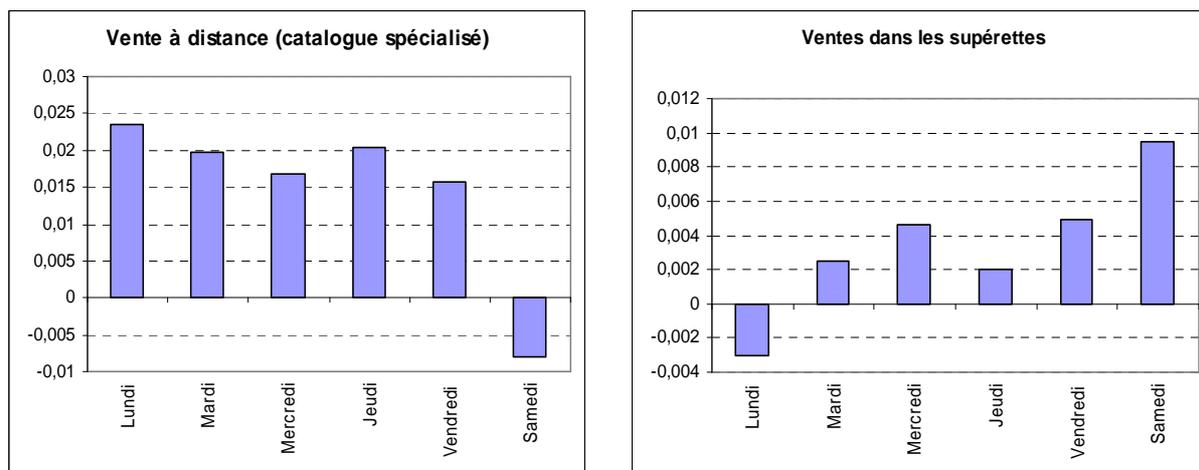
## 2. Comment choisir le meilleur jeu de régresseurs pour corriger les effets de calendrier ?

On a vu précédemment une formulation générale pour construire des jeux de régresseurs. Dans cette partie, on propose une méthode pour définir des jeux de régresseurs pertinents compte tenu du secteur d'activité sous revue et pour sélectionner le jeu de régresseurs le mieux adapté à chaque série parmi différents jeux de régresseurs.

### 2.1. Prendre en compte le secteur d'activité sous revue

Les effets des différents jours de la semaine peuvent être très variables selon les secteurs d'activité. Le graphique ci-dessous (graphique 7) représente les coefficients des jours de la semaine du lundi au samedi pour les indices mensuels de chiffres d'affaires dans deux secteurs du commerce de détail en France : la vente à distance sur catalogue spécialisé et les supérettes.

*Graphique 7* : Coefficients des régresseurs pour effets de calendrier des indices de chiffre d'affaires de la vente à distance sur catalogue spécialisé et des supérettes



Pour les supérettes, le coefficient du lundi est négatif, car en France de nombreux magasins sont fermés le lundi. Au contraire, le coefficient du samedi est élevé, l'activité étant plus intense ce jour-là. Les autres coefficients sont positifs, mais plus faibles. Pour la vente à distance, les coefficients du lundi au vendredi sont proches les uns des autres, tandis que le coefficient du samedi est négatif.

Dans les principaux logiciels de désaisonnalisation, Tramo-Seats et X12-ARIMA, des jeux de régresseurs prédéfinis sont proposés aux utilisateurs. Deux cas sont considérés :

- Le premier considère que chaque jour de la semaine a un effet distinct, le jeu de régresseurs comporte six variables correspondant aux lundis, mardis, ..., samedis.

<sup>10</sup> En fait, ils corrigent ainsi un artefact créé par la méthode d'interpolation du mois d'août dans les anciennes séries conjoncturelles (avant 2000, il n'y avait pas d'enquête de branche de l'indice de production industrielle en août).

- Le deuxième est fondé sur l'hypothèse que les jours de la semaine, du lundi au vendredi, sont similaires, alors que le samedi est considéré comme similaire au dimanche, le jeu de régresseurs est donc composé du contraste des jours de la semaine par rapport au week-end.

La vente à distance correspond au deuxième cas. En revanche, les supérettes ne correspondent à aucun des deux cas, puisque les coefficients du milieu de la semaine sont très proches. Il est donc utile de construire des jeux de régresseurs qui permettent de considérer des situations plus variées que les deux cas considérés par défaut dans les logiciels.

Cependant, si l'on part du modèle à 15 variables défini au paragraphe 1.5.1, on peut définir un très grand nombre de jeux de régresseurs, correspondant chacun à des hypothèses de similarité entre différents jours. Même si l'on construisait l'ensemble des ces jeux de régresseurs, on serait alors confrontés au problème compliqué de la sélection du meilleur jeu de régresseurs pour chaque série. Il convient donc de simplifier ce problème en définissant un nombre restreint de jeux de régresseurs pertinents et adaptés au secteur d'activité.

## 2.2. Définir quelques jeux de régresseurs pertinents

Dans ce paragraphe, on décrit la méthode qui a été mise au point pour définir les jeux de régresseurs pour la correction des effets de calendrier des séries produites au département des statistiques de court terme de l'Insee : indices de chiffres d'affaires, indice de la production industrielle, indices de commandes.

Pour ces séries, le modèle de base sur lequel on s'est appuyé est le modèle à 15 variables (modèle (3)). Une première simplification du modèle a été effectuée d'emblée, en s'appuyant sur des hypothèses apparaissant a priori comme pertinentes :

- les dimanches fériés ont été assimilés aux dimanches non fériés
- les jours fériés (autres que les dimanches) ont été considérés comme similaires entre eux.
- les jours fériés ont été assimilés aux dimanches.

Cette troisième hypothèse n'a en fait pas été formulée immédiatement. En effet, en France, de nombreux magasins sont ouverts les jours fériés (excepté le 1<sup>er</sup> mai, le 25 décembre et le 1<sup>er</sup> janvier), alors qu'ils sont fermés le dimanche. Cependant, les modèles distinguant les jours fériés des dimanches se sont révélés instables : des problèmes de colinéarité ont été décelés par l'observation des valeurs propres des jeux de régresseurs. La variable de contraste a donc été définie a priori comme le nombre de dimanches et de jours fériés.

Le premier jeu de régresseurs, noté  $S_1$ , comporte sept variables : nombre de lundis non fériés, nombre de mardis non fériés, ..., nombre de samedis non fériés, en contraste au nombre de dimanches et jours fériés ; Leap Year.

Le deuxième jeu de régresseurs retenu, noté  $S_2$ , est celui correspondant à la modélisation des effets de jours ouvrables sans distinction des types de jours : il comprend un régresseur défini comme le contraste entre le nombre de lundis, mardis, ..., vendredis non fériés et le nombre de dimanches et jours fériés, ainsi que la variable Leap Year.

Ces deux premiers jeux correspondent aux deux cas classiques modélisés avec les jeux de régresseurs prédéfinis dans les logiciels.

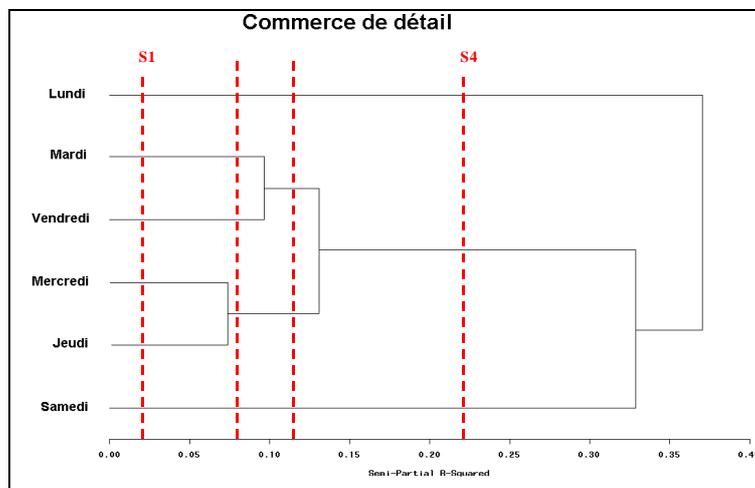
Un troisième jeu de régresseurs,  $S_3$ , a été défini en supposant que, dans certaines activités économiques, les jours de la semaine (du lundi au vendredi) sont différents les uns des autres mais que le samedi peut être assimilé au dimanche et jours fériés.

L'idée a ensuite été de s'appuyer sur un outil d'analyse des données, la classification ascendante hiérarchique (CAH), pour mettre en évidence les similarités entre les jours les plus fréquemment observés dans les différents secteurs d'activité. Les variables du jeu de régresseurs  $S_1$  ont été importées dans le logiciel Demetra+ en tant que « user-variables » (variables « utilisateur ») et les coefficients des variables ont été estimés par la méthode X12-ARIMA pour toutes les séries d'indices de chiffres d'affaires.

Une CAH a alors été effectuée pour chaque grand secteur d'activité (commerce de détail, commerce de gros, services, industrie). Le graphique suivant (graphique 8) représente l'arbre de classification obtenu dans le commerce de détail (57 séries).

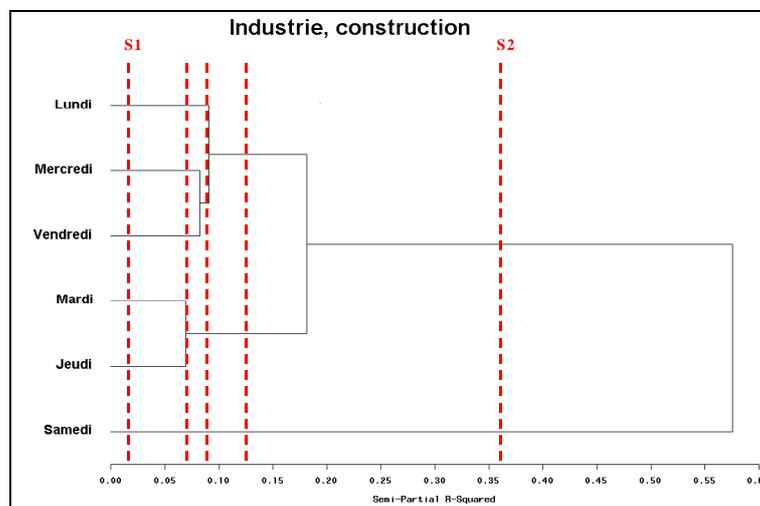
L'arbre fait apparaître une proximité entre les coefficients des mardis et des vendredis, des mercredis et des jeudis, puis entre ces quatre types de jours alors qu'au contraire le samedi et le lundi se distinguent nettement. En coupant l'arbre à différents niveaux (voir les lignes rouges verticales en pointillés), on peut définir plusieurs jeux de régresseurs. Si l'on coupe au niveau des « feuilles » (première ligne en partant de la gauche), on retrouve le jeu de régresseurs S1 qui distingue tous les jours de la semaine du lundi au samedi. Si l'on utilise la quatrième ligne en partant de la gauche, on définit un jeu de trois régresseurs, dans lequel les jours du mardi au vendredi sont regroupés, et le lundi et le samedi sont distingués, ce qui correspond au cas des supérettes, fréquent dans le commerce de détail.

Graphique 8 : Arbre de classification des coefficients de régresseurs pour jours ouvrables des séries d'indices de chiffres d'affaires dans le commerce de détail.



Le graphique 9 montre l'arbre de classification obtenu avec les indices de chiffres d'affaires dans le secteur de l'industrie et de la construction (122 séries). Celui-ci met en évidence une similarité entre les jours de la semaine du lundi au vendredi. Le régresseur S2 semble donc bien adapté à ce secteur d'activité.

Graphique 9 : Arbre de classification des coefficients de régresseurs pour jours ouvrables des séries d'indices de chiffres d'affaires dans l'industrie et la construction.



A l'issue de cette exploration, nous avons donc retenu cinq jeux de régresseurs définis ci-dessous :

- **S0**: Leap Year seul (hypothèse que tous les jours de la semaine sont similaires)
- **S1** : 7 variables : #Lundis non fériés, #Mardis non fériés, #Mercredis non fériés, #Jeudis non fériés, #Vendredis non fériés, #Samedis non fériés ; Contraste : (#Dimanches + #Jours fériés) ; Leap Year
- **S2** : 2 variables : (#Lundis non fériés + #Mardis non fériés + #Mercredis non fériés + #Jeudis non fériés + #Vendredis non fériés) ; Contraste : (#Samedis + #Dimanches + #Jours fériés) ; Leap Year
- **S3** : 6 variables : #Lundis non fériés, #Mardis non fériés, #Mercredis non fériés, #Jeudis non fériés, #Vendredis non fériés ; Contraste : (#Samedis + #Dimanches + #Jours fériés) ; Leap Year
- **S4** : 4 variables : #Lundis non fériés, (#Mardis non fériés + #Mercredis non fériés + #Jeudis non fériés + #Vendredis non fériés) ; #Samedis non fériés ; Contraste : (#Dimanches + #Jours fériés) ; Leap Year

Rappelons que la variable Leap Year est la longueur du mois désaisonnalisée. Le régresseur S0 correspond donc à la modélisation d'un effet longueur du mois seul, sans effet spécifique du type de jour de la semaine. Il a paru intéressant de retenir dans un premier temps ce jeu de régresseurs qui pourrait être adapté à des activités de type production en continu. Le jeu S4 n'a été retenu que pour les séries du secteur tertiaire.

La question qui se pose alors est de savoir comment sélectionner le jeu de régresseurs le mieux adapté à chaque série.

### 2.3. Algorithme de sélection du jeu de régresseurs pour chaque série

#### 2.3.1. Description de l'algorithme

Pour chaque série, un modèle Reg-ARIMA a été estimé avec X12-ARIMA, en utilisant le jeu de régresseurs S1 (jeu le plus complet parmi ceux retenus). Le modèle ARIMA, ainsi que les outliers (points extrêmes, ruptures de niveaux, changements transitoires de niveau) identifiés automatiquement par X12-ARIMA (avec une valeur critique fixée à 5 pour la détection des outliers) ont ensuite été utilisés pour estimer les coefficients des variables de chacun des autres jeux de régresseurs. Puis la sélection parmi les différents jeux de régresseurs s'est faite selon un algorithme qui combine deux critères :

- **Un test de Fisher** de contraintes sur les coefficients : si l'hypothèse sur l'égalité de plusieurs coefficients n'est pas rejetée, il est possible de regrouper les jours correspondants (par exemple le lundi avec le mardi, le mercredi, le jeudi et le vendredi)
- **Le critère d'Akaike (AICC)** qui fournit de l'information sur la qualité d'ajustement du modèle Reg-ARIMA pour chaque jeu de régresseurs.

La statistique de Fisher a été calculée comme pour le test de saisonnalité fixe dans X12-ARIMA (voir Lytras et Feldpausch [17]) :

$$F = \frac{\hat{\chi}^2}{p} \times \frac{n-d-k}{n-d}$$

où :

$$\hat{\chi}^2 = (R\hat{\beta} - c)' [RVAR(\hat{\beta})R']^{-1} (R\hat{\beta} - c)$$

$\beta$  = vecteur des coefficients des régresseurs du modèle Reg-ARIMA

R = matrice de contraintes sur les coefficients

c = vecteur défini de telle sorte que la contrainte sur les coefficients s'écrive :  $R\beta - c = 0$

n = nombre d'observations dans la série brute

d = degré de différenciation dans le modèle Reg-ARIMA

p = nombre de régresseurs CJO

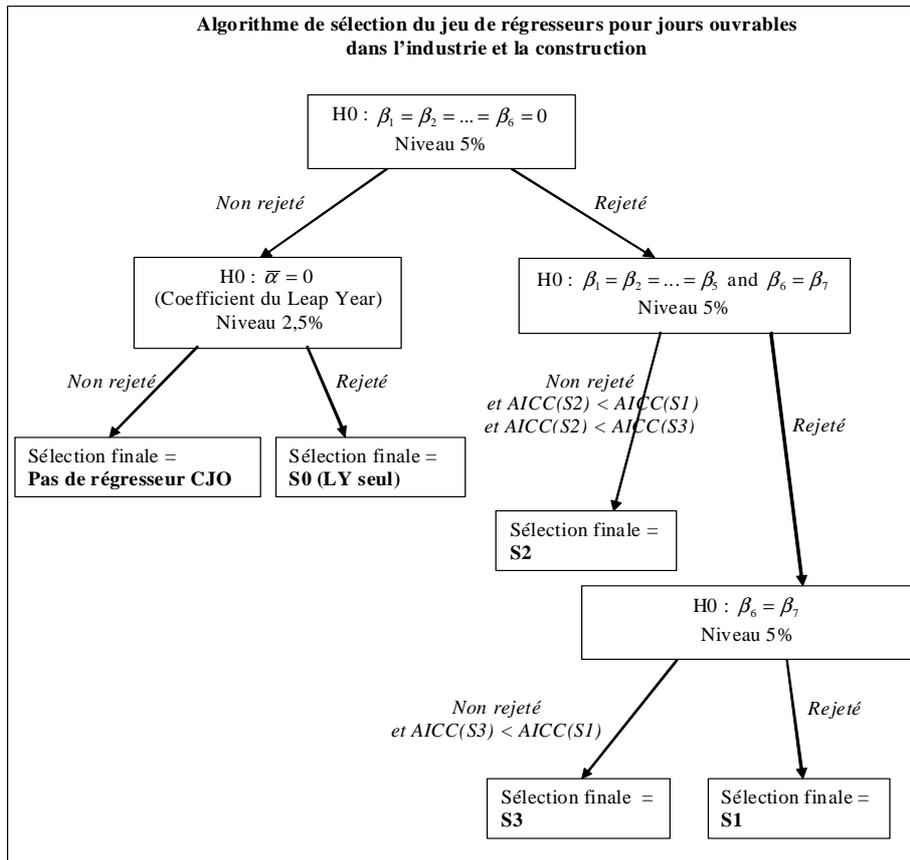
k = nombre total d'éléments dans le vecteur  $\beta$

k peut être différent de p si d'autres régresseurs sont présents dans le modèle Reg-ARIMA (outliers par exemple).

La version Dos de X12-ARIMA a été utilisée pour récupérer la matrice de variance-covariance des coefficients, car elle n'est pas fournie dans la version actuelle de Demetra+.

Le graphique 10 décrit l'algorithme complet de sélection, pour chaque série de l'industrie et de la construction, du meilleur jeu de régresseurs parmi S0, S1, S2, S3.

Graphique 10 : Algorithme de sélection, pour chaque série de l'industrie et de la construction, du meilleur jeu de régresseurs pour la correction des effets de calendrier



La robustesse de cet algorithme a été vérifiée sur une centaine de séries : les jeux de régresseurs obtenus en l'appliquant sur les séries complètes ont été comparés à ceux sélectionnés en tronquant les séries un mois auparavant, et en les tronquant un an plus tôt. Les résultats ne différaient que pour quelques séries. Par ailleurs, les résultats de la sélection ont été soumis aux producteurs d'indices, pour un examen de leur pertinence d'un point de vue économique.

### 2.3.2. Application aux indices de chiffres d'affaires

Les algorithmes de sélection ont été appliqués pour chacune des séries d'indices de chiffre d'affaires : celui du graphique 10 pour l'industrie et la construction, et un algorithme similaire, mais incluant le jeu S4, pour les autres secteurs d'activité. Dans le tableau 2 sont présentés les résultats de ce processus de sélection.

Tableau 2 : Répartition (en %) des jeux de régresseurs sélectionnés dans chaque secteur d'activité

	Pas d'effet de calendrier	S0 (LY seul)	S1	S2	S3	S4
Commerce de détail (57 séries)	5	2	21	4	7	61
Commerce de gros (89 séries)	17	4	8	27	16	28
Services (138 séries)	26	1	7	32	14	19
Industrie - construction (122 séries)	15	0	14	47	25	Non testé

On remarque que dans le commerce de détail, pour une large majorité de séries (61%), le jeu de régresseurs S4 a été sélectionné comme étant le meilleur, ce qui confirme les résultats de la classification. Dans le commerce de gros, S4 est également le jeu de régresseurs le plus fréquemment sélectionné, pratiquement à égalité avec le jeu S2. Dans les services et dans l'industrie-construction, c'est le régresseur S2 qui est le plus souvent sélectionné. A l'exception du commerce de détail, le pourcentage de séries sans effets de calendrier est important. En revanche, le jeu S0 (Leap Year seul) n'a été sélectionné pour pratiquement aucune série.

Finalement, on peut se demander si la correction des effets de jours ouvrables avec ces nouveaux jeux de régresseurs est performante. Pour ce contrôle de qualité, on a examiné le diagnostic de Demetra+ sur la présence ou non de pics à la fréquence des jours ouvrables dans la série CVS-CJO (corrigée des variations saisonnières et des effets de calendrier) et dans la composante irrégulière. Pour chaque jeu de régresseurs, le pourcentage de séries présentant un mauvais résultat pour ce diagnostic (« Bad » ou « Severe ») a été calculé. On a également calculé ce pourcentage dans le cas où chaque série avait été traitée avec le jeu sélectionné par l'algorithme défini plus haut. Les résultats sont présentés dans le tableau 3.

Tableau 3 : Pourcentage de séries d'indices de chiffres d'affaires présentant des pics spectraux résiduels à la fréquence des jours ouvrables.

	Pas d'effets de calendrier	S0	S1	S2	S3	S4	Jeu sélectionné
Commerce de détail	61%	67%	9%	39%	14%	16%	12%
Commerce de gros	71%	70%	11%	28%	19%	16%	19%
Services	62%	62%	8%	21%	13%	9%	14%
Industrie, construction	74%	75%	9%	17%	9%		13%

Comme on pouvait s'y attendre, la proportion de séries présentant des effets de calendrier résiduels est plus faible avec le jeu de régresseurs le plus complet (S1), et plus forte avec le jeu de régresseurs le plus parcimonieux (S2). Mais la proportion obtenue avec les jeux de régresseurs sélectionnés selon notre algorithme n'est pas beaucoup plus élevée que celle obtenue avec S1, notamment dans le commerce de détail, alors que les jeux sélectionnés sont le plus souvent des jeux plus parcimonieux.

## **Conclusion**

Dans cet article, nous avons décrit une méthode de construction de jeux de régresseurs pour effets de calendrier, tenant compte des spécificités du calendrier (jours fériés) et du secteur d'activité. Une formulation générale des régresseurs permet de construire des jeux variés de régresseurs mais il est conseillé de se limiter à quelques uns, les plus pertinents et les plus stables. Nous avons également proposé un algorithme permettant de sélectionner le meilleur jeu de régresseurs pour chaque série, en prenant en compte le critère de parcimonie et la qualité d'ajustement du modèle Reg-ARIMA. Une application sur les indices de chiffres d'affaires français a conduit à des résultats satisfaisants.

**Annexe : Hypothèses sur les effets des types des jours dans le modèle à 15 variables et régresseurs correspondants**

On fournit dans cette annexe les régresseurs obtenus avec quelques hypothèses a priori pertinentes pour les séries économiques.

Hypothèse 1 : Les dimanches fériés ont le même effet sur l'activité que les dimanches non fériés :

$$\beta_7 = \beta_{14}.$$

$$X_t = \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{2} D_t \right) + \sum_{i=8}^{13} \beta_i \left( N_{it} - \frac{1}{2} D_t \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14 H1})$$

où  $D_t$  est le nombre de dimanche (fériés ou non fériés) à la date t.

Hypothèse 2 : Les dimanches non fériés et tous les jours fériés ont le même effet sur l'activité :  $\beta_7 = \beta_8 = \beta_9 = \dots = \beta_{13} = \beta_{14}$

$$X_t = \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{8} \sum_{i=7}^{14} N_{it} \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H2})$$

$\sum_{i=7}^{14} N_{it}$  est la somme du nombre de dimanches et du nombre de jours fériés du lundi au samedi.

Hypothèse 3 : Les dimanches fériés ont le même effet sur l'activité que les dimanches non fériés :  $\beta_7 = \beta_{14}$  et les effets des lundis fériés, mardis fériés, ..., samedis fériés sont les mêmes :

$$\beta_8 = \beta_9 = \dots = \beta_{13}$$

On reprend le modèle (M14 H1) et on applique la contrainte sur les coefficients des jours fériés :

$$X_t = \sum_{i=1}^6 \beta_i \left( N_{it} - \frac{1}{2} D_t \right) + \beta_8 \left( \sum_{i=8}^{13} N_{it} - 3D_t \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H3})$$

$\sum_{i=8}^{13} N_{it}$  est le nombre de jours fériés du lundi au samedi.

Hypothèse 4 : Les dimanches fériés ont le même effet sur l'activité que les dimanches non fériés :  $\beta_7 = \beta_{14}$ , les effets des lundis fériés, mardis fériés, ..., samedis fériés sont les mêmes :

$$\beta_8 = \beta_9 = \dots = \beta_{13} \text{ et tous les jours de la semaine non fériés ont le même effet : } \beta_1 = \beta_2 = \dots = \beta_5$$

On reprend le modèle (M14H2) et on applique la contrainte sur les coefficients des jours de la semaine.

$$X_t = \beta_1 \left( \sum_{i=1}^5 N_{it} - \frac{5}{2} D_t \right) + \beta_6 \left( N_{6t} - \frac{1}{2} D_t \right) + \beta_8 \left( \sum_{i=8}^{13} N_{it} - 3D_t \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H4})$$

$\sum_{i=1}^5 N_{it}$  est le nombre de jours de la semaine (du lundi au vendredi) non fériés.

Hypothèse 5 : Les dimanches non fériés et tous les jours fériés ont le même effet l'activité :  $\beta_7 = \beta_8 = \beta_9 = \dots = \beta_{13} = \beta_{14}$  et tous les jours de la semaine non fériés ont le même effet :  $\beta_1 = \beta_2 = \dots = \beta_5$ .

On reprend le modèle (M14H4) et on applique la contrainte sur les jours de la semaine :

$$X_t = \beta_1 \left( \sum_{i=1}^5 N_{it} - \frac{5}{8} \sum_{i=7}^{14} N_{it} \right) + \beta_6 \left( N_{6t} - \frac{1}{8} \sum_{i=7}^{14} N_{it} \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H5})$$

Hypothèse 6 : Les samedis et les dimanches non fériés ont le même effet :  $\beta_6 = \beta_7$  , tous les jours de la semaine non fériés ont le même effet :  $\beta_1 = \beta_2 = \dots = \beta_5$  et tous les jours fériés ont le même effet :  $\beta_8 = \beta_9 = \dots = \beta_{14}$  .

$$X_t = \beta_1 \left( \sum_{i=1}^5 N_{it} - \frac{5}{2} (N_{6t} + N_{7t}) \right) + \beta_8 \left( \sum_{i=8}^{14} N_{it} - \frac{7}{2} (N_{6t} + N_{7t}) \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H6})$$

Hypothèse 7 : Les samedis et les dimanches non fériés et tous les jours fériés ont le même effet :  $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \dots = \beta_{14}$  , et tous les jours de la semaine non fériés ont le même effet :  $\beta_1 = \beta_2 = \dots = \beta_5$

$$X_t = \beta_1 \left( \sum_{i=1}^5 N_{it} - \frac{5}{9} \sum_{i=6}^{14} N_{it} \right) + \bar{\alpha} N_t + \varepsilon_t \quad (\text{M14H7})$$

$\sum_{i=6}^{14} N_{it}$  est la somme des samedis (fériés et non fériés), des dimanches (fériés et non fériés) et de tous les jours fériés du lundi au vendredi.

## Bibliographie

- [1] Bell W. R., « Seasonal Decomposition of Deterministic Effects », *Research Report*, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., RR84/01, 1984.
- [2] Bell W. R., « Correction to 'Seasonal Decomposition of Deterministic Effects' » (n° RR84/01), *Research Report*, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., RR95/01, 1995.
- [3] Bell W. R., Hillmer, S. C., « Modeling Time series with Calendar Variation », *Journal of the American Statistical Association*, 383, 78, 526-534, 1983.
- [4] Bell W. R., Martin D. E. K., « Modeling Time Varying Trading-Day Effects in Monthly time Series », *ASA proceedings*, November 2004.
- [5] Bessa M., Dhifalli R., Ladiray, D., Lassoued A., Maghrabi B., « Les effets de calendrier dans les séries tunisiennes », *STATECON* N°103, 2009.
- [6] Chen B-C., Findley D. F., « Multiplicative trading day adjustments: X-11 and RegARIMA compared », *American Statistical Association Annual Meetings*, Alexandria, 1993.
- [7] Cholette P., « A Note on the Treatment of Februaries in Trading-Day Adjustment », *Working Paper*, Time Series Research and Analysis Center, Statistics Canada, 2003.
- [8] Cleveland W. P., « Calendar Adjustment of Time Series », *Discussion Paper*, Special Studies Section, Board of Governors of the Federal Reserve System, 198, 1987
- [9] Cleveland W. P., Grupe M. R., « Modeling time series when calendar effects are present », *Applied Time Series Analysis of Economic Data*, Zellner A. (editor), U.S. Department of Commerce, U.S. Bureau of the Census, Washington D.C., pp 57-67, 1983.
- [10] Dagum E. B., « The X11-ARIMA Seasonal Adjustment Method », *Catalogue* N° 12- 64E, Statistics Canada, 1980
- [11] Dagum E. B., Quenneville B., Sutradhar B., « Trading-day variations multiple regression model with random parameters », *International Statistical Review*, 1, 60, 57-73, 1992.
- [12] Fournier J-M., Ladiray D., « Les effets de calendrier dans l'analyse conjoncturelle de la production », *Economie et Statistique*, 285-286, 115-126, 1995.
- [13] Findley D. F., Soukup R. J., « Detection and Modeling of Trading Day Effects », *Proceedings of the International Conference on Establishment Surveys*, 2000.
- [14] Findley D. F., Monsell B. C., Bell, Mark C. Otto, and Bor-Chung Chen, « New Capabilities and Methods of the X12-ARIMA Seasonal Adjutment Program », Preprint version of *Journal of Business and Economic Statistics*, Vol. 16, 1998
- [15] Gómez V., Maravall A., « Programs TRAMO (Time series Regression with Arima noise, Missing observations, and Outliers) and SEATS (Signal Extraction in Arima Time Series). Instructions for the User », *Working Paper* 9628, Research Department, Banco de España, 1996.
- [16] Ladiray D., Quenneville B., « Désaisonnaliser avec la Méthode X11 », *Methodologica*, numéro spécial 8- , Université Libre de Bruxelles, 2001
- [17] Lytras Demetra P., Feldpausch Roxanne M., and Bell William R., « Determining Seasonality: A Comparison of Diagnostics From X-12-ARIMA », U.S. Census Bureau, June 2007.
- [18] Maillard V., « Théorie et pratique de la correction des effets de jours ouvrables », *document de travail*, Direction des Etudes et Synthèses Economiques, INSEE, Paris, G9405, 1994.
- [19] Maravall A., « Notes on Programs Tramo and Seats », Bank of Spain, 2008.
- [20] Soukup R. J., Findley D. F., « On the Spectrum Diagnostics Used by X-12-ARIMA to Indicate the Presence of Trading Day Effects after Modeling or Adjustment », *ASA Proceedings of the Joint Statistical Meetings* (Business and Economic Statistics Section), 1999.
- [21] Tondering C., « Frequently Asked Questions about Calendars », Version 2.8, available on [www.tondering.dk/clus/calendar.html](http://www.tondering.dk/clus/calendar.html), 2005.
- [22] U.S. Census Bureau, « X-12-ARIMA Reference Manual, Version 0.3 », Time Series Staff, Statistical Research Division, Washington, DC, 2009.
- [23] Young A. H., « Estimating trading-day variations in monthly economic series », *Technical Paper*, U.S. Department of Commerce, U.S. Bureau of the Census, Washington D.C., 1965.
- [24] Young A. H., « Comment on 'Modeling Time Series when Calendar Effects are Present' », *Applied Time Series Analysis of Economic Data*, Zellner, A. (editor), U.S. Department of Commerce, U.S. Bureau of the Census, pp 68-71, 1983.