La désaisonnalisation en temps de crise : enjeux et stratégies illustrées avec la méthode X-13-ARIMA

Anna SMYK

Février 2021

Introduction

En raison de la a crise sanitaire provoquée par l'épidémie de covid-19, de très nombreux indicateurs économiques ont affiché des valeurs exceptionnelles dès le mois de mars 2020. Les valeurs brutes prises par ces séries ont en général reflété une baisse inédite de l'activité mesurée, avec toutefois dans quelques secteurs une hausse exceptionnelle.

Les indicateurs économiques sont publiés corrigés des variations saisonnières (cvs)[1]. L'ajustement saisonnier est utilisé pour purger les séries de mouvements périodiques afin d'éviter que les variations conjoncturelles et les tendances de fond ne soient masquées par ces phénomènes récurrents, souvent de grande ampleur. Pour rendre ces indicateurs lisibles il faut estimer les facteurs saisonniers S[2] qui seront ensuite enlevés de la série brute. Celle-ci est, au préalable, décomposée en saisonnalité, tendance et irrégulier ($Y = S + T + I$). Au sein du système statistique public (SSP) seules sont publiées les séries brutes Y et cvs $Y_{\text{cvs}} = Y - S$, comme le recommandent les guidelines d'Eurostat[1] sur l'ajustement saisonnier. Le partage final exact entre T et I n'est pas ici primordial.

La crise sanitaire étant un phénomène conjoncturel, son impact doit être attribué à la composante irrégulière de afin d'être entièrement reflété par la série désaisonnalisée. Il ne doit pas contribuer, à court terme du moins, à l'estimation des facteurs saisonniers, dont la variation est supposée lente.

Hors, dans la plupart des chaînes de production, les coefficients saisonniers sont graduellement mis à jour lors de l'actualisation infra-annuelle des séries, lorsque de nouvelles données brutes sont disponibles. Cette stratégie, valable en régime courant, peut engendrer une ré-estimation erronée des coefficients saisonniers dans une situation conjoncturelle exceptionnelle.

Pour éviter ce phénomène, Eurostat a édicté des recommandations[2], relayées par le Département des Méthodes statistiques, consistant à exclure les points impactés par la crise de la mise à jour des coefficients saisonniers. Les producteurs de séries cvs au sein du SSP ont depuis adopté diverses stratégies allant dans ce sens.

Avant de décrire les différentes stratégies de production infra-annuelle et annuelle recommandables dans le contexte actuel, nous mettons en évidence les conséquences d'une absence de correction dans la phase de pré-ajustement, lorsque la décomposition de la série est faite par l'algorithme X-11.

1 Les conséquences de l'absence de correction d'un choc

Lorsque les coefficients saisonniers sont susceptibles d'être ré-estimés à chaque nouveau point brut, il est nécessaire de traiter un choc dès la phase de pré-ajustement, conçue à cet effet, bien que l'algorithme X-11 comporte également un système de correction des points atypiques.

Pour produire une série cvs, on cherche une décomposition du type [3] :

$$Y = T + S + I$$

Pour estimer ces composantes, la méthode X13-arima procède en deux étapes : une phase de pré-ajustement paramétrique et une méthode de décomposition (X-11) non paramétrique, par moyennes mobiles.

[1] Nous ferons le raccourci usuel en désignant par corrigées des variations saisonnières (cvs) des séries aussi, éventuellement corrigées des jours ouvrables (cvs-cjo)

[2] S contient aussi un éventuel effet de calendrier

[3] dans le cas d'un modèle additif, pris ici en exemple

La phase de pré-ajustement produit une série linéarisée, par régression linéaire avec résidus Arima. Elle est nettoyée des effets déterministes qui nuiraient à la qualité de décomposition, notamment des outliers (O) et des effets de calendrier C, et prolongée par un an de prévisions[4]

$$Y_{\text{lin}} = Y - \sum \hat{\alpha}_i O_{it} - \sum \hat{\beta}_j C_{jt}$$

La décomposition par X-11 intervient sur la série linéarisée et fournit des composantes provisoires.

$$Y_{\text{lin}} = T_{\text{decomp}} + S_{\text{decomp}} + I_{\text{decomp}}$$

Les effets estimés lors de la phase de pré-ajustement sont ensuite ré-alloués aux composantes adéquates. Les effets de calendrier seront *in fine* attribués à la composante saisonnière. L'allocation des outliers dépend de leur nature : les additive outliers (AO) et les transitory (TC) seront alloués à la composante irrégulière et les Level Shifts (LS) à la tendance.(cf. Annexe n°2)

$$I = I_{\text{decomp}} + Outliers_{\text{Irregulier}}$$

$$T = T_{\text{decomp}} + Outliers_{\text{Tendance}}$$

$$S = S_{\text{decomp}} + Cal[5]$$

Les composantes finales sont calculées en deux étapes. Les révisions de la série linéarisée, prévisions comprises détermineront les révisions du profil saisonnier. La philosophie des deux phases est très différente : lors du pré-ajustement, on peut modéliser les effets déterministes affectant la série, en choisissant leur période et leur intensité, et prendre en compte une partie plus ou moins grande de l'information apportée par un nouveau point brut, en une choisissant une politique de rafraîchissement adaptée (cf.Annexe n°1). A l'inverse, lors de la décomposition, la série est retraitée dans son intégralité, les paramétrages s'appliquant aux points passées, actuels et prévus, sans distinction.

1.1 Traitement par X-13 d'un choc ponctuel

Comparons une série où un choc ponctuel, par exemple une baisse brutale en mars 2020 a été corrigé par un additive outlier (AO) avec la même série non corrigée. L'absence de correction peut se produire en utilisant un mode d'actualisation "fixed model" (sans re-identification d'outliers sur la fin de la période) ou bien si le choc n'a pas été identifié automatiquement par l'algorithme en mode "last outliers". La Figure 1 montre que l'ajout d'un AO en mars évite une importante révision de la série linéarisée.

La série linéarisée est ensuite éventuellement à nouveau corrigée par l'algorithme X-11, qui repose sur un principe de lissage par moyennes mobiles successives, sous l'hypothèse d'une saisonnalité localement stable. Schématiquement, une première tendance est estimée par moyenne mobile d'ordre la égal à la périodicité ($2 * 12$, dans la cas mensuel), la composante saisonnière est extraite par moyenne mobile sur la série détrendée, ce qui permet d'obtenir une estimation de la série cvs. Celle-ci est de nouveau lissée pour améliorer l'estimation de la tendance et calculer plus précisément les facteurs saisonniers et la série cvs.

Cette succession de calculs, décrite dans l'annexe n°2, est appliquée trois fois, les deux premiers passages ne servant qu'à corriger la série des point atypiques. Cette correction repose sur le repérage de points dont l'estimation provisoire de l'irrégulier s'écarte trop, en valeur absolue et en nombre d'écart-types de sa valeur théorique[6].

La figure 2 montre que le choc intervenu en mars 2020 est lissé sur janvier, février et mars, avec par conséquent une intensité moindre en mars que ce qu'estime la phase de pré-ajustement.

Ce mode de correction peut-être désactivé, mais ce sera alors le cas pour l'intégralité de la série, avec un changement de mode de calcul en cours d'année. *In fine*, une révision de la série linéarise entraîne une revision du profil saisonnier, y compris passé (par effet de lissage). Si l'on pense que le profil saisonnier n'a pas de raisons d'avoir été sensiblement affecté, il faut minimiser l'impact sur la série linéarisée des nouveaux points bruts.

[4] horizon par défaut, les prévisions permettent d'utiliser un plus grand nombre de filtres symétriques, ayant de meilleures propriétés, dans la décomposition

[5] il peut aussi y avoir des effets affectés à la composante saisonnière autres que les effets de calendrier, notamment les seasonal outliers (SO), mais ces cas sont très rares et nous les ignorons ici

[6] 0 dans un schéma additif, 1 pour un schéma multiplicatif

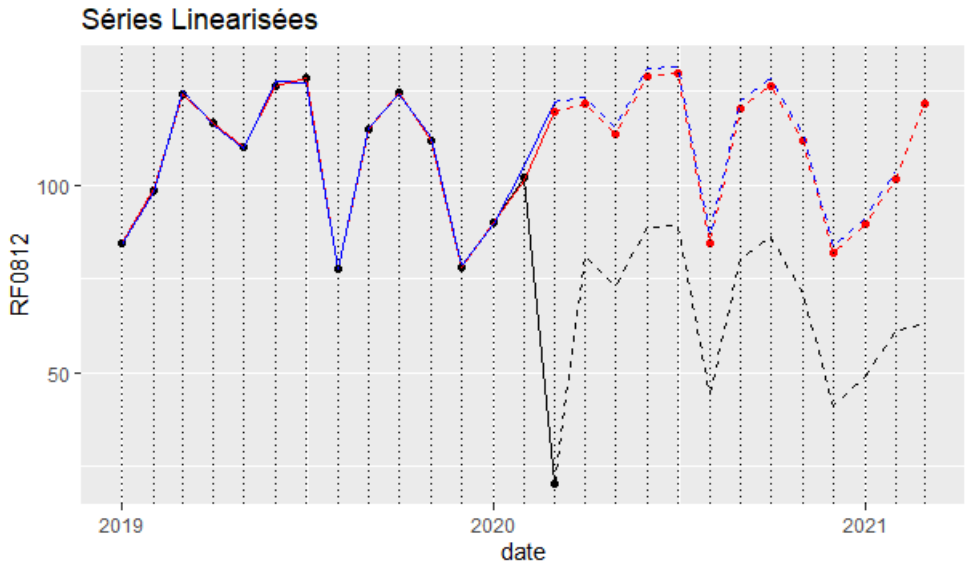


Figure 1: En bleu la série linéarisée avant avant que le point brut de mars 2020 ne soit disponible, en rouge la série linéarisée où le point de mars a été corrigé par un AO et en noir celle où aucune correction n'a été faite. La série brute correspond à la série linéarisée non corrigée, à l'effet de calendrier (ici faible) près. Les prévisions sont en pointillés.

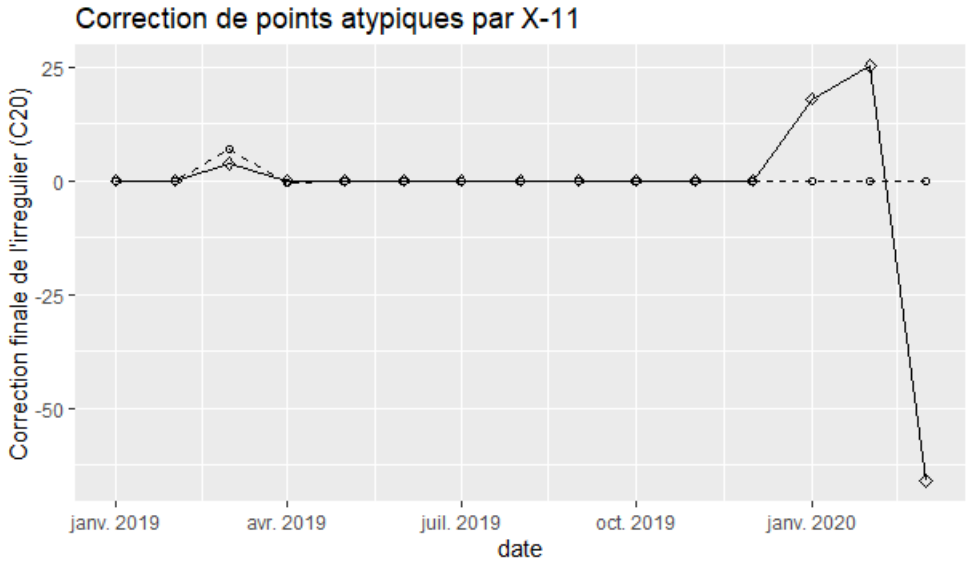


Figure 2: En trait plein les corrections de la série non corrigée dans la phase de pré-ajustement. En pointillés, les corrections, quasi-nulles, de la série corrigée avec un AO

1.2 Points de vigilance

L'arrivée d'une nouvelle valeur brute va plus ou moins modifier l'estimation de la série linéarisée, selon la modélisation des chocs retenue, le modèle Arima choisi [7] et le mode d'actualisation infra-annuelle.

Les changements de paramètres que l'on pourrait faire sur X-11 (modulation ou désactivation des points atypiques, longueur des moyennes mobiles) s'appliqueront à toute la série et auront toujours pour effet de repartir le choc enregistré en un point sur les points voisins.

En utilisant la méthode X-13-Arima, il faut porter son attention aux révisions de la série linéarisée, prévisions comprises. Une série linéarisée peu révisée garantit un profil saisonnier peu révisé également. Des corrections maîtrisées, ne peuvent se faire que dans la phase de pré-ajustement, les effets sont difficiles à distinguer dans la phase de décomposition.

[7] celui-ci déterminera notamment les prévisions

Existe-t-il, toutefois, un risque sur-correction ? Le problème est très asymétrique : l'oubli d'un outlier peut conduire à une revision conséquente et injustifiée du profil saisonnier y compris passé, alors que l'ajout à tort outlier ne va que "geler" l'estimation de la saisonalité, cette information pourra être prise ne compte ultérieurement, avec plus de visibilité sur les évènements. A court terme, la correction systématique n'est pas un problème, cela le devient à plus long terme quand on pense que le profil saisonnier mérite d'être revu.

Après avoir décrit les conséquences de la non correction d'un choc, nous allons rappeler les stratégies de désaisonnalisation recommandées dans le contexte actuel.

2 Actualisation infra-annuelle des séries

On se place ici dans un cadre de production mensuel ou trimestriel. Le cas bimestriel est plus rare. A l'Insee, il est traité avec la méthode Tramo-Seats car X-13-Arima n'autorise pas encore cette périodicité. Les recommandations ci-après restent entièrement valables. L'idée sous-jacente est que dans l'incertitude, comme celle qui a marqué le début de la crise sanitaire et qui ne s'est pas totalement estompée aujourd'hui, on préfère ne pas prendre en compte l'information nouvelle pour l'estimation de la saisonnalité.

2.1 Utiliser de coefficients saisonniers projetés

Une solution valable est d'utiliser des coefficients saisonniers estimés dans la période d'avant crise, c'est à dire ceux prévus lors campagne infra annuelle non impactée par la crise. Ils sont composés de deux éléments : les coefficients de calendrier, nuls en l'absence d'effet de calendrier, et les coefficients saisonniers hors effets de calendrier. $S_{final} = S + Cal$

Les coefficients saisonniers ont pour chaque période de l'année en cours des valeurs proches de la même période de l'année précédente, car on a supposé pour les calculer que l'évolution de la saisonnalité était lente. Les facteurs liés au calendrier eux sont variables, propres à chaque période du calendrier. L'intérêt d'utiliser des coefficients saisonniers finaux prévus prévus réside dans le fait qu'ils contiennent un effet de calendrier calculé pour la période d'intérêt.

L'avantage d'utiliser des coefficients exogènes est que l'on peut se passer du logiciel JDemetra+ et alléger ainsi la chaîne de production. Mais devoir modifier la chaîne de calcul en cours d'année peut aussi apparaître comme l'inconvénient majeur de cette solution, d'autant qu'il sera alors impossible de de repasser à une stratégie, plus fine, utilisant JDemetra + en cours d'année, sans recalculer le passé.

Si au bout de quelques mois, on se retrouve à court de coefficients projetés, deux cas sont à considérer. Si les effets de calendrier sont quasiment nuls (enquetes de conjoncture, données salariales) on peut se contenter d'utiliser les coefficients de la dernière année non impactée par la crise. Si on contraire on a besoin de ré-estimer les effets de calendrier, on peut prolonger ses modèles par des AO à chaque mois depuis mars 2020 et générer une nouvelle année de coefficients, ou bien se tourner vers cette méthode de mise en outlier du dernier point, qui est l'objet du paragraphe suivant.

2.2 Considérer les nouveaux points comme des Additive Outliers (AO)

Ce choix va produire un résultat numérique proche de celui obtenu avec les coefficients projetés car la série linéarisée sera très peu révisée. Le fait de privilégier des additive outliers, permet de faire les hypothèses les plus faibles possibles sur le profil d'impact de la crise. Le cumul d'AO est ce qui entraînera in fine la revision la plus faible du profil saisonnier et permettra de répercuter l'ensemble du choc conjoncturel dans la série désaisonnalisée. Comme on en cherche pas à estimer de tendance, le fait d'allouer le choc systématiquement à l'irrégulier n'a pas ici d'importance.

La version 2.2.3 de JDemetra+, sortie au début de l'été 2020, permet d'implémenter ce choix de façon automatique ("current AO approach"),y compris en production avec le cruncher, pour un ensemble de séries. L'avantage est que la chaîne de production n'aura pas à être modifiée et qu'un retour à une actualisation "last outliers" classique peut se faire immédiatement.

Cette approche peut être, bien entendu, également mise en oeuvre avec la méthode "last outliers", si l'on force à dire d'expert les points ayant besoin d'être mis en AO et non détectés par l'algorithme. Cela implique toutefois une intervention manuelle.

Si la série cvs prend des valeurs négatives, suite à de très fortes baisses de la série brute dans le cas d'un modèle additif, et que celles-ci n'ont pas de sens étant donnée la grandeur mesurée, par exemple un indice de production, il faut changer le schéma de décomposition de la série. Un schéma multiplicatif évitera par

définition les valeurs négatives, mais obligera à ré-estimer l'ensemble des paramètres sur la période d'avant crise. L'actualisation pourra ensuite être refaite avec l'ajout d'un outlier à chaque point depuis mars 2020. Les revisions sur le passé de la période peuvent être conséquentes.

Estimer une douzaine, voire plus, d'outliers consécutifs soulève un problème de parcimonie et de robustesse des estimateurs des coefficients de régression. Dans la pratique les séries sont en général suffisamment longues pour que la perte de degrés de liberté dans l'estimation puisse être négligée. Des tests seraient à mener en présence de séries courtes, schématiquement de moins de 5 ans. Pour la production de la plupart des indicateurs, on peut donc négliger cet inconvénient encore quelques mois, voire en 2021 dans son intégralité, mais il faudra à un moment revenir à une estimation plus robuste en faisant des hypothèses plus explicites sur l'impact de la crise. Tel sera l'enjeu de la première campagne annuelle d'après mars 2020.

3 Stratégies pour les campagnes annuelles

Les campagnes annuelles sont des opérations de re-évaluation complète des paramètres d'estimation des séries désaisonnalisées. Il s'agit, dans le contexte actuel,de ne pas trop fortement réviser le profil saisonnier du fait d'un évènement conjoncturel. Les choix devront être validés à l'aune de l'ampleur des revisions du profil saisonnier passé qui paraissent acceptables.

Une ré-estimation complète ne vaut le coup que s'il l'on est décidé à prendre des décisions plus tranchées quant à l'impact de de la crise. Ré-estimer en mettant en AO tous les points depuis mars va produire le même résultat que faire une ré-estimation complète jusqu'en février 2020 puis actualiser par la current-AO-approach.

Toutefois, ré-estimer jusqu'en février 2020, puis suivre les stratégies d'actualisation infra-annuelles à partir de mars peut être une solution d'attente envisageable pour les séries dont la dernière re-estimation complète est suffisamment ancienne. C'est une solution provisoire qui consiste à intégrer une douzaine, ou plus, de points d'avant crise et permet d'attendre encore un peu avant de se décider sur la modélisation à retenir.

L'enjeu principal est de déterminer quels points apportent une information devant contribuer à l'estimation du profil saisonnier et quelle information doit rester considérée comme de la perturbation conjoncturelle.La construction d'un profil d'impact de la crise devra être ajustée et testée sectoriellement. Une approche indirecte devient plus appropriée, car pour les séries désagrégées il sera plus simple de déterminer un mode de correction selon les perturbations connues par un secteur donné.

Par exemple, modéliser l'intégralité de la période récente avec un level shift (LS), revient à supposer que le niveau récent est bien le nouveau niveau de la série, ce qui est une hypothèse forte, celle d'une baisse durable. Le profil est en général plus heurté et différencié entre secteurs d'activité : on constate souvent une forte baisse puis une reprise, avec plus ou moins d'effet de compensation, éventuellement suivie d'autres phases de baisse inhabituelles. A contrario, la modélisation avec des additive outliers est celle qui revient à faire le moins d'hypothèses sur ce profil et permet d'estimer des effets différenciés pour chaque période.

Toutefois, pour gagner en robustesse et en précision, il sera à un moment nécessaire de regrouper les outliers utilisés pour l'actualisation infra-annuelle dans un regresseur unique[8], où l'on pourra moduler l'intensité de l'effet de la crise, par exemple selon la durée du confinement dans le mois. Le défaut de cette approche est d'estimer un effet moyen, alors que les outliers ponctuels pouvaient capter des effets différenciés pour chaque mois. Ce genre de regresseur est donc à tester et à comparer à une suite d'outliers.

JDemetra+ permet d'implémenter ce type de correction en utilisant les "user-defined variables". On affectera le regresseur à la composante irrégulière[9]. Tester un grand nombre de régresseurs correspondant à différents profils d'impact, peut être fait avec le package RJDemetra[3], qui permet d'automatiser la modification des spécifications via un script R, ainsi que de récupérer les coefficients et les statistiques.

Conclusion

La crise sanitaire ne doit pas à court terme impacter le profil saisonnier, par définition localement stable, des séries. Dans le contexte actuel, une attention particulière doit être portée sur les révisions de la série

[8] ou en tout cas dans un moins grand nombre de regresseurs

[9] par défaut, ou à la tendance si l'objectif du regresseur considéré est uniquement de corriger la tendance

linéarisée, prévisions comprises. Des corrections maîtrisées ne peuvent se faire que dans la phase de pré-ajustement, dans le cas contraire la décomposition par moyenne mobile lissera l'impact du choc. Les choix de modélisation seront validés à l'aune de l'ampleur de la révision du profil saisonnier qui semble acceptable pour le producteur.

Lors de la production infra-annuelle, l'utilisation de coefficients saisonniers projetés ou l'ajout systématique d'additive outliers (AO) sont recommandés. Les ré-estimations complètes de l'ensemble de paramètres, usuellement annuelles, peuvent être retardées, afin de gagner en visibilité et de tester des modélisations plus parcimonieuses de l'impact de la crise.

References

[1] Eurostat. *ESS Guidelines on Seasonal Adjustment*. Tech. rep. Eurostat Methodologies and Working Papers, European Commission, 2015. DOI: 10.2785/317290. URL: http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-15-001.

[2] URL: https://ec.europa.eu/eurostat/cros/content/%20treatment-covid19-seasonal-adjustmentmethodological-note_en.

[3] Anna Smyk and Alice Tchang. *R Tools for JDemetra+, seasonal adjustment made easier*. Tech. rep. Documents de travail de méthodologie statistique, Insee, 2021. URL: https://www.insee.fr/en/statistiques/5019812.

Annexe n°1

Actualisation infra-annuelle : les "Refresh Policies" de JDemetra+

Approche	Option JD+	Option cruncher
Ré-utilisation des coefficients saisonniers de l'année passée Utilisation des coefficients saisonniers projetés		
On applique le modèle identifié et estimé sans les nouveaux points et on classe tous les nouveaux points en AO*	Current adjustment (AO approach)*	current
On applique le modèle identifié et estimé sans les nouveaux points à la série prolongée	Fixed model	fixed (f)
Les paramètres du modèle sont inchangés et seuls les coefficients de la régression linéaire sont ré-estimés	Estimate regression coefficients	fixedparameters(ou fp)
Les coefficients du modèle ARIMA sont aussi ré-estimés	+ Arima parameters	parameters (ou p)
Les outliers de la dernière année sont aussi ré-identifiés	+ Last outliers	lastoutliers (ou l)
Tous les outliers de la série sont aussi ré-identifiés	+ All outliers	outliers (ou o)
Les ordres du modèle arima sont aussi ré-identifiés	+ Arima model	stochastic (ou s)
Tous les paramètres du modèle sont ré-estimés	Concurrent	complete/concurrent (ou c)

* à partir de la version 2.2.3

Paramètres du modèle: jeux de régresseurs CJO, outliers, éventuelles autres variables incluses dans la linéarisation ET ordres de l'ARIMA

Identification : choix/détermination des outliers, du jeux de régresseurs de JO, des ordres de l'ARIMA
Estimation : estimation des coefficients du modèle (déjà identifié)
Tous ces paramétrages concernent la série linéarisée, qui est ensuite de nouveau décomposée par X-11

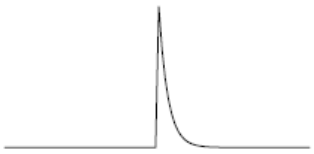
Annexe n°2

Les principaux types d'outliers

Choc ponctuel
Additive outlier (AO)
Affecte l'Irrégulier

Changement de niveau
Level Shift (LS)
Affecte la Tendance

Changement de niveau transitoire
Transitory Change (TC)
Affecte l'Irrégulier



Annexe n°3

Principe itératif de X11

L'algorithme X-11 se fonde principalement sur le bloc de calculs ci-dessous, appliqué trois fois à la série. Les deux premiers passages ont pour objet de corriger les points dont la composante irrégulière semble atypique et le troisième d'estimer les composantes S, T, I finales.

Première estimation de la cvs

1. Estimation de la **tendance-cyle** par moyenne mobile 2×12 :

$$TC_t^{(1)} = M_{2\times 12}(X_t)$$

2. Estimation de la composante **saisonnier-irrégulier** :

$$(S_t + I_t)^{(1)} = X_t - TC_t^{(1)}$$

3. Estimation de la composante **saisonnière** par moyenne mobile 3×3 **chaque mois/trimestre** :

$$S_t^{(1)} = M_{3\times 3}\left[(S_t + I_t)^{(1)}\right] \text{ et normalisation } Snorm_t^{(1)} = S_t^{(1)} - M_{2\times 12}\left(S_t^{(1)}\right)$$

4. Première estimation de la série corrigée des variations saisonnières :

$$Xsa_t^{(1)} = (TC_t + I_t)^{(1)} = X_t - Snorm_t^{(1)}$$

Seconde estimation de la CVS :

1. Estimation de la **tendance-cyle** par moyenne de Henderson (généralement 13 termes, meilleur pouvoir de lissage que la $M2 * 12$, mais ne peut s'appliquer qu'à une série désaisonnalisée) :

$$TC_t^{(2)} = H_{13}(Xsa_t^{(1)})$$

2. Estimation de la composante **saisonnier-irrégulier** :

$$(S_t + I_t)^{(2)} = X_t - TC_t^{(2)}$$

3. Estimation de la composante **saisonnière** par moyenne mobile 3×5 (généralement) pour **chaque mois/trimestre** :

$$S_t^{(2)} = M_{3\times 5}\left[(S_t + I_t)^{(2)}\right] \text{ et normalisation } Snorm_t^{(2)} = S_t^{(2)} - M_{2\times 12}\left(S_t^{(2)}\right)$$

4. Estimation de la série corrigée des variations saisonnières :

$$Xsa_t^{(2)} = X_t - Snorm_t^{(2)}$$