

Désaisonnalisation de séries longues : l'apport des modèles structurels

Anna Smyk, Département des Méthodes Statistiques

Journées de méthodologie statistique de l'Insee (JMS)
Mars 2022 / PARIS

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Décomposer pour désaisonnaliser

- Objectif de l'ajustement saisonnier : purger les séries de mouvements infra-annuels périodiques
- Il faut estimer des facteurs saisonniers (S) qui seront enlevés de la série brute $Y_{CVS} = Y - S$
- Pour cela, on décompose la série en 3 composantes inobservables : saisonnalité, tendance et irrégulier ($Y = S + T + I$)
- divers algorithmes, trois présentés dans la suite



Difficultés d'estimation liées à longueur de la série

séries trop courtes (nouvelles mesures)

- instabilité des estimations
- voire impossibilité d'utiliser les algorithmes

séries très longues maintenues par les institutions statistiques

- le processus générateur de données évolue...
- ...en même temps que les institutions et les comportements (ex: travail du dimanche, consommation de produits frais)
- ruptures possibles en cours de série (éventuelles rétropolations)

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels**
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Les algorithmes de désaisonnalisation utilisés dans les INS

- Deux familles : X13-ARIMA et Tramo-Seats
- Algorithmes en deux étapes :
 - une étape pré-ajustement destinée à linéariser la série (enlever les effets de calendrier, les outliers..) à l'aide d'une modélisation Reg-ARIMA
 - une étape de décomposition de la série linéarisée pour déterminer les composantes S , T , I (auxquelles seront réinjectées les corrections du pré-ajustement)
 - SEATS : décomposition paramétrique à partir du modèle Arima
 - X-13 : décomposition non paramétrique à l'aide de moyennes mobiles ad-hoc (par itération)

Face à des séries longues : quelles rigidités ?

Pré-ajustement fait "à coefficients fixes" sur l'intégralité de la période

- effet de calendrier
- détection de points aberrants (outliers)
- structure d'auto-corrélation (modèles Arima)

Décompositions plus ou moins rigides

- SEATS : filtres fixés à partir du modèle Arima du pré-ajustement (rigide)
- X-13 : moyennes mobiles (fenêtres mobiles, souple)
- X-13 : utilisation d'une prévision en fin de série avec modèle Arima (rigide)

Face à ces rigidités : quel traitement pratique ?

Que disent les Guidelines d'Eurostat sur l'ajustement saisonnier ?

- sous optimal de désaisonnaliser les séries de plus de 20 ans (alternative C)
- recommandation : estimer des cvs par sous-périodes (stratégies et coûts cf. JMS 2018)

Mais si l'on ne veut pas couper

- utiliser des algorithmes fondés sur modélisation plus souple
- choisir la méthode la moins pénalisante

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels**
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Le modèles structurels (Basic Structural Models (BSM))

Nouvelle classe d'algorithmes de désaisonnalisation

- utilisation du cadre général de modèles "espaces d'états"
- permet d'estimer des composantes inobservables d'une variable mesurée (équation de mesure) en modélisant explicitement la dynamique sous-jacente de celles-ci (équations d'état ou de transition)
- application à la désaisonnalisation formalisée par Harvey (1989) (références dans l'article)

Équation de mesure

On exprime les données observées en fonction des composantes inobservables dans une équation de mesure.

On note y_t la série observée

$$y_t = T_t + S_t + X_t + \nu_t, \quad t = 1, \dots, n,$$

avec

T_t = tendance

S_t = composante saisonnière

X_t = effets déterministes : calendrier, outliers

ν_t = Irrégulier

Équation d'état

Dynamique des composantes inobservables

$$T_t = T_{t-1} + \beta_{t-1} + w_{T,t}$$

$$\beta_t = \beta_{t-1} + w_{\beta,t}$$

$$S_t = - \sum_{i=1}^{s-1} S_{t-i} + w_{S,t}$$

On introduit la pente β pour forcer une tendance linéaire locale. La représentation de la saisonnalité est ici très simple, de nombreuses variantes sont possibles. On peut aussi modéliser ici des effets déterministes ("pré-ajustement") à coefficients variables.

Représentation espace d'états

Le modèle admet une représentation espace d'états:
(SSF: "state space framework")

$$y_t = A_t x_t + \nu_t, \quad \nu_t \sim \text{NID}(0, R)$$

$$x_t = \Phi_t x_{t-1} + w_t, \quad w_t \sim \text{NID}(0, Q)$$

Vecteur d'état :

$$x_t = (T_t, \beta_t, S_t)'$$

Matrice de mesure :

$$A_t = (1 \ 0 \ 1)$$

Représentation espace d'états (suite)

Équation d'état (ou de transition):

$$\begin{bmatrix} T_t \\ \beta_t \\ S_t \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -(B + \dots B^{s-1}) \end{bmatrix} \begin{bmatrix} T_{t-1} \\ \beta_{t-1} \\ S_t \end{bmatrix} + \begin{bmatrix} w_{Tt} \\ w_{\beta t} \\ w_{S_t} \end{bmatrix} \quad \begin{cases} w_T \sim N(0, \sigma_T^2) \\ w_\beta \sim N(0, \sigma_\beta^2) \\ w_S \sim N(0, \sigma_S^2) \end{cases}$$

avec B : opérateur retard

Estimation

On utilise

- Filtre (et lisseur) de Kalman (KFS)
- maximum de vraisemblance

Idées

- prévoir les observations pas à pas (erreur de mesure)
- prévoir les variables d'état pas à pas conditionnellement aux observations
- actualiser ces prévisions (filtrage, en fonction de l'erreur de mesure)
- lissage : actualisation des prévisions à rebours quand toutes les observations sont connues
- vraisemblance : produit de densités conditionnelles calculables récursivement

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes**
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Approche structurelle vs Modèles Arima

L'approche BSM a de nombreux avantages par rapport à la modélisation Reg-arima (Box-Jenkins; BJ) :

- BSM: modélisation explicite, pré-ajustement et décomposition en une seule étape
- BJ : approche boîte noire, la tendance et la saisonnalité sont éliminées par différenciation, il est nécessaire de stationnariser les séries (question délicate de la sur-différenciation), puis mettre en oeuvre une méthode de décomposition
- BSM : flexibilité due au caractère récursif de l'estimation : se prête aux changements de structure, coefficients variables (CJO) et traitement des valeurs manquantes
- BJ : homogènes dans le temps (défaut pour traitement des séries longues)

Approche structurelle vs Modèles Arima (2)

- BSM : un traitement multivarié est beaucoup plus simple (procédures identiques)
- héritage commun pour les deux approches : lissage exponentiel
- si la structure d'auto-corrélation de la série est simple (airline $(0,1,1)(0,1,1)$ par exemple, on trouve une modélisation structurelle équivalente (cf illustration 2)

Pourquoi l'approche structurelle pas (ou très peu) utilisée dans les statistiques officielles ?

Quelques raisons :

- Historiques : développement plus tardif des BSM (80's), même si SSF(60's). Approche X-11 (50's), enrichie par les modèles Arima (70's). Les BSM sont arrivés en "terrain conquis".
- Techniques : l'approche BSM requiert un traitement assez complexe à optimiser en pratique, avec outils informatiques "lents". Contrainte complètement levée aujourd'hui.
- Conjoncturelles : poids des personnalités et des institutions qui ont porté les projets "concurrents" (Dagum, Findley, Maravall)

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier**
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Rupture de profil saisonnier : "Seasonality break" : définition

Étude d'un cas particulier d'outlier :

- "Seasonality break" : rupture soudaine du profil saisonnier dans son ensemble (différent du "seasonal break" que l'on traite usuellement avec un "seasonal outlier" : rupture sur une période, "reste" uniformément ré-alloué)
- détection d'un choc la composante saisonnière dans son ensemble (résidus lissés)
- approche De Jong(1998) généralisée par Palate (2021)
- la position de la rupture de profil saisonnier est automatiquement identifiée (max de la statistique de test)

Modélisation de la rupture

Une fois que la rupture a été détectée, on remodelise la série : une seule tendance mais deux saisonnalités différentes

- Forme uni-variée (équation de mesure)

$$y_t = T_t + 1_{t < t_b} S_{1_t} + 1_{t \geq t_b} S_{2_t} + \nu_t$$

- Forme multi-variée équivalente

$$y_{1_t} = T_t + S_{1,t} + \nu_t$$

$$y_{2_t} = T_t + S_{2,t} + \nu_t$$

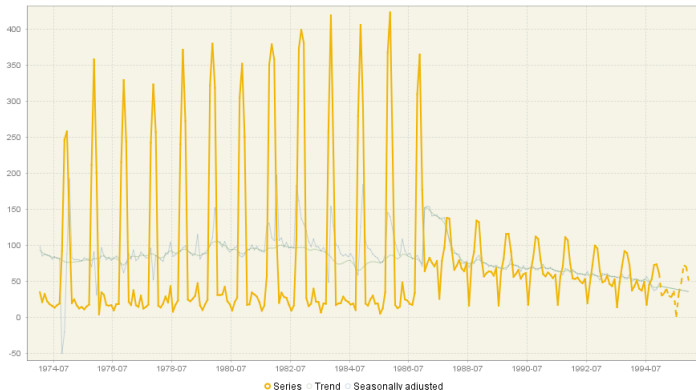
avec

$$y_{1_t} = \begin{cases} y_t & \text{if } t < t_b \\ \text{na} & \text{if } t \geq t_b \end{cases} \quad y_{2_t} = \begin{cases} \text{na} & \text{if } t < t_b \\ y_t & \text{if } t \geq t_b \end{cases}$$

- Pas de problème de raccordement, tendance et série cvs estimées sur l'ensemble de la période

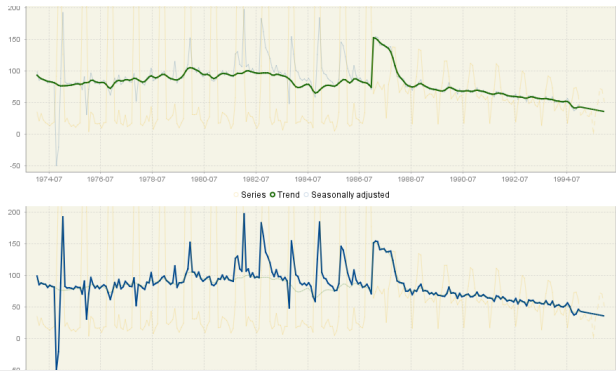
Un exemple de rupture violente du profil saisonnier (1)

La production de sucre en Belgique connaît une rupture majeure du profil saisonnier à la fin de l'année 1986.



Tentative de désaisonnalisation classique

Si l'on essaie d'obtenir une cvs classique avec Tramo-Seats, en autorisant les Seasonal Outliers. Il faudrait couper la série et la traiter par sous-périodes.



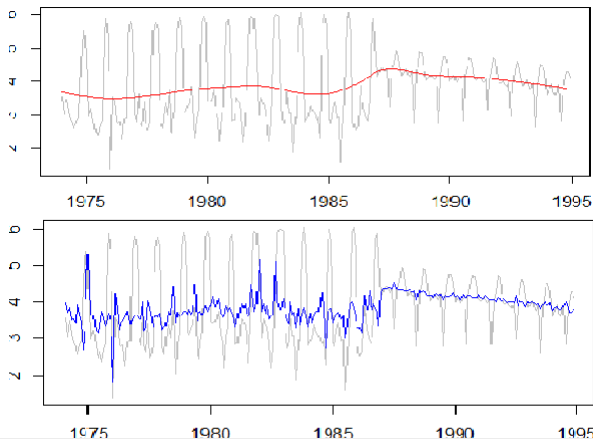
Regression model			
Outliers			
	Coefficients	T-Stat	P[T > t]
SO (11-1987)	-278,7919	-35,68	0,0000
SO (10-1987)	-215,1439	-29,14	0,0000
AO (1-1975)	107,8418	13,27	0,0000
SO (12-1987)	-153,6606	-19,72	0,0000
TC (10-1974)	-132,9650	-14,22	0,0000
SO (12-1983)	-143,1907	-19,72	0,0000
LS (1-1987)	80,3075	8,71	0,0000
TC (9-1982)	87,7097	10,32	0,0000
AO (12-1975)	-50,7437	-6,18	0,0000
SO (10-1979)	71,8113	11,06	0,0000
AO (1-1982)	90,7650	12,11	0,0000
SO (12-1981)	88,9883	12,39	0,0000
AO (12-1984)	71,9510	9,09	0,0000
TC (10-1985)	65,6748	7,56	0,0000
TC (11-1983)	94,6324	10,88	0,0000
AO (12-1979)	47,0435	6,18	0,0000
TC (12-1974)	76,8249	7,54	0,0000
TC (11-1984)	60,2597	6,88	0,0000
SO (7-1987)	-41,9155	-6,53	0,0000
TC (9-1981)	36,7089	4,32	0,0000
AO (11-1977)	-35,5716	-4,81	0,0000
TC (10-1983)	-39,6326	-4,67	0,0000
AO (5-1979)	29,0058	3,96	0,0001



Tendance et cvs dans le cadre BSM

Ajustement saisonnier en utilisant l'approche BSM multi-variée.

Tendance et cvs :



Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion

Stabilité du message conjoncturel en fin de période

- le message conjoncturel en fin de période (3 dernières années) est souvent l'information la plus importante pour les utilisateurs de séries cvs-cjo
- désaisonnaliser des séries longues expose à un message conjoncturel erroné (évolution du PGD et rigidités)
- longueur optimale d'estimation entre 8 et 12 ans (consensus)
- si on laisse ses séries vieillir, sans les couper : quel algorithme serait le plus pénalisant (X-13, Tramo-Seats ou BSM)

Comparaisons empiriques (1/2)

- estimation de la série cvs sur longue période (L) et sur période optimale (O), pour les trois algorithmes
- mesure de l'écart entre ces estimations sur l'intégralité de la période commune (optimale) et sur les 3 dernières années

Comparaisons empiriques (2/2)

Trois distances utilisées

- moyenne des écarts en niveau

$$EN(Y_t) = \frac{1}{T} \sum_{k=1}^T \left| \frac{cvs_k^L - cvs_k^O}{cvs_k^L} \right| \times 100$$

- moyenne des écarts en taux d'évolution

$$ETC(Y_t) = \frac{1}{T-1} \sum_{k=2}^T \left| \frac{cvs_k^L - cvs_{k-1}^L}{cvs_{k-1}^L} - \frac{cvs_k^O - cvs_{k-1}^O}{cvs_{k-1}^O} \right| \times 100$$

- proportion de points avec sens d'évolution contraires

$$CohS(Y_t) = \frac{1}{T-1} \sum_{k=2}^T I_{(cvs_k^L - cvs_{k-1}^L)(cvs_k^O - cvs_{k-1}^O) < 0} \times 100$$

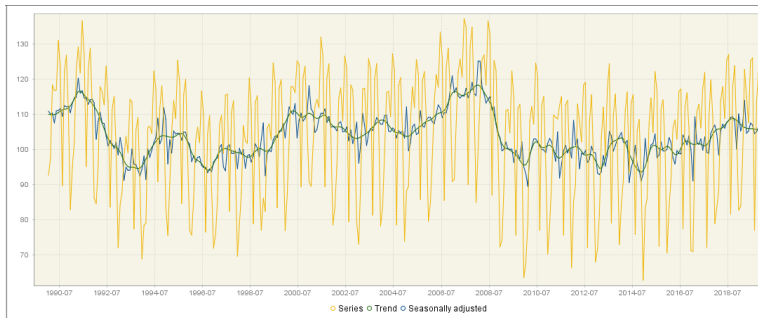
Données et outils

- séries mensuelles d'indices Ipi
- séries mensuelles de soldes d'opinion de l'enquête de conjoncture dans l'industrie
- période 1990 -2019 (2012 -2019), coupées avant le Covid
- séries linéarisées, pour mesurer l'effet des différences de décomposition
- calculs faits avec les algorithmes disponibles dans le logiciel JDemetra+ et notamment le package rjd3sts

(sts: "structural time series")

Quelques faits stylisés (1/5)

Profil 1 : Série mensuelle d'indices très saisonnière, structure d'autocorrélation Airline $(0,1,1)(0,1,1)$ (Ipi RF0812)



Quelques faits stylisés (2/5)

Méthode	EN	ETC	CohS	EN 3ans	ETC 3ans	CohS 3ans
X-13	0.43	0.57	5.26	0.10	0.16	0.00
BSM	0.65	0.80	12.63	0.46	0.71	8.33
Tramo-Seats	0.78	1.11	12.63	0.60	1.01	8.33

Sur ce type de série : écarts faibles, résultats BSM et Tramo-Seats proches, plus grande stabilité avec X-13.

Quelques faits stylisés (3/5)

Profil 2 : Série mensuelle d'indices très saisonnière, structure d'autocorrélation plus complexe (3,1,1)(0,1,1) (Ipi RF1051)

Méthode	EN	ETC	CohS	EN 3ans	ETC 3ans	CohS 3ans
X-13	0.28	0.36	4.21	0.10	0.16	2.78
BSM	0.27	0.37	6.32	0.24	0.31	11.11
Tramo-Seats	0.29	0.37	9.47	0.26	0.28	16.67

Sur ce type de série : plus grande stabilité avec X-13, BSM (flexible) produit moins d'incohérences que Tramo-Seats (modèle fixe)

Quelques faits stylisés (4/5)

Profil 3 : Série mensuelle de soldes d'opinion, production passée dans l'industrie (C5), structure d'autocorrélation $(3,1,1)(0,1,1)$

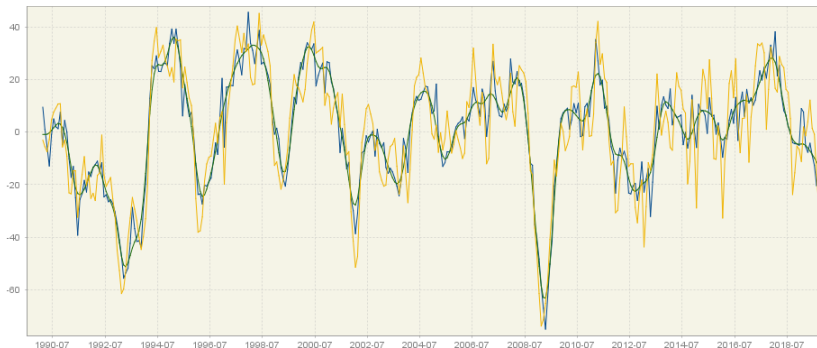


Figure: Solde production passée secteur C5 : brute (en jaune), cvs X-13

Quelques faits stylisés (5/5)

Méthode	EN	CohS	EN 3ans	CohS 3ans
X-13	0.81	4.21	0.33	2.78
BSM	2.86	16.84	2.55	16.67
Tramo-Seats	1.60	9.47	0.99	8.33

"Mauvais cas" pour le BSM qui produit les résultats les moins stables.

Stabilité du message en fin de période

Si l'on se focalise sur les différences de décomposition entre les algorithmes :

- X-13, non paramétrique, donne de meilleurs résultats que les algorithmes dont la décomposition est fondée sur une modélisation explicite
- les BSM et Tramo-Seats donnent des résultats proches pour des structures d'autocorrélation simples
- des résultats plus stables (vs modélisation Arima fixe) auraient pu être attendus, mais on trouve facilement des "contre-exemples"

(Généralisation impossible, bien entendu, même si l'on étudie un grand nombre de séries réelles ou simulées, juste faits stylisés sur profils type)

Plan

- 1 Désaisonnalisation et longueur de série
- 2 Structure et rigidités des algorithmes usuels
- 3 Cadre alternatif : les modèles structurels
- 4 Éléments de comparaison entre algorithmes
- 5 Illustration 1 : rupture de profil saisonnier
- 6 Illustration 2 : stabilité du message en fin de période
- 7 Conclusion**

Conclusion

- face aux rigidités des algorithmes usuels face à l'évolution du processus utilisation d'un cadre alternatif pour la désaisonnalisation : les modèles structurels (BSM)
- BSM : traitement unifié de la saisonnalité pré-ajustement et décomposition en une seule étape avec explicite des composantes
- BSM : solution intéressante pour traiter les séries longues avec rupture de profil saisonnier (et effets JO variables)
- le message conjoncturel en fin de période est sensible à la période d'estimation, une décomposition non paramétrique (X-13) semble produire les estimations les plus stables

Merci de votre attention

Annonce en avant-première :

- séminaire de Méthodologie sur la désaisonnalisation de données haute-fréquence (quotidiennes, hebdomadaires)
- le **jeudi 30 juin après-midi**, à l'Insee (White, Montrouge)
- ouvert à tous
- programme non définitif, annonce à venir