Désaisonnalisation de séries haute fréquence avec le package rjd3highfreq de JDemetra+ 3.0

Anna Smyk, Insee, Département des Méthodes Statistiques

Séminaire de méthodologie statistique de l'Insee
29 juin 2022 / Montrouge

Plan

Données haute-fréquence

On nomme ainsi les séries temporelles d'une fréquence plus haute que mensuelle, ce sont en général des données

- hebdomadaires (victimes d'accidents de la route)
- journalières (naissances, décès)
- horaires (consommation d'électricité, débit d'un fleuve)

Ces séries peuvent être saisonnières

Objectif de la présentation

- montrer comment les algorithmes de désaisonnalisation développés pour des séries mensuelles et trimestrielles ont été adaptés dans JDemetra+ v3.0 pour les traiter

Ces algorithmes sont accessibles avec le package {rjd3highfreq}

Ajustement saisonnier : objectif et processus (1/2)

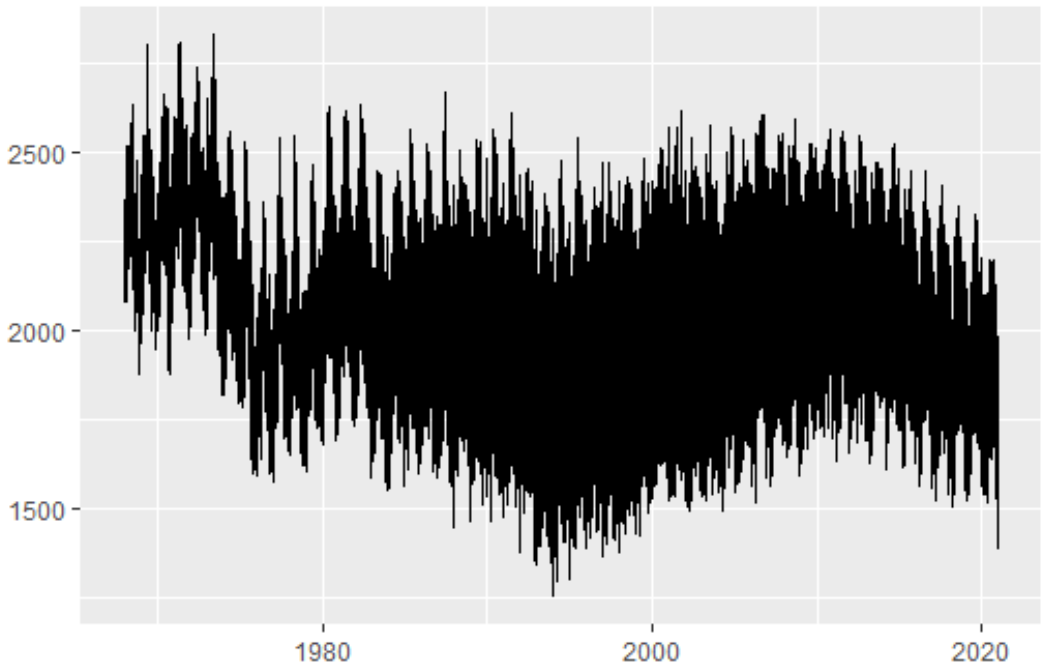
- purger les séries des variations infra-annuelles périodiques
- estimation de facteurs saisonniers (S) et de calendrier (C) qui seront enlevés de la série brute $Y_{cvs} = Y - S - C$
- décomposition la série en composantes inobservables : saisonnalité, tendance et irrégulier ($Y = T + S + I$), après avoir corrigé C
- deux algorithmes très utilisés : X13-Arima et Tramo-Seats

Ajustement saisonnier : objectif et processus (2/2)

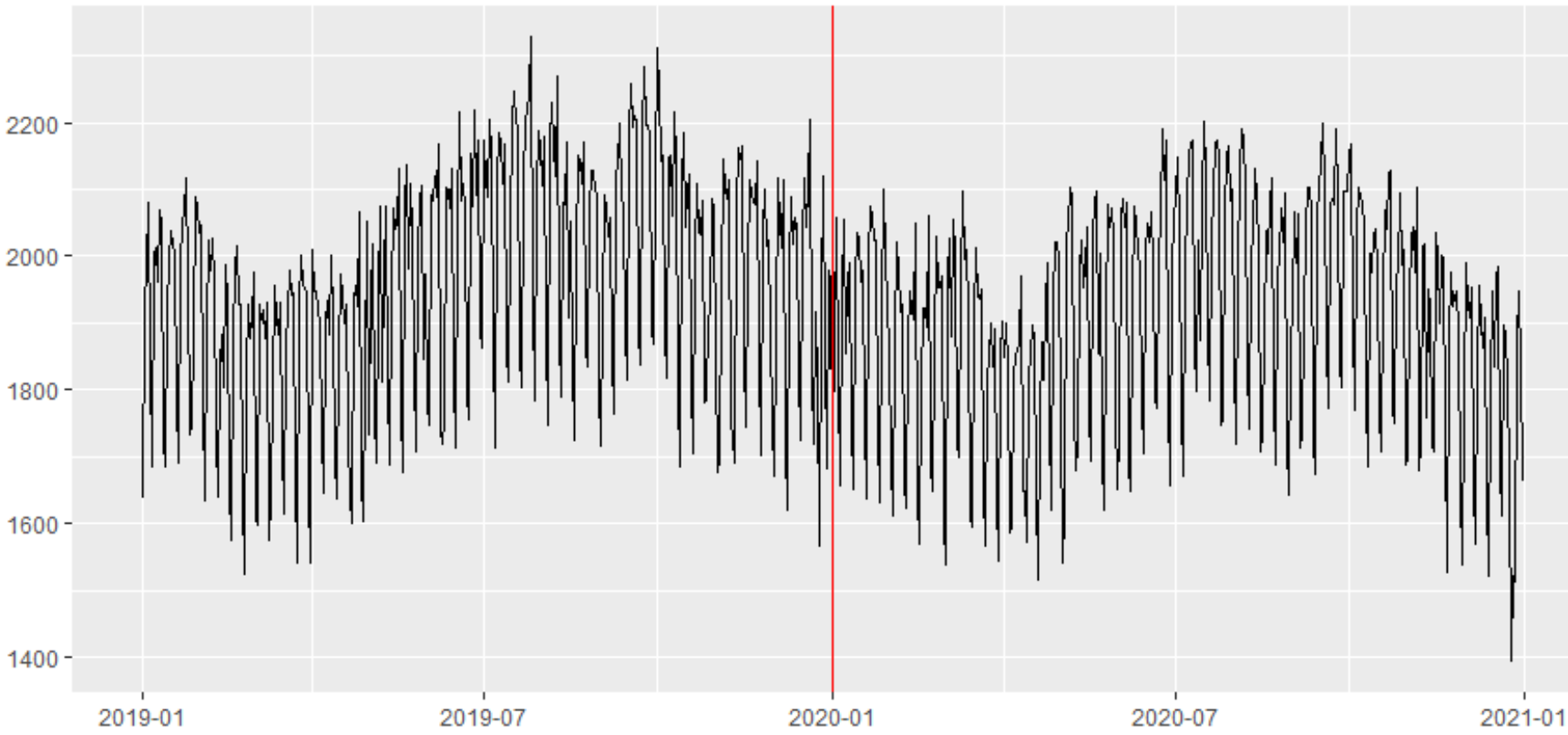
- identification de la saisonnalité (graphiques, tests)
- linéarisation (outliers, calendrier)
- décomposition (par moyennes mobiles ou paramétrique)
- calcul de la série cvs-cjo
- recherche de saisonnalité résiduelle

Plan

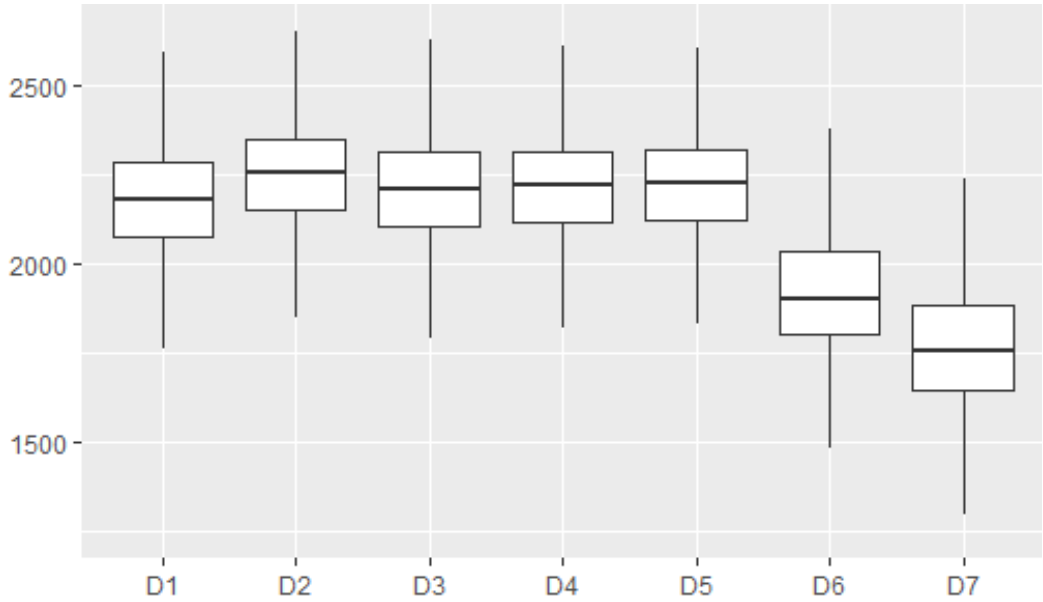
Nombre de naissances quotidiennes 1968-2020



Nombre de naissances quotidiennes zoom 2019-2020

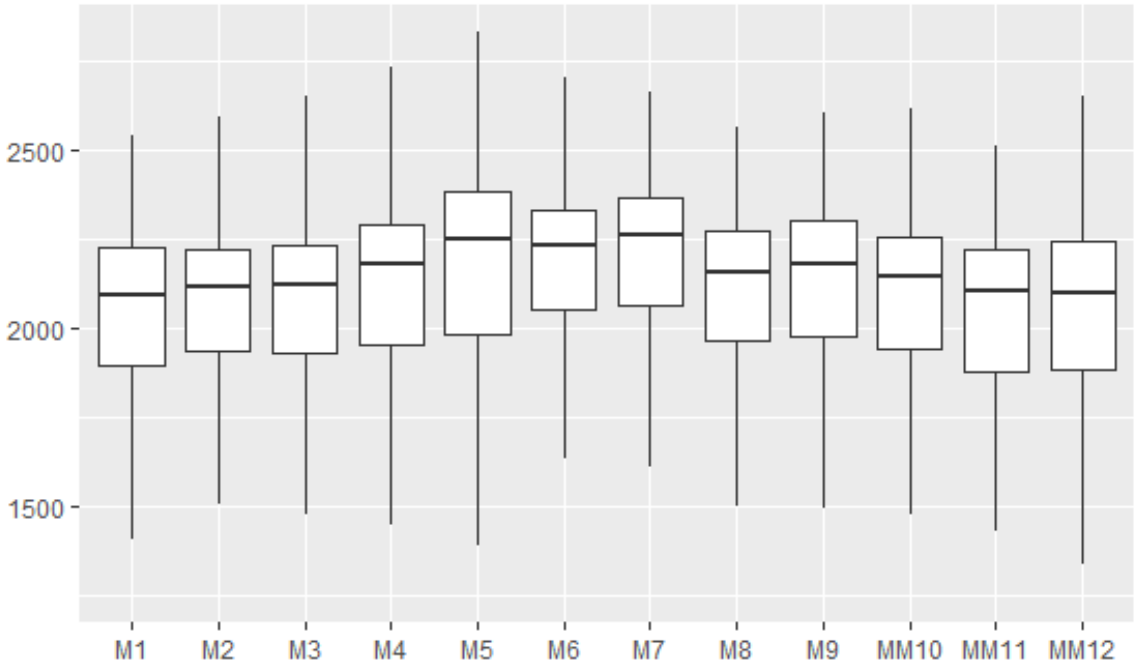


Répartition selon le jour de la semaine (1968-2020)



Mise en évidence d'une périodicité hebdomadaire ($p = 7$)

Répartition selon le mois de l'année (1986-2020)



Mise en évidence d'une périodicité annuelle ($p = 365.25$)

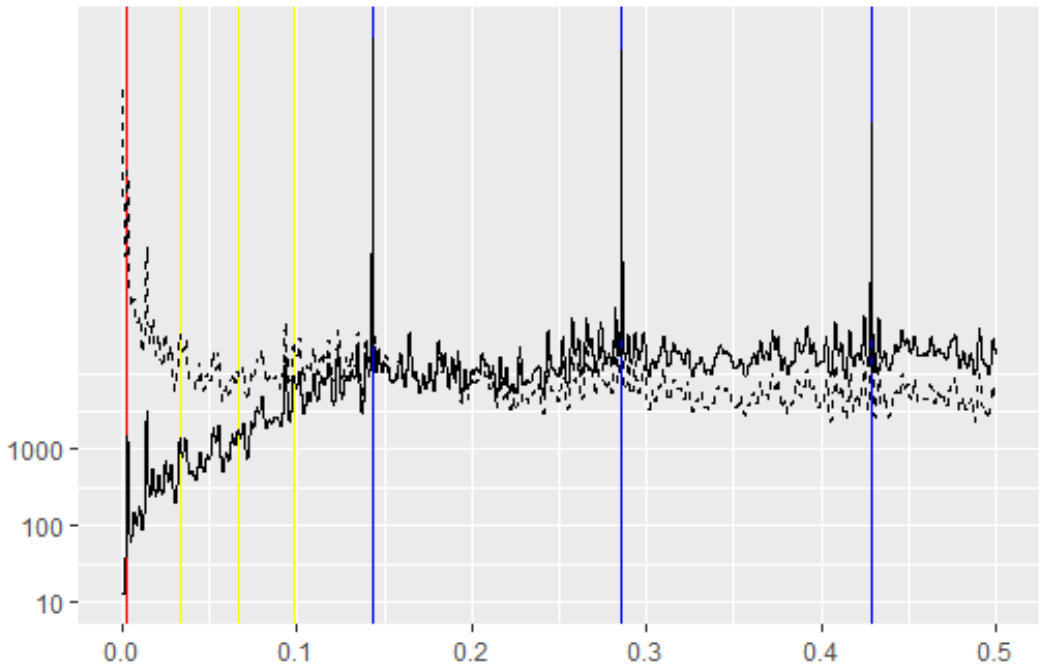
Périodicités infra-annuelles : multiples et non entières

	Périodes (nombre d'observations par cycle)			
Données	Jour	Semaine	Mois	Année
Trimestrielles				4
Mensuelles				12
Hebdomadaires			4.348125	52.1775
Journalières		7	30.436875	365.2425
Horaires	24	168	730.485	8765.82

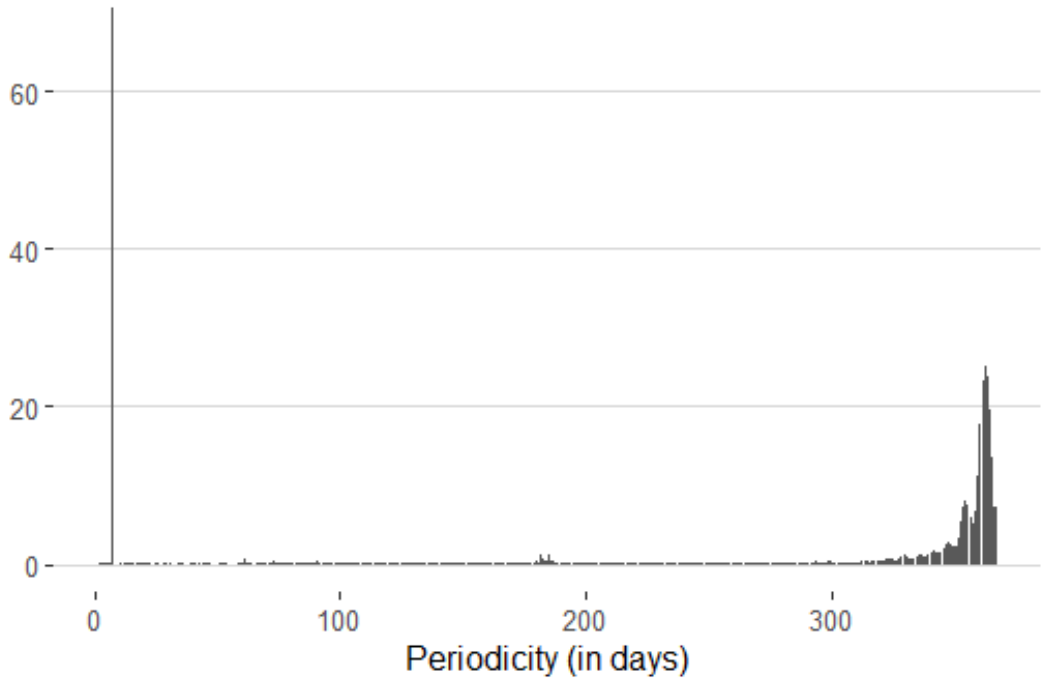
Une série journalière peut présenter 3 périodicités

- hebdomadaire ($p = 7$) : les lundis se ressemblent et sont différents des dimanches (DOW)
- intra-mensuelle ($p = 30.44$) : les derniers jours de chaque mois sont différents des premiers (DOM)
- une périodicité annuelle ($p = 365.25$) : d'une année à l'autre les 15/06 se ressemblent, les jours d'été se ressemblent...(DOY)

Spectre naissances quotidiennes



Test Canova-Hansen



Décomposition en composantes inobservables

série brute	=	Tendance	∘	Saisonnalité	∘	Calendrier	∘	Irrégulier
↓		↓		↓		↓		↓
y_t	=	t_t	∘	s_t	∘	c_t	∘	i_t
		↓		↓		↓		↓
		Variations de long terme		Variations infra-annuelles périodiques		Effets de Calendrier		Résidu

↘ ↙

Corrigée des variations saisonnières (CVS) data

- Décomposition Additive ($\circ = +$), multiplicative ($\circ = \times$)

Facteurs Saisonniers multiples

Nouvelle écriture pour les données haute-fréquence :

$$S_t = S_{t,7} \circ S_{t,30.44} \circ S_{t,365.25}$$

La décomposition se fera itérativement selon les différentes périodes, en commençant par la plus petite (plus haute fréquence) car

- c'est celle qui a en général les mouvements de plus forte ampleur et les plus stables
- les cycles des périodes les plus courtes peuvent se mélanger aux périodes les plus longues

Saisonnalité et effets de calendrier (1/2)

- Les effets de calendrier perturbent la comparaison de périodes de même type, c'est un effet déterministe que l'on corrige par régression
- la définition d'un effet de calendrier dépend de la fréquence des données et de la période selon la quelle on décompose
- pour des données mensuelles ou trimestrielles
  - effet type de jours (ex: nombre de jours ouvrables)
  - effet année bissextile (éventuellement)
  - les regresseurs sont des contrastes

Saisonnalité et effets de calendrier (2/2)

- pour des séries journalières
  - neutraliser la particularité des jours fériés (si pertinente) pour rendre les différents types de jours comparables..
  - ... lors de l'estimation de S_7
  - ... lors de l'estimation de $S_{365.25}$
  - si l'on distingue jours fériés fixes et mobiles : ambiguïté allocation des effets des jours fixes entre calendrier et $S_{365.25}$
  - les regresseurs sont des indicatrices de jours fériés

Plan

Linéarisation

Dans x13-Arima et Tramo-Seats

- étape de modélisation Reg-Arima
- pour enlever les effets déterministes : outliers et calendrier
- les outliers seront réinjectés dans la série CVS

Le modèle Reg-ARIMA s'écrit comme suit :

$$\left(Y_t - \sum \alpha_i X_{it}\right) \sim ARIMA(p,d,q)(P,D,Q)$$

Ces modèles s'écrivent avec des opérateurs retard saisonniers $B^s(y_t) = y_{t-s}$

Adaptation du modèle Airline (1/2)

Le modèle dit "Airline" ARIMA(0,1,1)(0,1,1), qui est le plus simple et le plus courant s'écrit pour des données mensuelles ou trimestrielles :

$$(1-B)(1-B^s)y_t = (1-\theta_1 B)(1-\theta_2 B^s)\epsilon_t \qquad \epsilon_t \sim \text{NID}(0,\sigma_\epsilon^2)$$

Pour des données haute fréquence :

- le modèle peut comporter plusieurs différenciations $\Delta_s = 1 - B^s$ et des polynômes retard en B^s avec s non entier
- on écrit $s = s' + \alpha$, avec α un nombre réel appartenant à l'intervalle $]0, 1[$ (par exemple $52.18 = 52 + 0.18$ périodicité annuelle pour des données hebdomadaires)

Adaptation du modèle Airline (2/2)

- Avec un développement limité au voisinage de 1 de $f(x) = x^\alpha$

$$\begin{array}{lcl} x^\alpha & = & 1 + \alpha(x-1) + \frac{\alpha(\alpha+1)}{2!}(x-1)^2 + \frac{\alpha(\alpha+1)(\alpha+2)}{3!}(x-1)^3 + \cdots \\ B^\alpha & \cong & (1-\alpha) + \alpha B \end{array}$$

- On obtient un écriture approchée de l'opérateur retard $B^{s+\alpha}$

$$B^{s+\alpha} \cong (1-\alpha)B^s + \alpha B^{s+1}$$

Modélisation de la série de naissances quotidiennes

deux périodicités $p_1 = 7$ et $p_2 = 365.25$

$$(1-B)(1-B^7)(1-B^{365.25)}(Y_t - \sum \alpha_i X_{it}) = (1-\theta_1 B)(1-\theta_2 B^7)(1-\theta_3 B^{365.25})\epsilon_t$$

$$\epsilon_t \sim \text{NID}(0, \sigma_\epsilon^2)$$

avec

$$1 - B^{365.25} = (1 - 0.75B^{365} - 0.25B^{366})$$

Linéarisation : paramétrages de la fonction

```
pre.mult<- rjd3highfreq::fractionalAirlineEstimation
                        (df_daily$log_births, # ici série en log
                  x = q, # q= calendrier
                  periods = 7, # approx de c(7,365.25)
                  ndiff = 2, ar = FALSE, mean = FALSE,
                  outliers = c("ao","wo"),
                  # WO compensation, LS peu crédible
                  criticalValue = 0, #determinée par l'algorithme
                  precision = 1e-9, approximateHessian = TRUE)

# définition des regresseurs de calendrier
```

avec le package {rjd3modelling}

Linéarisation : résultats

Variable	Coef	Coef_SE	Tstat
14jt	-0.12	0.00	-26.00
8mai	-0.15	0.01	-28.71
asc	-0.17	0.00	-38.72
01jan	-0.26	0.00	-52.87
e_mon	-0.19	0.00	-42.44
1mai	-0.12	0.00	-24.81
l_pen	-0.19	0.00	-42.64
15aou	-0.12	0.00	-26.29
1nov	-0.15	0.00	-33.75
11nov	-0.13	0.00	-27.52
25dec	-0.28	0.00	-55.91

Linéarisation : résultats

Variable	Coef	Coef_SE	Tstat
AO.1993-12-24	-0.19	0.03	-5.77
WO.2001-03-19	0.12	0.02	5.72
AO.1995-08-14	-0.19	0.03	-5.74
AO.1997-08-15	-0.18	0.03	-5.54
AO.1970-12-25	0.19	0.03	5.85
AO.2011-05-08	0.19	0.03	5.75
AO.2018-11-11	0.18	0.03	5.50
AO.2017-01-01	0.23	0.03	7.06
AO.1978-01-01	0.23	0.03	6.93
AO.1997-12-24	-0.18	0.03	-5.49
WO.2006-01-01	0.14	0.02	6.38
AO.1998-05-01	0.19	0.03	5.74

Plan

X-11 : généralités

- X-11 est le module de décomposition de X-13-Arima
- on décompose (S, T, I) la série linéarisée à l'étape précédente
- on procède période par période en commençant par la plus petite
- la structure globale des itérations est conservée
- adaptations pour traiter les périodicités non entières

Décomposition avec X-11 (1/2)

Une première estimation de la CVS à partir de la série linéarisée

1. Estimation de la Tendance par une moyenne mobile qui supprime la saisonnalité, avec un ordre égal à la périodicité :

$$T_t^{(1)} = M_p(X_t) = M_p(T_t + S_t + I_t)$$

2. Estimation de la composante saisonnier-irrégulier par différence

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

3. Estimation de la composante saisonnière par moyenne mobile 3×3 (p ex) sur chaque type de période

$$S_t^{(1)} = M_{3\times 3}\left[(S_t + I_t)^{(1)}\right]$$

4. Première estimation de la série corrigée des variations saisonnières

$$Xsa_t^{(1)} = X_t - S_t^{(1)}$$

Décomposition avec X-11 (2/2)

Une seconde estimation de la CVS :

1. Amélioration de l'estimation de la Tendance par MM de Henderson d'ordre q, à partir d'une série non saisonnière

$$T_t^{(2)} = H_q(Xsa_t^{(1)})$$

2. Estimation de la composante saisonnier-irrégulier

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

3. Estimation de la composante saisonnière par moyenne mobile 3×5 (généralement) sur chaque type de période:

$$S_t^{(2)} = M_{3\times 5}\left[(S_t + I_t)^{(2)}\right]$$

4. Estimation de la série corrigée des variations saisonnières :

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

X-11 utilise ces deux boucles plusieurs fois avec une correction de l'irrégulier entre les itérations

Adaptation du filtre de suppression de la saisonnalité

Pour la première estimation de la tendance, généralisation des moyennes mobiles centrées symétriques d'ordre égal à la période p;

- longueur l du filtre : plus petit entier impair supérieur à p
- ex : p=7, l=7, p=12 l=13, p=365,25, l=367, p=52.18 l=53
- coefficients centraux $1/p$ (1/12,1/7, 1/365.25)
- coefficients extrêmes $\mathbb{I}\{E(p) \text{ pair}\} + (p - E(p))/2p$
- ex : p=12 (1/12 et 1/24) (on a bien $M_{2\times 12}$ du cas mensuel classique)
- ex : p=365.25 (1/365.25 et 0.25/(2*365.25)

Adaptation du filtre d'extraction de la saisonnalité (1/2)

le filtrage se fait type de période par type de période
exemple $M_{3\times3}$

$$M_{3\times3}X = \frac{1}{9}(X_{t-2p}) + \frac{2}{9}(X_{t-p}) + \frac{3}{9}(X_t) + \frac{2}{9}(X_{t+p}) + \frac{1}{9}(X_{t+2p})$$

si p entier, rien à changer
si p pas entier on utilise l'approximation de Taylor du polynôme retard

$$B^{s+\alpha} \cong (1-\alpha)B^s + \alpha B^{s+1}$$

Adaptation du filtre d'extraction de la saisonnalité (2/2)

par exemple, pour $p = 30.44$ un filtre 3×3 s'écrit:

$$\begin{aligned}
\hat{s}_t &= \frac{1}{9}\left[0.88 \times (\widehat{si})_{t-61} + 0.12 \times (\widehat{si})_{t-60}\right] \\
&+ \frac{2}{9}\left[0.44 \times (\widehat{si})_{t-31} + 0.56 \times (\widehat{si})_{t-30}\right] \\
&+ \frac{3}{9}(\widehat{si})_t \qquad (1) \\
&+ \frac{2}{9}\left[0.56 \times (\widehat{si})_{t+30} + 0.44 \times (\widehat{si})_{t+31}\right] \\
&+ \frac{1}{9}\left[0.12 \times (\widehat{si})_{t+60} + 0.88 \times (\widehat{si})_{t+61}\right]
\end{aligned}$$

avantage de cette approximation : pas d'imputation de données

Adaptation du filtre d'estimation finale de la tendance

La saisonnalité a été supprimée à la première étape, il n'y a plus de problème de périodicité lors de l'estimation finale de la tendance. Toutefois l'algorithme X-11 a été enrichi dans {rjd3hihgfreq} :

- X-11 d'origine : filtres de Henderson (+ Musgrave asymétriques pour les fins de séries)
- dans{rjd3highfreq} généralisation de cette méthode : approximation polynomiale locale de la série avec différentes distributions de poids

Extended X-11 pour p=7 : paramétrages

```
x11.dow <- rjd3highfreq::x11(exp(pre.mult$model$linearized),
        period = 7,                          # DOW pattern
        mul = TRUE,
        trend.horizon = 9,  # 1/2 Filter length : not too long vs p
        trend.degree = 3,                                # Polynomial degree
        trend.kernel = "Henderson",                      # Kernel function
        trend.asymmetric = "CutAndNormalize",            # Truncation method
        seas.s0 = "S3X9", seas.s1 = "S3X9",              # Seasonal filters
        extreme.lsig = 1.5, extreme.usig = 2.5)          # Sigma-limits
```

Extended X-11 pour p=365.25 : paramétrages

```
x11.doy <- rjd3highfreq::x11(x11.dow$decomposition$sa,  # previous sa
                             period = 365.2425,          # DOY pattern
                             mul = TRUE,
                             trend.horizon = 371, # 1/2 final filter length
                             trend.degree = 3,
                             trend.kernel = "Henderson",
                             trend.asymmetric = "CutAndNormalize",
                             seas.s0 = "S3X15", seas.s1 = "S3X5",
                             extreme.lsig = 1.5, extreme.usig = 2.5)
```

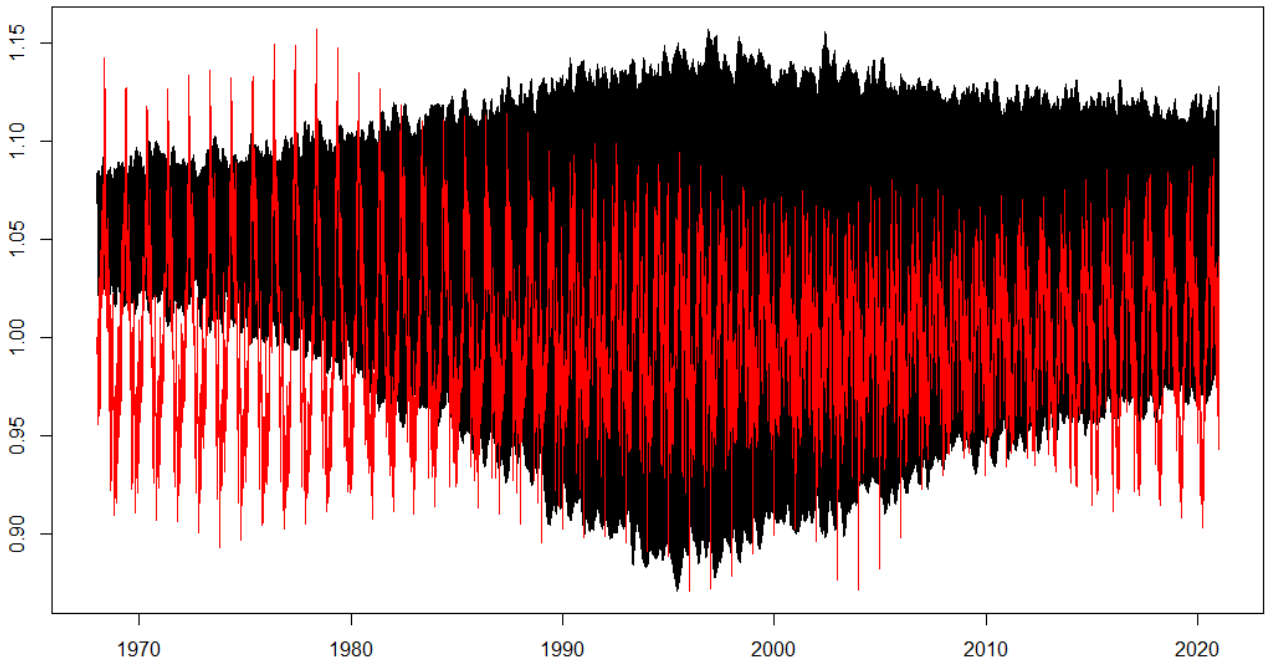
Problème non encore résolu : critères pour déterminer la longueur des filtres et les seuils

Résumé des opérations

- linéarisation : $Y_{lin} = FracAirline(Y)$, calcul de Y_{cal}
- $cvs_7 = X11_7(Y_{lin})$, calcul de S_7
- $cvs_{365.25} = X11_{365.25}(cvs_7)$, calcul de $S_{365.25}$
- $cvs_{finale} = Y_{cal}/S_7/S_{365.25}$

Décomposition de la série des naissances (1/2)

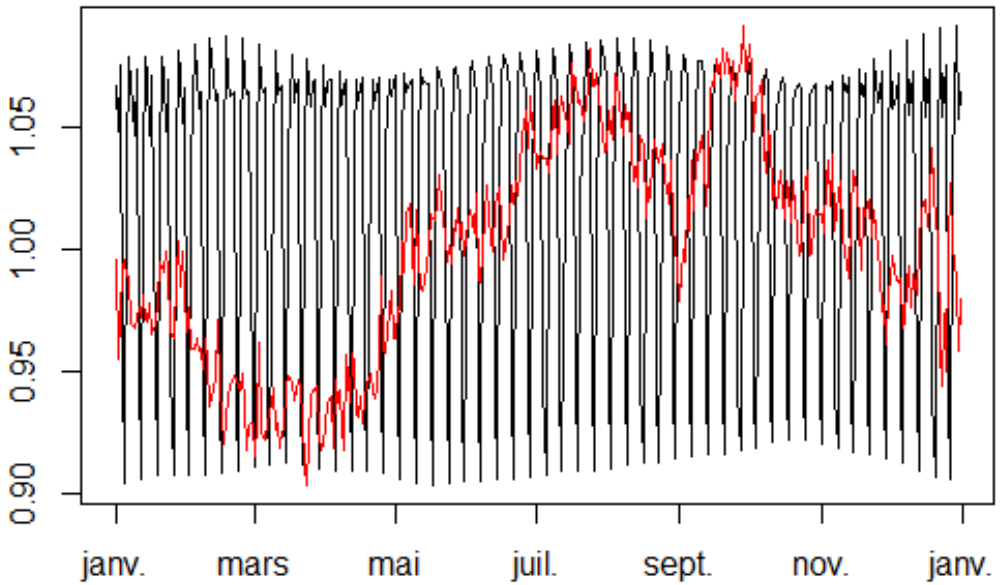
Facteurs saisonniers estimés : p=7 (noir) et p=365.25 (rouge)



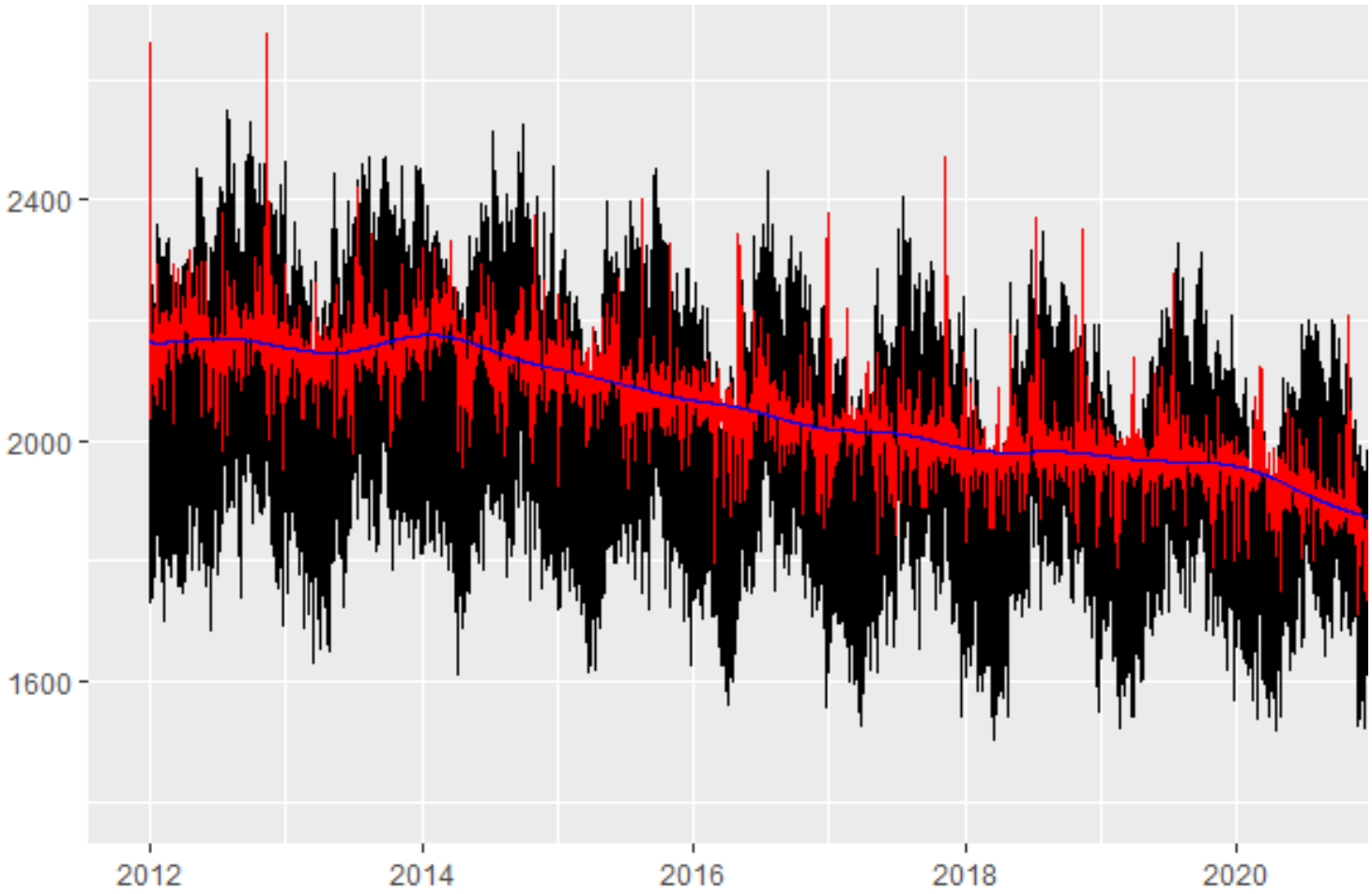
saisonnalité évolutive sur longue période

Décomposition de la série des naissances (2/2)

Zoom sur l'année 2000 : Facteurs saisonniers estimés : p=7 (noir) et p=365.25 (rouge)



Naissances : brute, cvs et tendance



Plan

Extended SEATS

Décomposition paramétrique (AMB, extension de Seats) également possible avec {rjd3highfreq}
Paramétrage 2ème étape de décomposition :

```
amb.doy <- rjd3highfreq::fractionalAirlineDecomposition(
  amb.dow$decomposition$sa,   # DOW-adjusted linearised data
  period = 365.2425,          # DOY pattern
  sn = FALSE,                 # Signal (SA)-noise decomposition
  stde = FALSE,               # Calculate standard deviations
  nbcasts = 0, nfcasts = 0)   # Numbers of back- and forecasts
```

AMB filtres optimaux par construction

Résultats légèrement différents de X-11 (comme pour basses fréquences...)

Plan

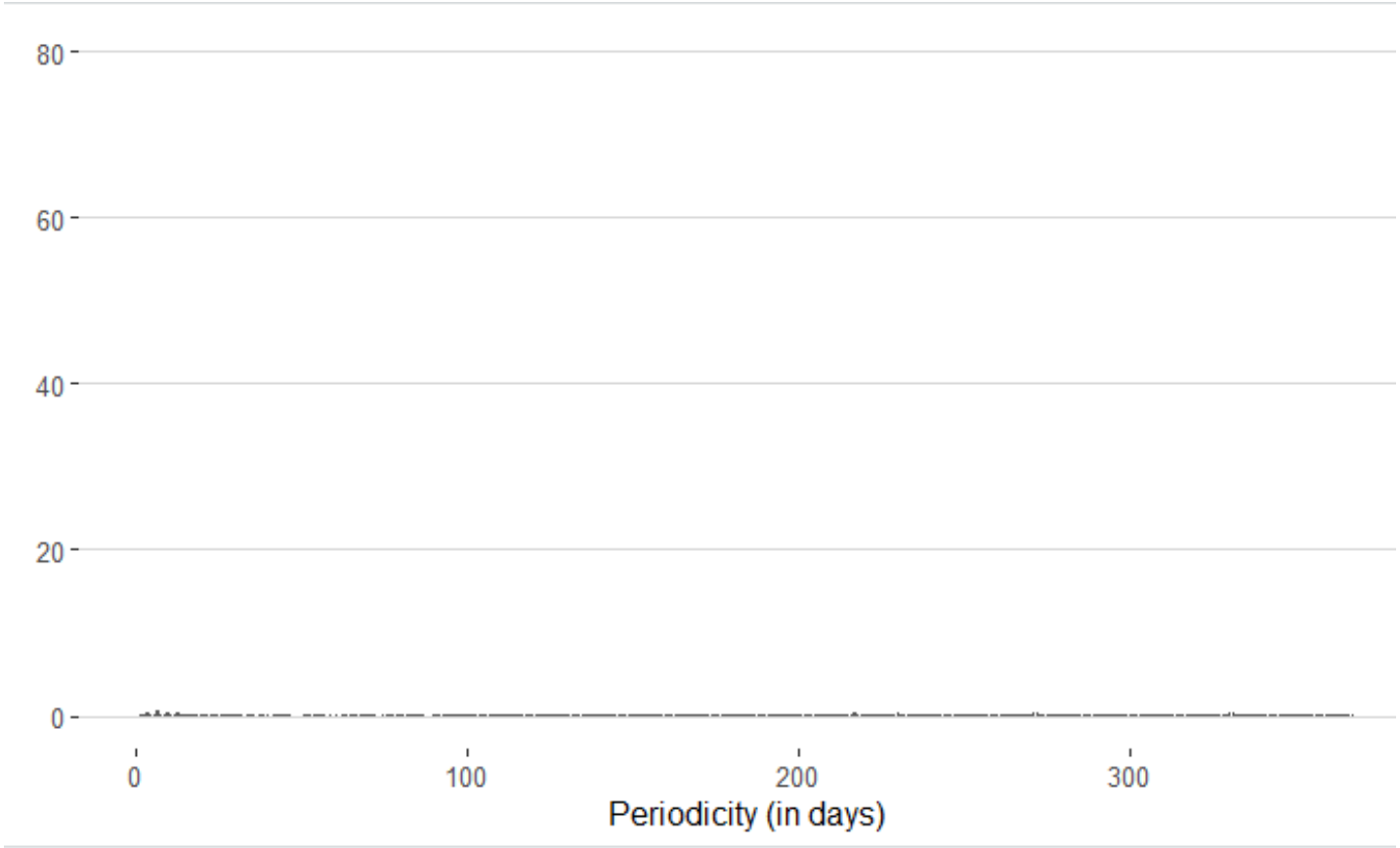
Saisonnalité résiduelle (1/2)

On souhaite vérifier que la série cvs-cjo n'est plus saisonnière, pour toutes les périodicités traitées

- tests de saisonnalité usuels pas forcément transposables (ex : si Anova, problèmes avec périodicités multiples et non entières)
- test spectraux : distributions souvent non calculées

Saisonnalité résiduelle (2/2)

Test de Canova-Hansen sur CVS finale avec X-11



Plan

Conclusion

- difficultés pour désaisonnaliser des données HF : périodicités multiples et non entières
- JDemetra+ v3.0 propose plusieurs algorithmes adaptés : X-13-Arima, Tramo-Seats, vus ici, mais aussi STL et modèles structurels (BSM)
- une interface graphique (prototype) est également disponible
- les algorithmes adaptés pour les données HF sont en évolution depuis plusieurs années
- manquent encore : tests, critères pour choix filtres

Merci de votre attention

- {rjd3highfreq} : `https://github.com/palatej`
- interface graphique (prototype) : `https://github.com/nnbrd`
- article détaillé à venir en octobre (JSM 2022, proceedings)