Séries à haute fréquence

Principaux défis et traitement dans JDemetra+ 3.0

jean.palate@nbb.be

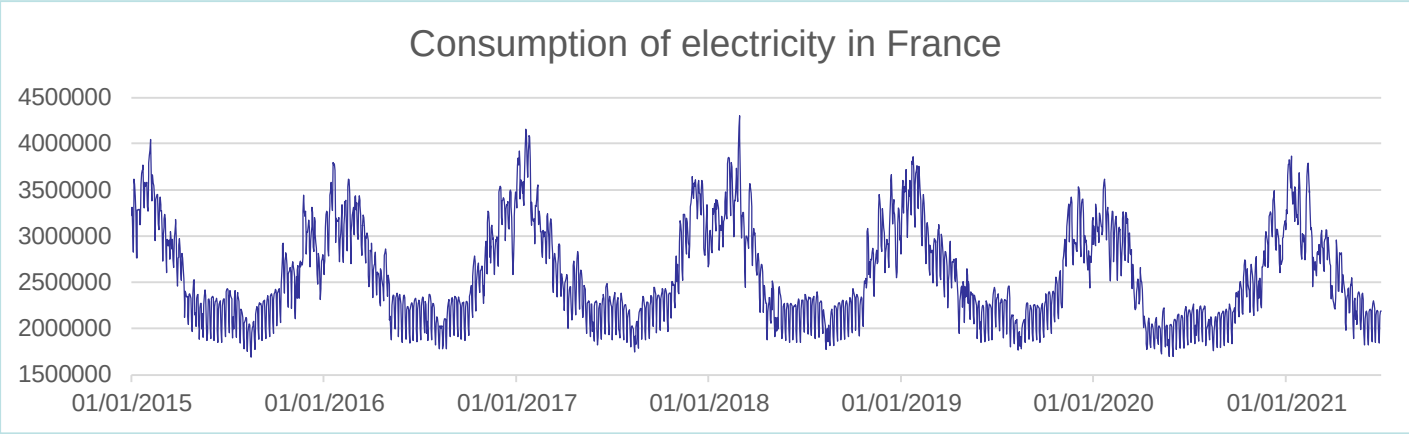
Plan

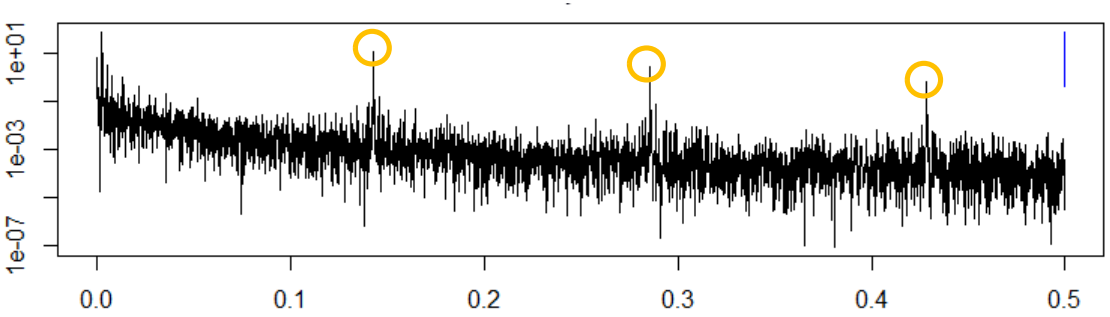
- Motivations
- Exemples de séries à haute fréquence
- Principaux défis
- Implémentation
  - Transposition des algorithmes actuels
    - RegArima, décomposition canonique, décomposition non paramétrique
    - Exemples, constatations
- Recherche en cours
- Modules disponibles

0. Motivations

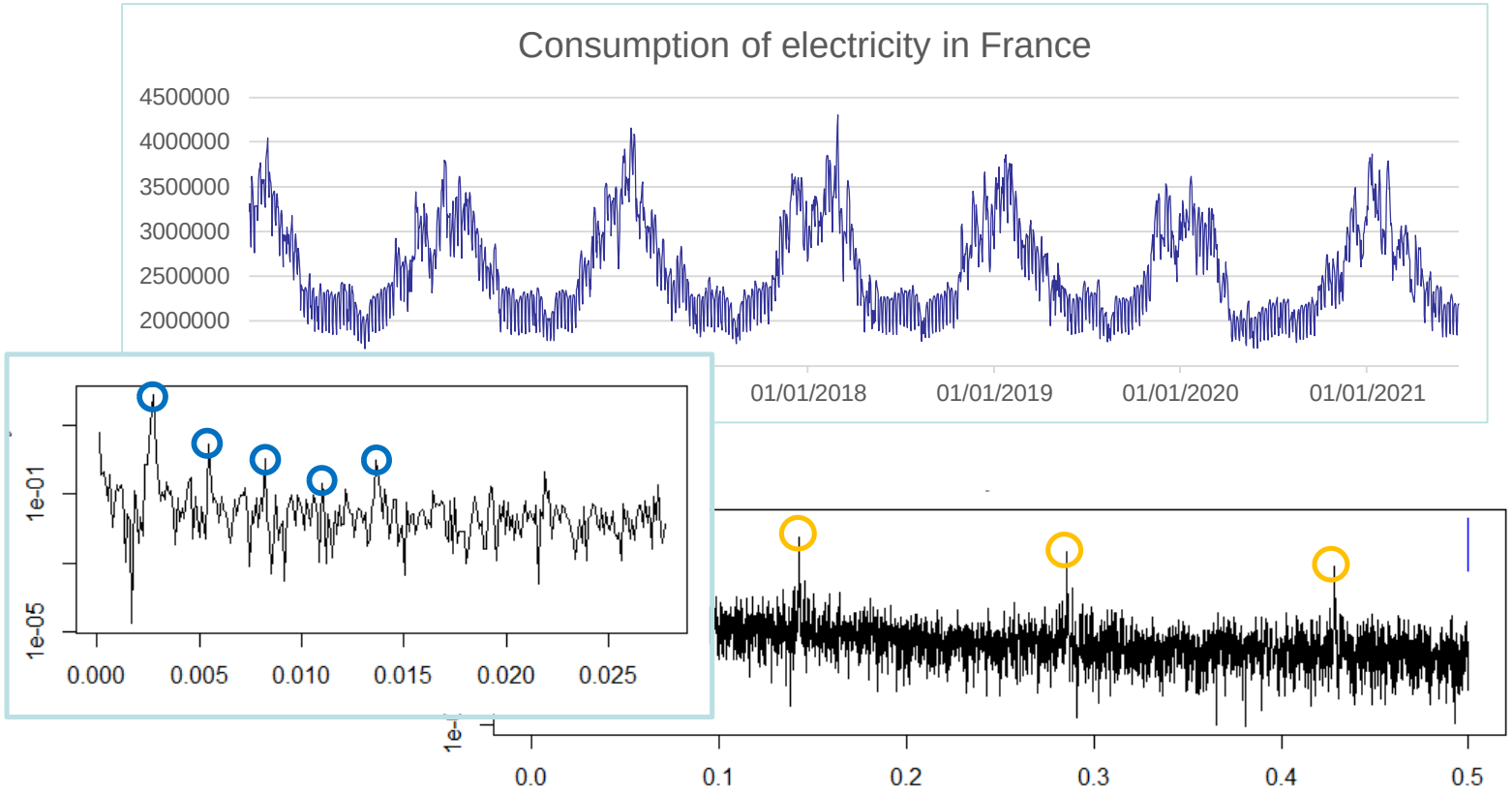
- Forte demande au moment de la crise du Covid
  - Données journalières/hebdomadaires
  - Analyse de l'évolution à très court terme
- Augmentation des données quasi en temps réel
  - → nowcasting
- Meilleure compréhension des séries mensuelles/trimestrielles (en particulier effets de calendrier)

1. Exemple I (série journalière)

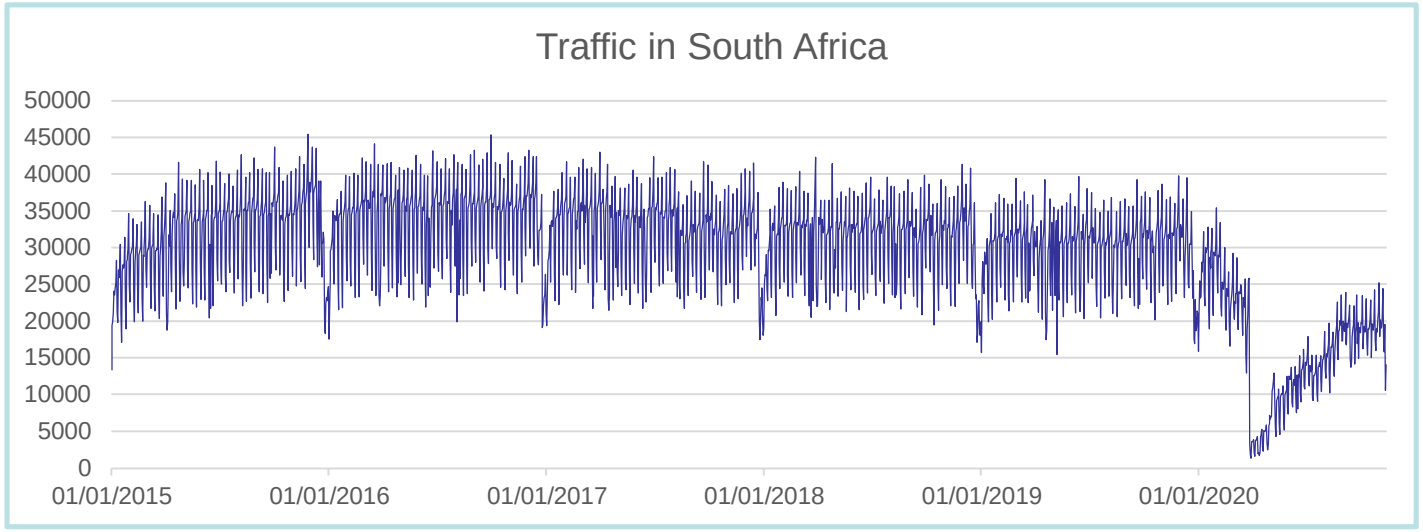


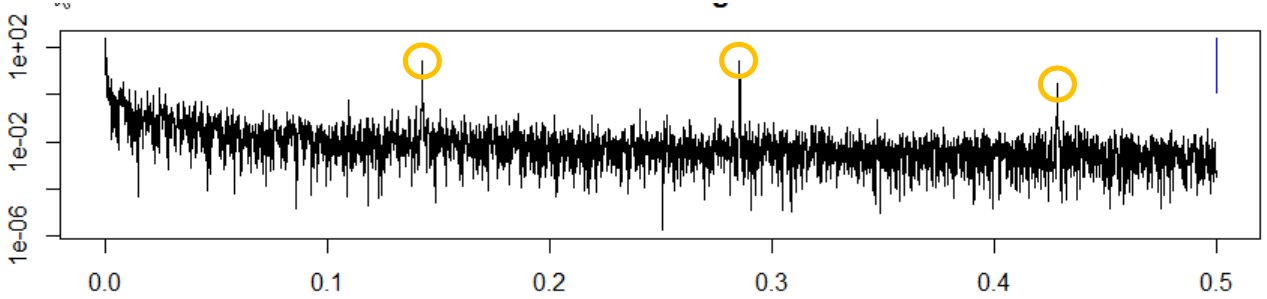


1. Exemple I (série journalière)

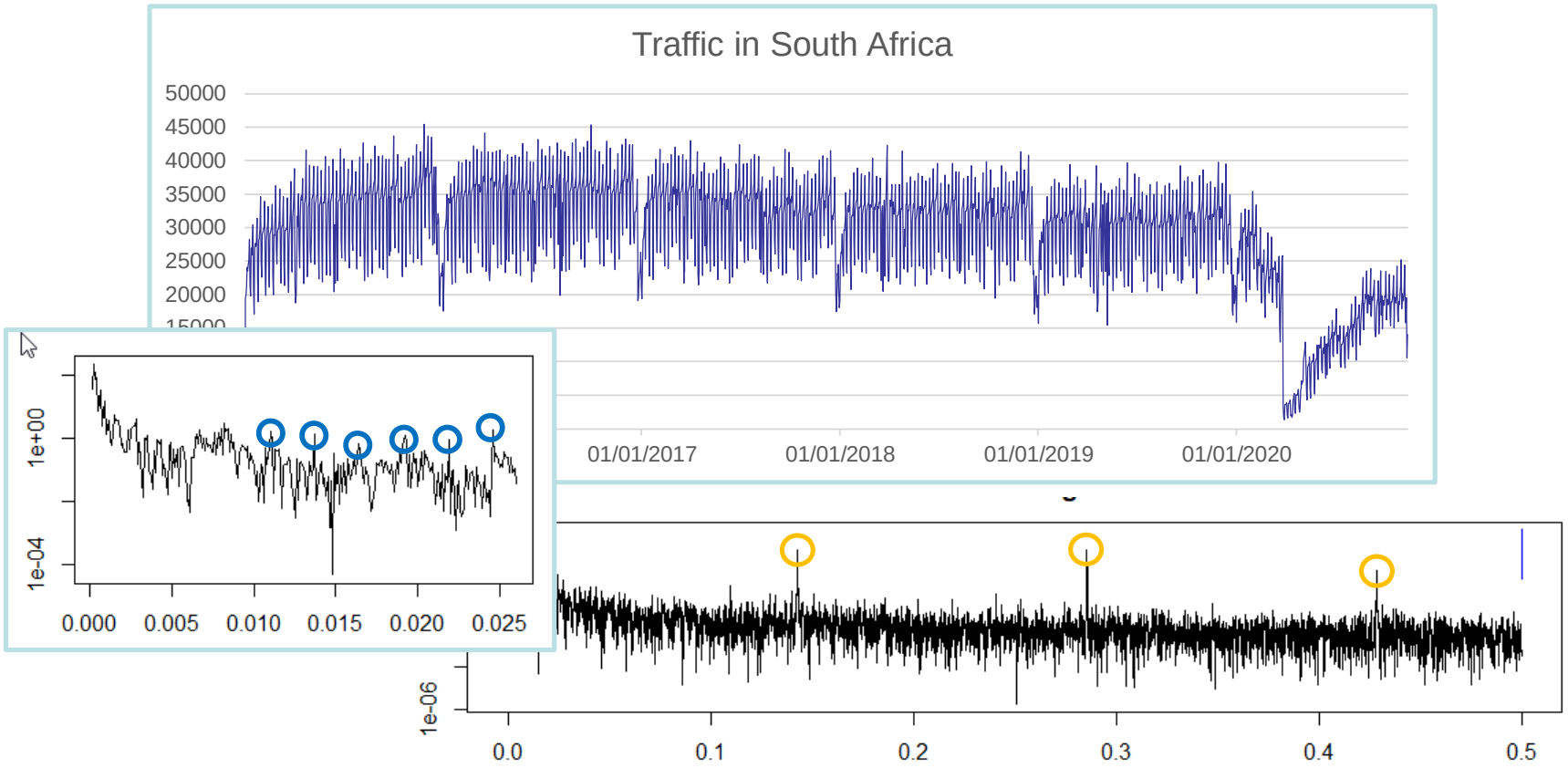


1. Exemple II (série journalière)



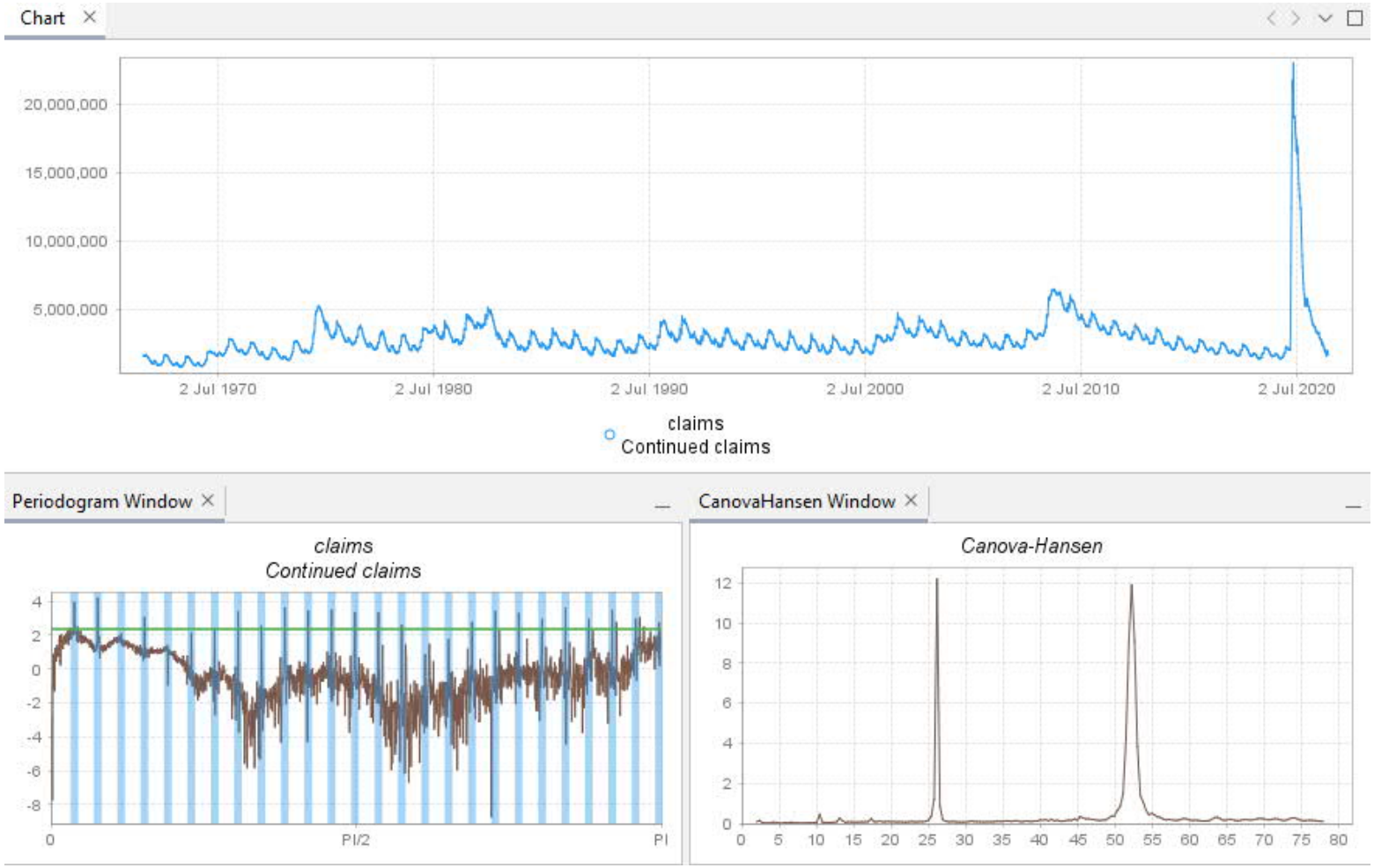


1. Exemple II (série journalière)



1. Exemple III (série hebdomadaire)

US claims (continued)

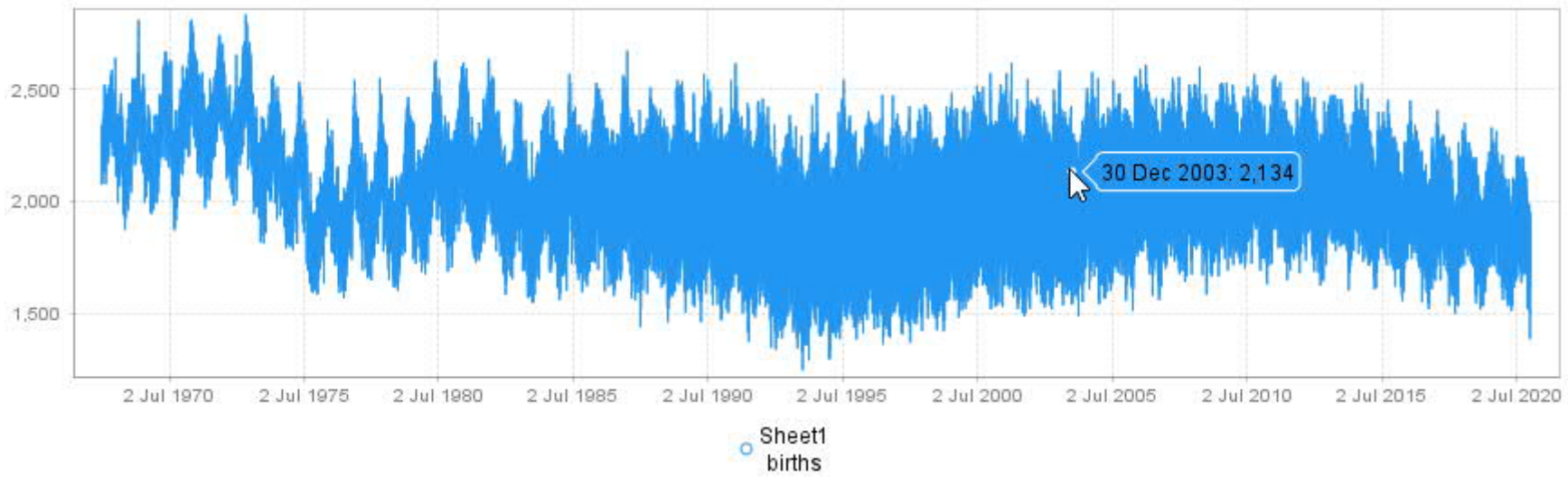


2. Principaux défis (I)

- Grande variété des données
  - Fréquence des observations, séries continues, données uniquement les jours ouvrés…
- [Multiple] périodicités, entières ou non
  - Journalière, hebdomadaire, intra-mensuelle, annuelle…
- Définition de la “saisonnalité”
  - Séparation des effets calendriers et des effets saisonniers, séparation des effets intra-mensuels et des effets annuels, régularité des effets saisonniers…

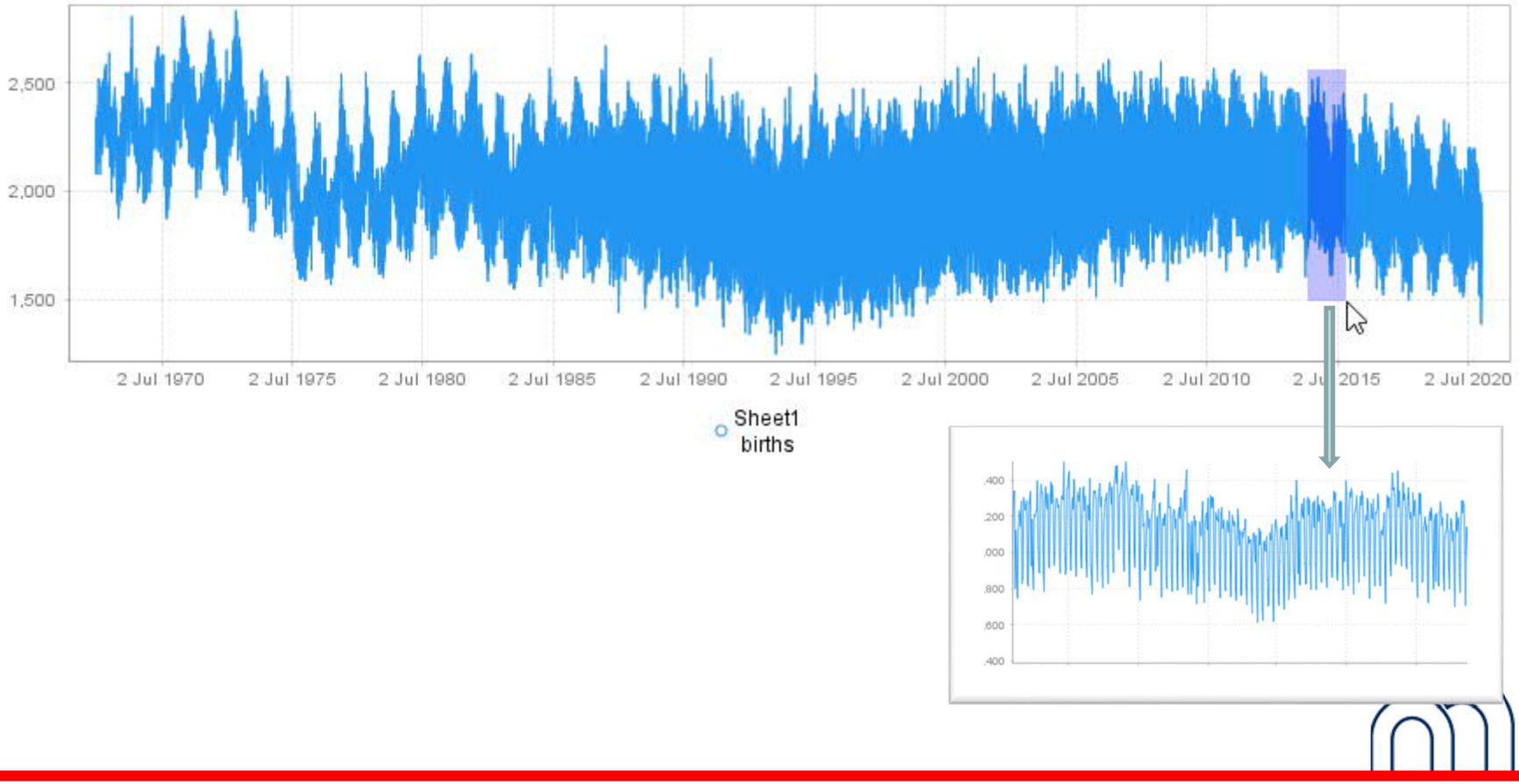
2. Principaux défis (II)

- Volume des données
  - Traitement, présentation…



2. Principaux défis (II)

- Volume des données
  - Traitement, présentation…

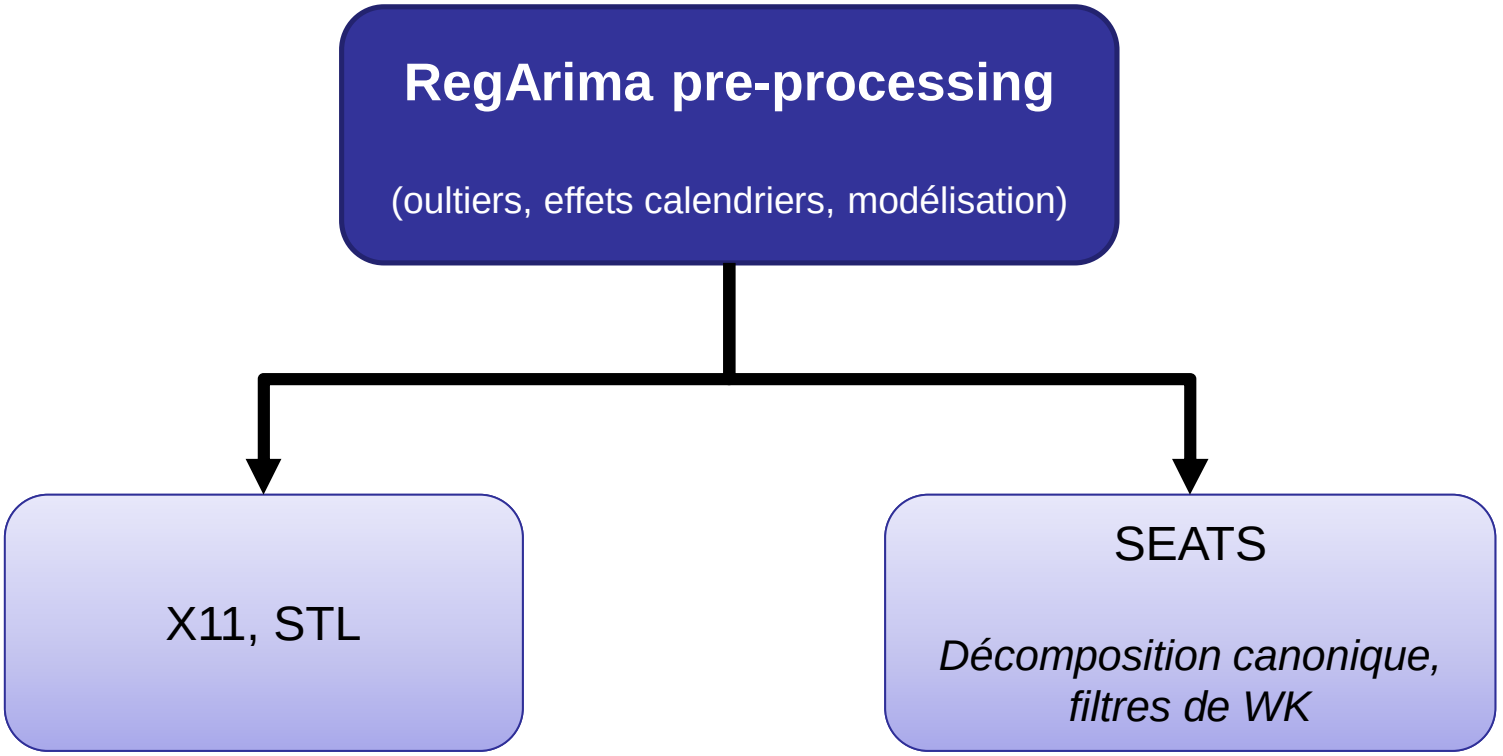


2. Principaux défis (II)

- Algorithmes
  - Stratégies de décomposition
    - Transposition des algorithmes existants (X13, Tramo-Seats, STL, modèles structurels…)
    - Autres approches?
  - Adaptation du traitement des effets calendriers
  - Adaptation des algorithmes aux contraintes de périodicités et de volume

- Travail en cours…

3. Transposition des algorithmes actuels



3.1 Pre-processing de type RegArima

- Approche similaire à la solution actuelle (Tramo, X13)
- Importance des variables de calendrier
- Détection automatique des outliers
  - Similaire à Tramo (+ rapide)
  - Switch outlier (0 1 -1 0)
- Modélisation ARIMA « simple » (extension du modèle airline aux périodicités non entières et/ou multiples
- Pas d'identification automatique du modèle

3.11 Modèles ARIMA (I)

- Airline “fractionnaire”
  - Notations

$$\Delta_s y_t = y_t - y_{t-s}$$
$$\Theta_{s,\theta}\varepsilon_t = \varepsilon_t - \theta\varepsilon_{t-s}$$
$$\tilde{\Delta}_{s+\alpha} y_t = y_t - (1-\alpha) y_{t-s} - \alpha\, y_{t-s-1}$$
$$\tilde{\Theta}_{s+\alpha,\theta}\varepsilon_t = \varepsilon_t - (1-\alpha)\theta\varepsilon_{t-s} - \alpha\theta\varepsilon_{t-s-1}$$
$$s \in \mathbb{N}\ and\ \alpha \in [0,1[$$

  - “Fractional airline” (séries hebdomadaires)

$$\Delta_1\Delta_{52.18} y_t = \Theta_{1,\theta}\Theta_{52.18,\theta_w}\varepsilon_t$$

$$\Delta_{52.18} y_t = y_t - (0.82 \cdot y_{t-52} + 0.18 \cdot y_{t-53})$$
$$= (1-B)(1 + B + \cdots + B^{51} + 0.18 \cdot B^{52}) y_t$$

3.11 ARIMA models (II)

- Multiples périodicités
  - $\Delta_1 \Delta_{p_1} \cdots \tilde{\Delta}_{p_n} y_t = \Theta_{1,\theta_1} \Theta_{p_1,\theta_{p_1}} \cdots \tilde{\Theta}_{p_n,\theta_{p_n}} \varepsilon_t$
  - $\Delta_1^k S_{p_1} \cdots \tilde{S}_{p_n} y_t = \Theta_{1,\theta_1} \Theta_{p_1,\theta_{p_1}} \cdots \tilde{\Theta}_{p_n,\theta_{p_n}} \varepsilon_t$

  With: $S_p y_t = y_t + \cdots + y_{t-s+1}$

  $\tilde{S}_{p+\alpha} y_t = y_t + \cdots + y_{t-s+1} + \alpha\, y_{t-s}$
- Exemples
  - W model: $\Delta_1 \Delta_7 (y_t - X_t \beta) = \Theta_{1,\theta_1} \Theta_{7,\theta_2} \varepsilon_t$, $\varepsilon_t \sim N(0, \sigma^2)$
  - YW(k) model: $\Delta_1^{k \leq 3} S_7 \tilde{S}_{365.25} (y_t - X_t \beta) = \Theta_{1,\theta_1} \Theta_{7,\theta_2} \tilde{\Theta}_{365.25,\theta_3} \varepsilon_t$

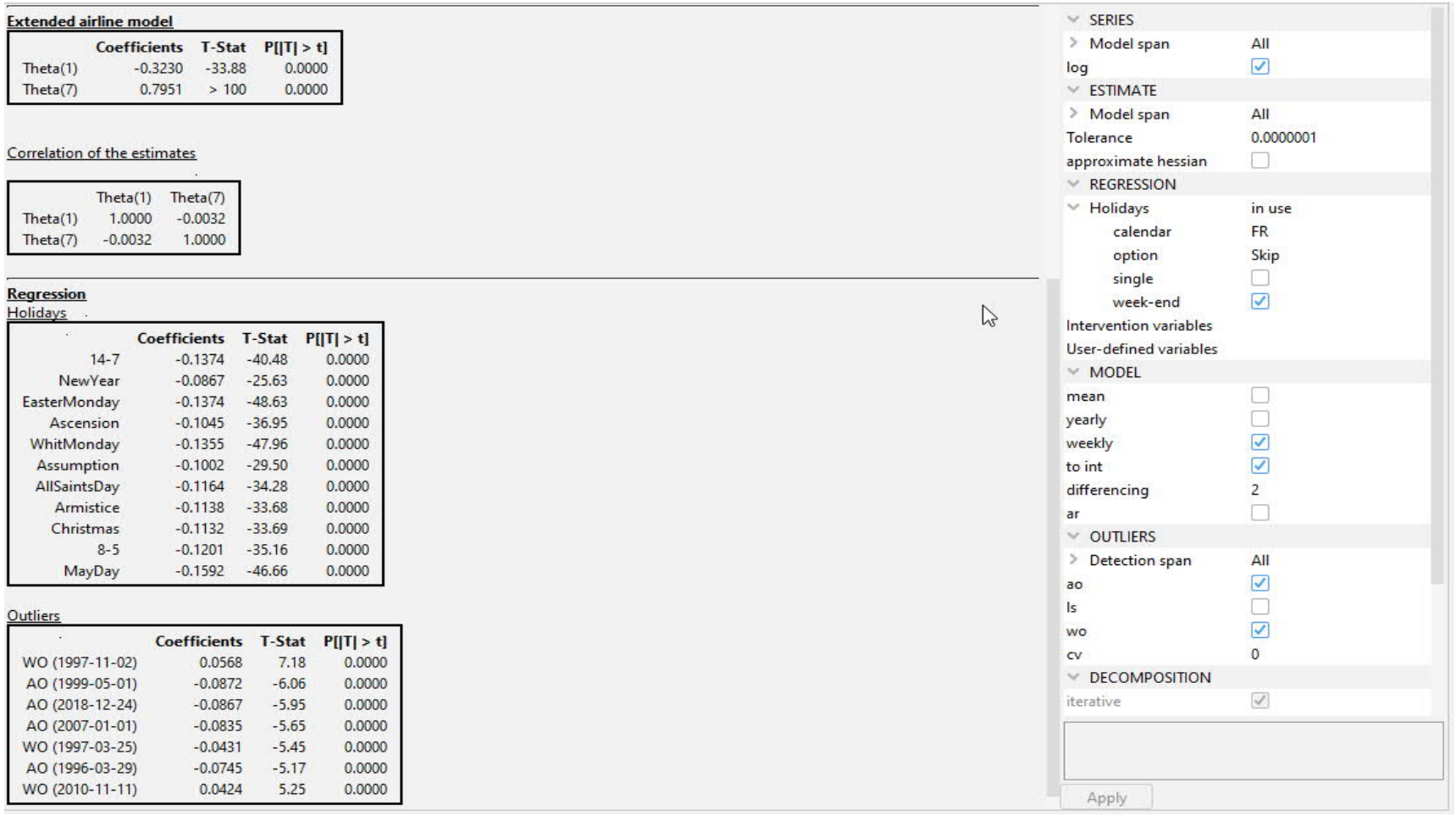
3.12 Estimation des modèles RegArima

- Approche comparable à celle utilisée dans Tramo
  - Décomposition du modèle en parties stationnaire/non stationnaire ⚠
  - Différenciation de la séries / variables explicatives
  - Application du filtre de Kalman (récursions de Chandrasekhar, rapides et suffisamment stables)
  - Calcul de la vraisemblance(σ^2, β), conditionnellement aux observations initiales
- Performance très acceptable
  - L'algorithme est
    - Linéaire en terme de la longueur de la série
    - Quasi-linéaire en terme de la dimension du modèle « state-space » *(dim=8 pour modèle W, =374 pour modèle YW)*

3.13 Constatations/conseils

- Utiliser un modèle simple pour la modélisation ARIMA
  - « airline » hebdomadaire pour les séries journalières
    - Beaucoup plus performant
    - Peu d'impact sur les coefficients des variables de régression (par rapport à un modèle complexe)
    - Meilleure identification des effets calendriers (pas d'interférence avec les effets saisonniers annuels)
- Importance dans la plupart des séries des effets calendriers
  - Pas nécessairement observables après agrégation
  - Nécessité d'une modélisation adéquate
  - Hypothèse de coefficients constants pas crédible sur longues séries *(→ state space avec coefficients variables, de type random walk)*

3.14 Consommation d’électricité



Extended airline model

	Coefficients	T-Stat	P[|T| > t]
Theta(1)	-0.3230	-33.88	0.0000
Theta(7)	0.7951	> 100	0.0000

Correlation of the estimates

	Theta(1)	Theta(7)
Theta(1)	1.0000	-0.0032
Theta(7)	-0.0032	1.0000

Regression

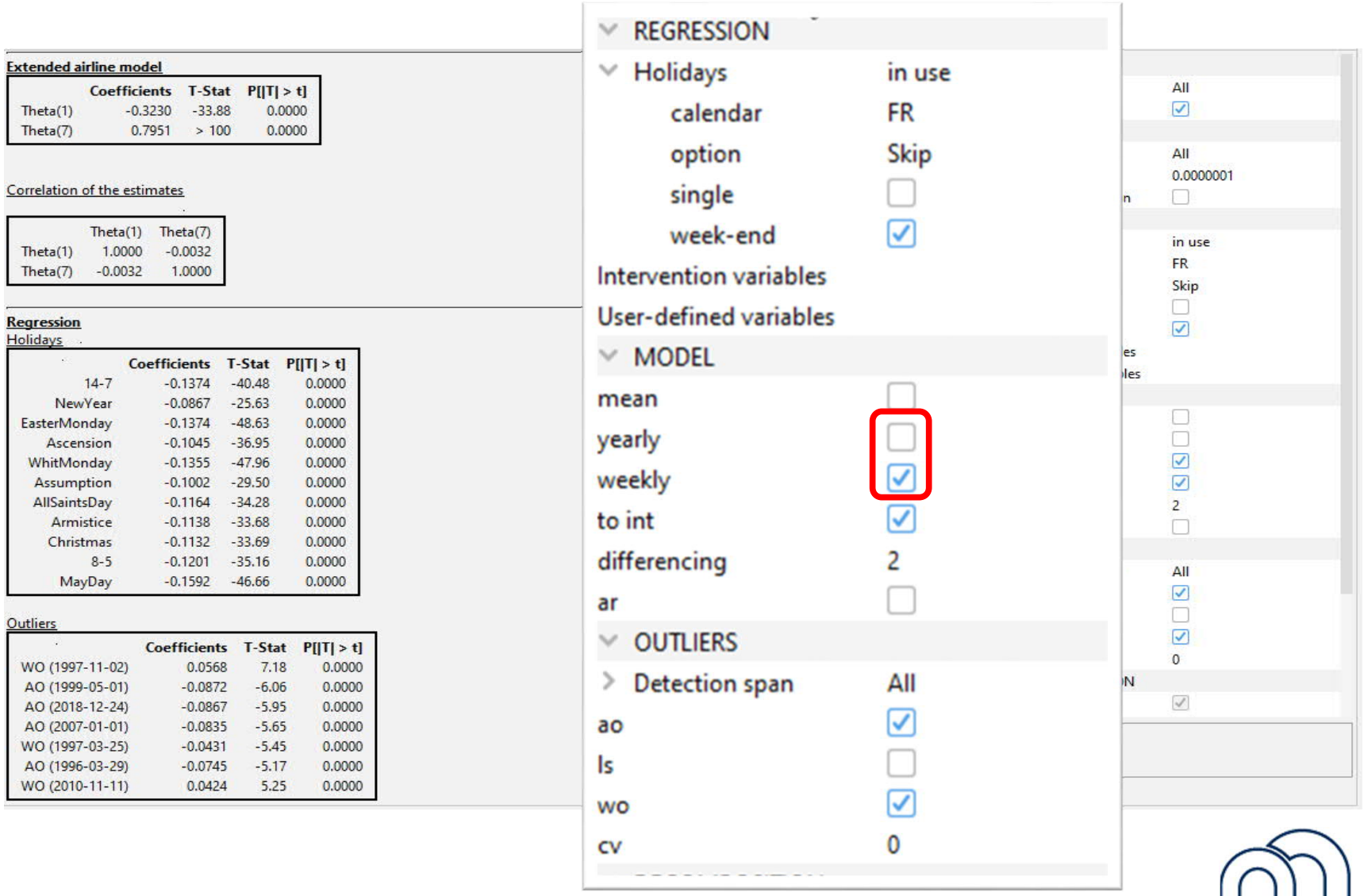
Holidays

	Coefficients	T-Stat	P[|T| > t]
14-7	-0.1374	-40.48	0.0000
NewYear	-0.0867	-25.63	0.0000
EasterMonday	-0.1374	-48.63	0.0000
Ascension	-0.1045	-36.95	0.0000
WhitMonday	-0.1355	-47.96	0.0000
Assumption	-0.1002	-29.50	0.0000
AllSaintsDay	-0.1164	-34.28	0.0000
Armistice	-0.1138	-33.68	0.0000
Christmas	-0.1132	-33.69	0.0000
8-5	-0.1201	-35.16	0.0000
MayDay	-0.1592	-46.66	0.0000

Outliers

	Coefficients	T-Stat	P[|T| > t]
WO (1997-11-02)	0.0568	7.18	0.0000
AO (1999-05-01)	-0.0872	-6.06	0.0000
AO (2018-12-24)	-0.0867	-5.95	0.0000
AO (2007-01-01)	-0.0835	-5.65	0.0000
WO (1997-03-25)	-0.0431	-5.45	0.0000
AO (1996-03-29)	-0.0745	-5.17	0.0000
WO (2010-11-11)	0.0424	5.25	0.0000

3.14 Consommation d’électricité (journ.)



Extended airline model

	Coefficients	T-Stat	P[|T| > t]
Theta(1)	-0.3230	-33.88	0.0000
Theta(7)	0.7951	> 100	0.0000

Correlation of the estimates

	Theta(1)	Theta(7)
Theta(1)	1.0000	-0.0032
Theta(7)	-0.0032	1.0000

Regression

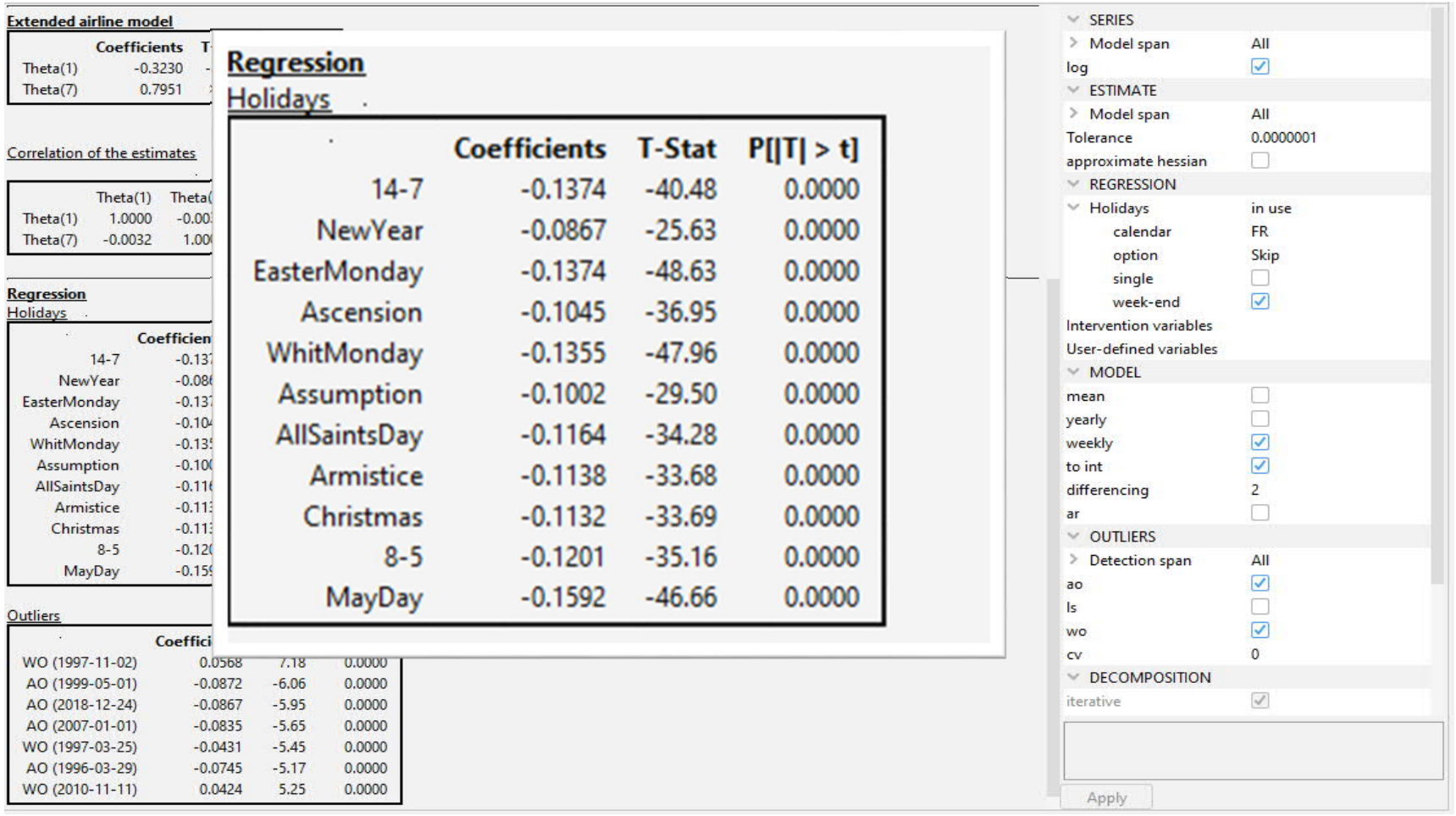
Holidays

	Coefficients	T-Stat	P[|T| > t]
14-7	-0.1374	-40.48	0.0000
NewYear	-0.0867	-25.63	0.0000
EasterMonday	-0.1374	-48.63	0.0000
Ascension	-0.1045	-36.95	0.0000
WhitMonday	-0.1355	-47.96	0.0000
Assumption	-0.1002	-29.50	0.0000
AllSaintsDay	-0.1164	-34.28	0.0000
Armistice	-0.1138	-33.68	0.0000
Christmas	-0.1132	-33.69	0.0000
8-5	-0.1201	-35.16	0.0000
MayDay	-0.1592	-46.66	0.0000

Outliers

	Coefficients	T-Stat	P[|T| > t]
WO (1997-11-02)	0.0568	7.18	0.0000
AO (1999-05-01)	-0.0872	-6.06	0.0000
AO (2018-12-24)	-0.0867	-5.95	0.0000
AO (2007-01-01)	-0.0835	-5.65	0.0000
WO (1997-03-25)	-0.0431	-5.45	0.0000
AO (1996-03-29)	-0.0745	-5.17	0.0000
WO (2010-11-11)	0.0424	5.25	0.0000

3.14 Consommation d’électricité



Extended airline model

	Coefficients	T
Theta(1)	-0.3230	
Theta(7)	0.7951	

Correlation of the estimates

	Theta(1)	Theta(
Theta(1)	1.0000	-0.00
Theta(7)	-0.0032	1.00

Regression

Holidays

	Coefficien
14-7	-0.13
NewYear	-0.08
EasterMonday	-0.13
Ascension	-0.10
WhitMonday	-0.13
Assumption	-0.10
AllSaintsDay	-0.11
Armistice	-0.11
Christmas	-0.11
8-5	-0.12
MayDay	-0.15

Outliers

	Coeffici		
WO (1997-11-02)	0.0568	7.18	0.0000
AO (1999-05-01)	-0.0872	-6.06	0.0000
AO (2018-12-24)	-0.0867	-5.95	0.0000
AO (2007-01-01)	-0.0835	-5.65	0.0000
WO (1997-03-25)	-0.0431	-5.45	0.0000
AO (1996-03-29)	-0.0745	-5.17	0.0000
WO (2010-11-11)	0.0424	5.25	0.0000

Regression

Holidays

	Coefficients	T-Stat	P[|T| > t]
14-7	-0.1374	-40.48	0.0000
NewYear	-0.0867	-25.63	0.0000
EasterMonday	-0.1374	-48.63	0.0000
Ascension	-0.1045	-36.95	0.0000
WhitMonday	-0.1355	-47.96	0.0000
Assumption	-0.1002	-29.50	0.0000
AllSaintsDay	-0.1164	-34.28	0.0000
Armistice	-0.1138	-33.68	0.0000
Christmas	-0.1132	-33.69	0.0000
8-5	-0.1201	-35.16	0.0000
MayDay	-0.1592	-46.66	0.0000

SERIES
- Model span: All
- log: ☑
ESTIMATE
- Model span: All
- Tolerance: 0.0000001
- approximate hessian: ☐
REGRESSION
- Holidays: in use
  - calendar: FR
  - option: Skip
  - single: ☐
  - week-end: ☑
- Intervention variables
- User-defined variables
MODEL
- mean: ☐
- yearly: ☐
- weekly: ☑
- to int: ☑
- differencing: 2
- ar: ☐
OUTLIERS
- Detection span: All
- ao: ☑
- ls: ☐
- wo: ☑
- cv: 0
DECOMPOSITION
- iterative: ☑

Apply

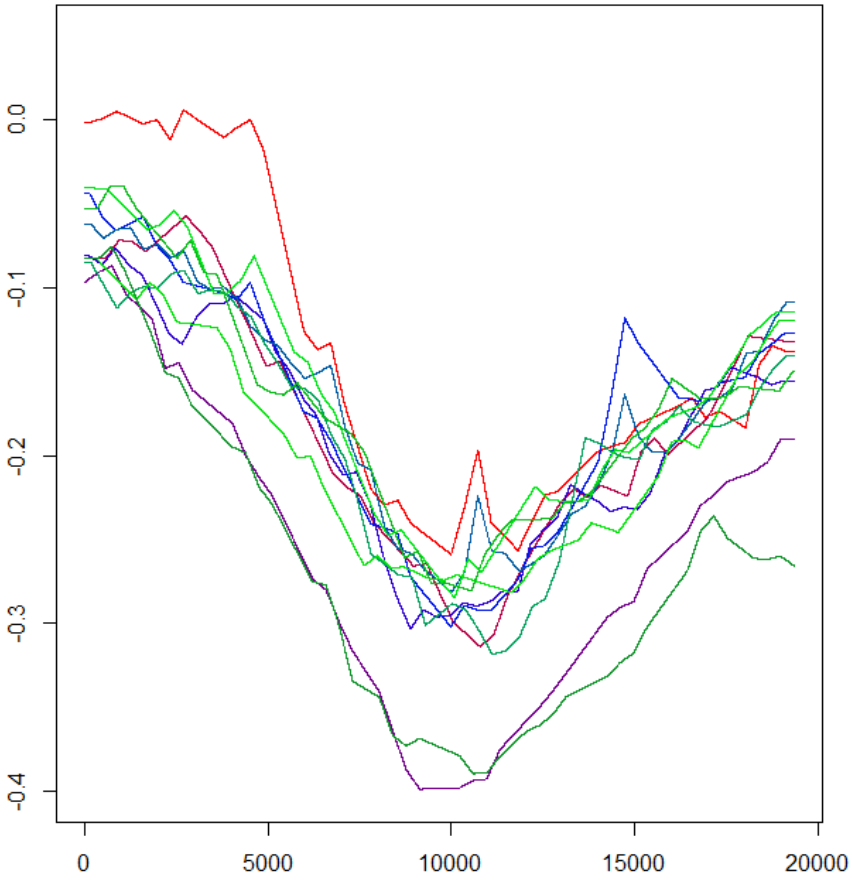
3.14 Naissances en France (1968-2021), Time-varying calendar effects

R code (rjd3modelling, rjd3sts)

```
hol<-rjd3modelling::holidays(jhol, "1968-01-01",
                             length = length(y), type = "Skip")

# create the model
sm<-model()
eq<-equation("eq")
# create the components and add them to the model
add(sm, noise("n"))
add(sm, seasonal("s", 7, type='HarrisonStevens'))
add(sm, locallineartrend("ll"))
add(sm, reg("cal", hol, .1))
add.equation(eq, "n")
add.equation(eq, "ll")
add.equation(eq, "s")
add.equation(eq, "cal")
add(sm, eq)
#estimate the model
rslt<-estimate(sm, log(y), marginal=F, initialization="SqrtDiffuse",
               optimizer="LevenbergMarquardt",
               concentrated=TRUE, precision = 1e-10)

fs<-result(rslt, "ssf.filtered.states")
ss<-result(rslt, "ssf.smoothing.states")
```



Processing time: 30 sec

3.21 Décomposition de type « model-based »

- Décomposition canonique (SEATS)
  - Calcul du modèle UCARIMA
    - Pas de problème théorique, complexité numérique dans le cas de séries journalières → traitement de larges polynômes → algorithmes spécialisés (calcul de racines…)
  - Estimation des composants
    - Algorithme de Burman pas suffisamment robuste
    - Kalman smoother → problèmes de mémoire → filtre de Chandrasekhar + Disturbance smoother → problème de robustesse
    - **→ Processing iteratif: élimination successive des effets de la plus haute à la plus basse fréquence**
  - **Limite: données journalières**

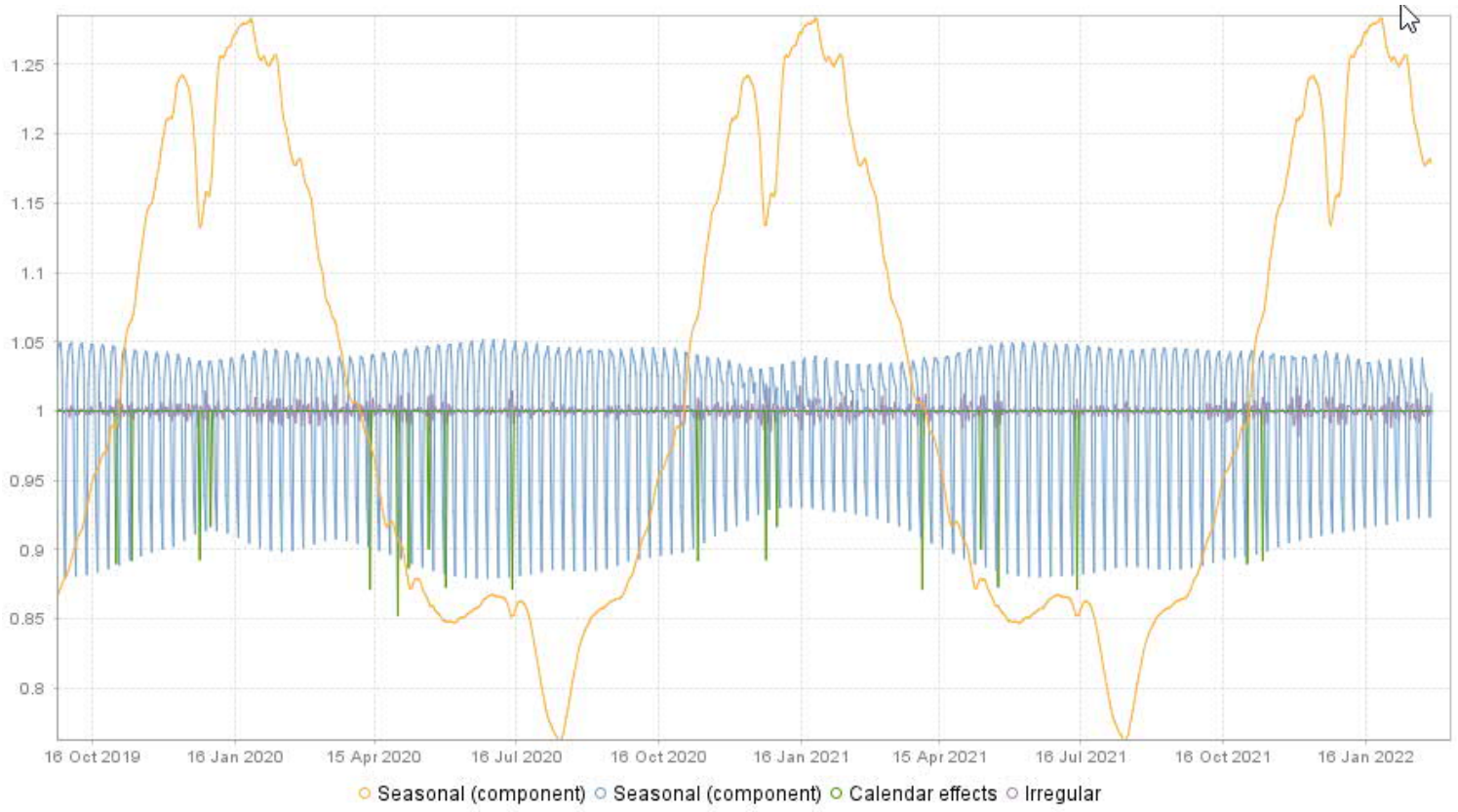
3.22 Décomposition de type non paramétrique

- X11, modifié pour les périodicités non entières (approche comparable au modèle airline fractionel)
- STL: idem + options additionnelles
- Autres filtres: polynômes locaux avec différents noyaux et différents traitement des filtres asymétriques

- **Processus itératif dans le cas de périodicités multiples**
- **Pas d'identification automatique des filtres optimaux (I/C ratio…)**

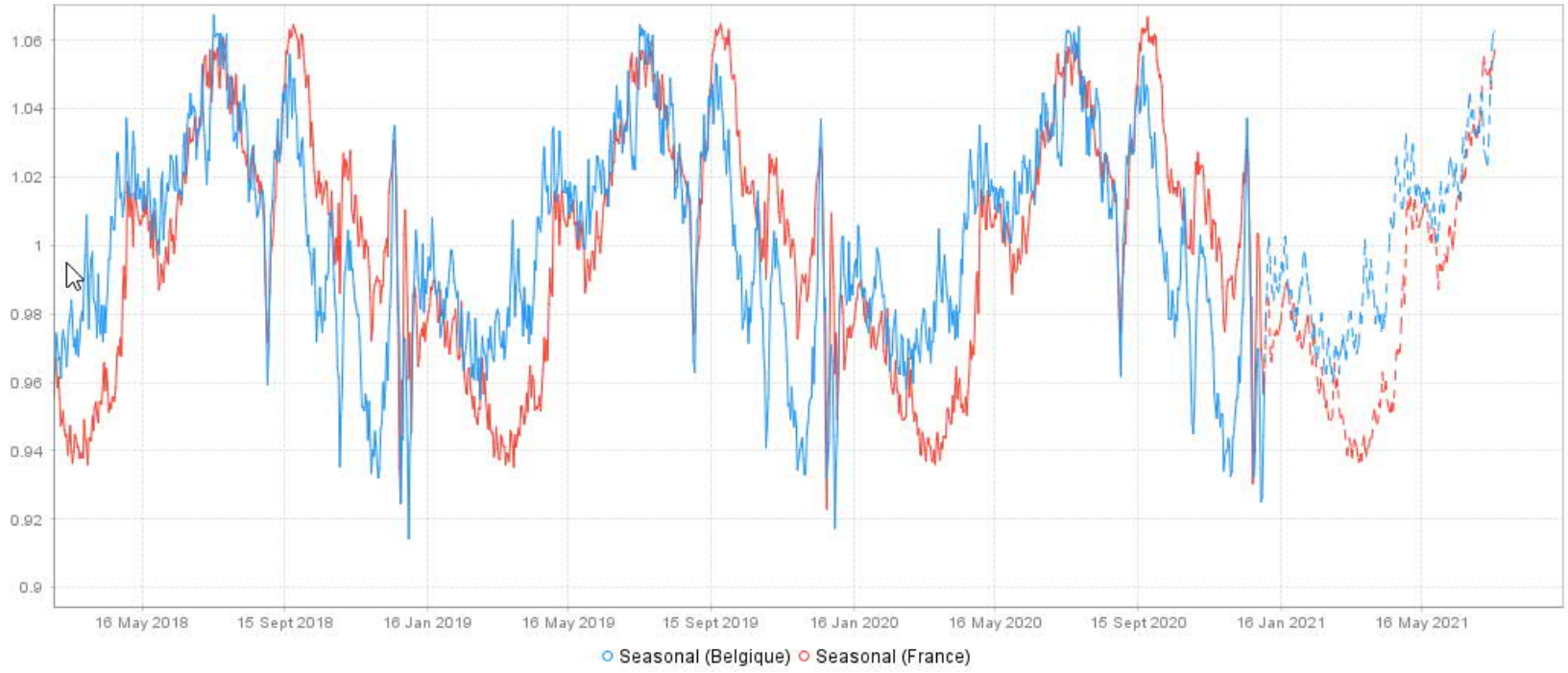
3.23 Exemple I (fractional airline)

- Consommation d’électricité en France



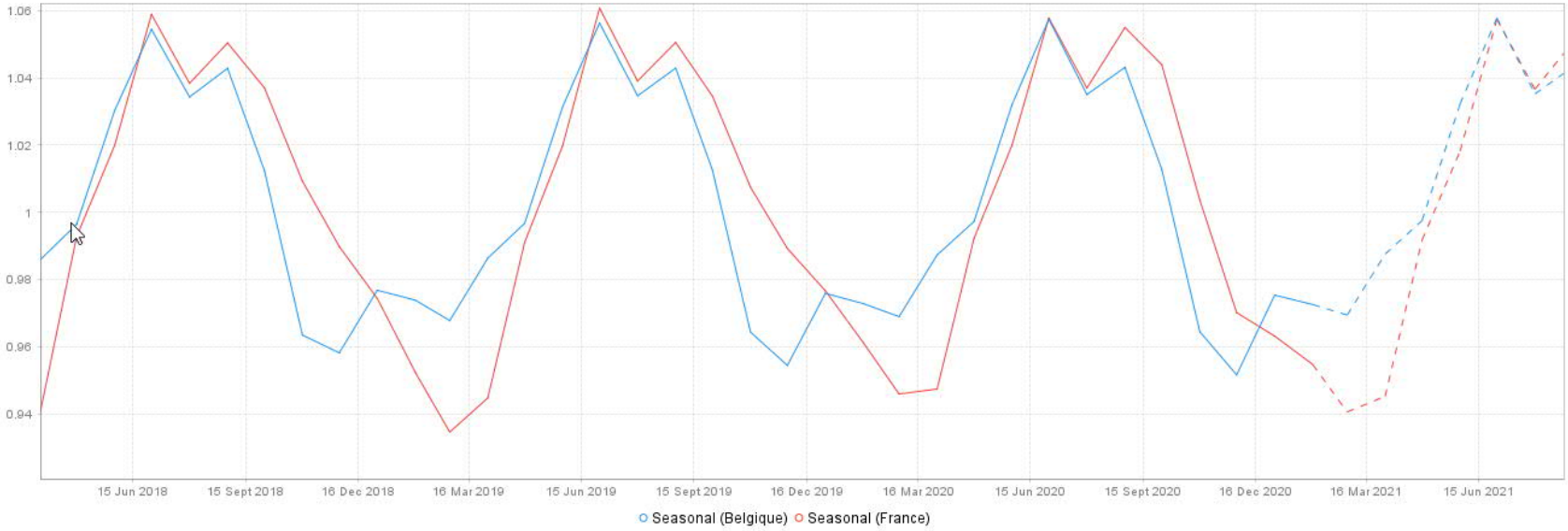
3.23 Exemple II (fractional airline)

- Naissances journalières. Saisonnalité annuelle

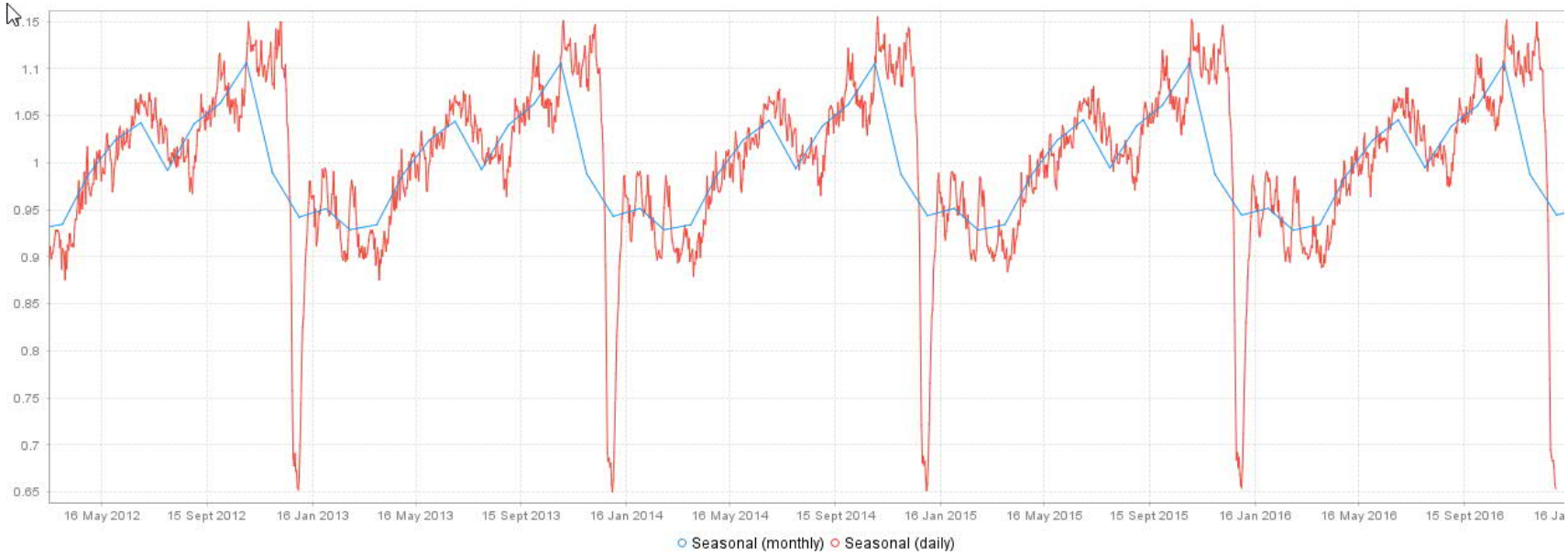


3.23 Exemple II (Tramo-Seats)

- Naissances mensuelles. Saisonnalité



3.23 Exemple III Accidents de la route (GB)



3.24 Constatations

- Algorithmes actuels:
  - La « saisonnalité » de chaque période (jour, semaine…) est estimée de manière quasi-indépendante →
    - Effets intra-mensuels non nécessaire
    - Distinction saisonnalité/jours fériés compliquée
    - Composante(s) saisonnières potentiellement très irrégulières
  - Problèmes d'overfitting, de rapport signal/bruit
  - Problèmes d'estimation (state spaces trop larges)
- Apports de l'analyse des séries à haute fréquence
  - Validation (ou non) des pratiques actuelles (jours fériés = dimanche)
  - Evolutions intra-mensuelles parfois très marquées, non visible au niveau agrégé
  - …

4. Recherche en cours

- Modélisation
  - RegArima (ou modèles structurels) + régression avec coefficients variant dans le temps (voir plus haut)
- Décomposition
  - Model-based (state-spaces):
    - Composantes saisonnières modélisées par des periodic cubic splines
    - Nœuds des splines:
      - Nombre et positions identifiés de manière automatique (adaptative cubic splines)
      - Valeur variant dans le temps (random walk)

4. Recherche en cours (II)

- Voir notamment:
  - Vivien Goepp, Olivier Bouaziz, Grégory Nuel. Spline Regression with Automatic Knot Selection. 2018. hal-01853459
  - Andrew Harvey, Siem Jan Koopman, Marco Riani. The Modeling and Seasonal Adjustment of Weekly Observations. 1997. JASA
- Adaptation aux méthodes non paramétriques
  - Valeurs des nœuds estimées par lissage sur un ensemble de points correctement identifiés (à investiguer)

5. Modules disponibles

- JDemetra+ 3.0 snapshot
  - https://github.com/nbbrd/jdemetra-app-snapshot/releases
- JDemetra+ 3.0 RC1
  - 15/7/2022
- R packages
  - https://github.com/palatej?tab=repositories