Modélisation de séries temporelles à haute fréquence : aspects logiciels

Guy Mélard [1]

[1] *ECARES, Université libre de Bruxelles, CP114/4 Avenue Franklin Roosevelt, 50, B-1050 Bruxelles, & ITSE sprl, rue Georges Huygens, B-1160 Bruxelles.*
gmelard@ulb.ac.be

Résumé. Les données temporelles en haute fréquence (semaine, jour, heure, etc.) sont de plus en plus souvent disponibles alors que les méthodes classiques traitent essentiellement des données trimestrielles ou mensuelles. En conséquence, les séries temporelles sont plus longues. A première vue, c'est excellent du point de vue statistique excepté que le nombre de paramètres des modèles ARIMA ou en espace d'état peut être accru. En réalité, plus de paramètres ne sont pas nécessairement requis mais l'utilisation de données hebdomadaires, journalières ou en haute fréquence implique des difficultés. Mélard (2013) a montré quelques exemples (vitesse du vent au sommet d'une éolienne, trafic dans une cellule d'un réseau GSM, ventes journalières de produits en magasin, flux monétaires journaliers dans des organisations, consommation d'énergie par heure dans des bâtiments de bureaux) et discuté quelques solutions comme l'utilisation des jours ouvrables (affectés par les congés et les ouvertures exceptionnelles), les effets de calendrier (jours fériés, promotions, longueur de mois), la flexibilité de l'horaire de travail, des variables explicatives. Une autre implication des séries en haute fréquence est que les variables de comptage prennent souvent des petites valeurs entières positives alors que la plupart des techniques de modélisation supposent des variables continues. Cet article présente une revue des aspects logiciels comme les modèles à coefficients dépendant du temps, le traitement de données manquantes et le traitement automatique des données aberrantes puisque les données manquantes et aberrantes surviennent plus souvent pour les séries en haute fréquence. Des références à la littérature et des indications d'implémentation logicielle seront données.

Mots-clés. série chronologique, haute fréquence, logiciel statistique.

Abstract. High frequency data (weekly, daily, hourly, ...) are more and more available whereas classical methods handle essentially quarterly or monthly time series. As a consequence, time series are longer. At first, this is fine from a statistical point of view, except that ARIMA or state-space models with more parameters can be expected. In reality, more parameters are not necessarily required, but the use of weekly, daily or higher frequency data implies some difficulties. Mélard (2013) has shown some examples (wind speed at the top of a windmill, traffic in a cell of a GSM network, daily sales of products

in stores, daily cash flows in organizations, energy consumption by hours in office buildings) and discussed some issues like store opening days (affected by holidays, exceptional Sunday openings), calendar effects (holidays, promotions, month length), work schedule flexibility, explanatory variables. Another implication of high frequency data is that counting variables often take some small positive integer values whereas most modeling techniques suppose continuous variables. This paper is a review covering all these aspects and some others, like models with time-dependent coefficients, missing data handling and automatic outlier treatment since missing data and outliers occur more often with high frequency data. References to the literature and indications on software implementations will be given.

Keywords. time series, high frequency, statistical software.

1 Introduction

Nous vivons dans un monde d'information avec de plus en plus de données. Dans le passé, les applications de l'analyse des séries chronologiques se contentaient généralement de données mensuelles et trimestrielles, notamment en économie et dans le monde des affaires. Il suffit de consulter les livres et les articles. De nos jours, nous devons traiter des données hebdomadaires, journalières, horaires ou à plus haute fréquence.

Dans un article récent, Mélard (2013), nous avons traité à un petit nombre de séries temporelles collectées pendant une période de dix ans. Elles sont typiquement plus longues que les exemples des ouvrages de base et nous avons d'abord cru que la méthodologie de Box et Jenkins conduirait à des modèles ARIMA avec un grand nombre de paramètres. Cette impression s'est révélée fausse puisque nous avons eu rarement besoin de plus de 5 paramètres, à l'exception des interventions.

Mais d'autres problèmes surviennent quand on traite des séries de trafic de réseaux, de balance de trésorerie, de données de ventes journalières, ou de consommation d'énergie sur un intervalle de 15 minutes. Selon les séries, vous rencontrez des problèmes de pannes, congés, grèves, plus les nombres différents de jours dans une semaine ou dans un mois qui perturbent la quasi périodicité sur base journalière, hebdomadaire, mensuelle ou annuelle. De plus, dans le cas de comptages, comme la plupart des séries de ventes, la variable peut prendre un petit nombre de valeurs naturelles, même être souvent égale à 0. Ceci n'est pas compatible avec l'hypothèse de distribution continue dans les modèles ARIMA. Nous allons insister ici sur les problèmes suivants : le traitement des données manquantes et des données aberrantes, et la modélisation de données de comptage.

2 Les méthodes de base

Il s'agit

- des modèles ARIMA saisonniers avec interventions, Box et al. (2015);
- d'alignement des données sur des périodes typiques (par exemple 22 jours par mois pour des données de flux financiers journaliers);
- de régression avec variables explicatives appropriées (par exemple l'occupation des locaux pour la consommation d'énergie).

3 Les méthodes avancées

Les modèles ARIMA sont complétés des extensions appropriées :

- Traitement automatique des données aberrantes ("outliers") avec 4 types au moins (parmi lesquels AO ou "additive outliers", IO, LS et TC), Chen et Liu (1993), Gómez et Maravall (2001);
- Traitement automatique des données manquantes par extrapolation, Jones (1980), interpolation, Gómez et al. (2001), ou remplacement par des AO, Gómez et al. (2001) et Prioetti (2008);
- Modèles de données de comptage, Davis et al. (2003, 2005), Dunsmuir (2010);
- Modèles coefficients dépendant du temps, Van Bellegem et von Sachs (2004), Azrak et Mélard (2006);
- erreurs GARCH (pour données financières), Francq et Zakoïan (2009).

4 Les aspects logiciels

Les logiciels statistiques considérés sont

- R arima standard et package forecast Arima;
- R glarma package (pour données de comptage);
- SAS;
- SPSS (ARIMA et TSMODEL);
- Stata;
- Tramo-Seats (TS).

5 Conclusions

- Tous ont les possibilités de date et de temps (mais pas l'alignement) sauf TS exclusivement conçu pour des données mensuelles et trimestrielles;
- Stata possède une possibilité de gestion de calendrier;
- Plus d'une période saisonnière n'est possible que dans SAS;
- Le traitement automatique des données aberrantes ("outliers") dans SAS (3 types), SPSS TSMODEL (7 types), TS (4 types), mais pas dans R ou Stata;
- Il existe des procédures différentes (et non équivalentes) pour le traitement automatique des données manquantes par extrapolation (SAS et Stata), interpolation (R, SPSS ARIMA, TRAMO-SEATS) ou remplacement par des AO (TRAMO-SEATS) et rien pour SPSS TSMODEL;
- Modèles de comptage : seulement dans R glarma, et pas dans les logiciels courants.

Bibliographie

[1] Azrak, R., Mélard, G. (2006), Asymptotic properties of quasi-maximum likelihood estimators for ARMA models with time-dependent coefficients, *Statistical Inference for Stochastic Processes* **9**, 279-330, 2006.
[2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015) *Time Series Analysis, Forecasting and Control*, 5th edn., Wiley, Hoboken NJ.
[3] Chen, C., Liu, L.-M. (1993) Joint estimation of model parameters and outliers effects in time series, *J. Amer. Statist. Assoc.* **88**, 284-297.
[4] Davis, R. A., Dunsmuir, W. T. M., Streett, S. B. (2003) Observation driven models for Poisson counts, *Biometrika* **90**, 777-790.
[5] Davis, R. A., Dunsmuir, W. T. M., Streett, S. B. (2005) Maximum likelihood estimation for an observation driven model for Poisson counts, *Methodol. Comput. Appl. Probab.* **7**, 149-159.
[6] Dunsmuir, W. T. (2010) R software for fitting observation driven regression models for univariate time series.
[7] Francq, C., Zakoïan, J.-M. (2009) Modèles Garch : structure, inférence statistique et applications financières Economica, Paris.
[8] Gómez, V., Maravall, A. (2001) Automatic modeling methods for univariate series, Chapter 7 in Peña, D., Tiao, G.C., Tsay, R.S. (eds), *A Course in Time Series Analysis*, Wiley, New York, pp. 171-201.
[9] Gómez, V., Maravall, A., Peña, D. (2001) Missing observations in ARIMA models: skipping strategy versus additive outlier approach, *J. Econometrics* **88**, 341-363.

[10] Jones, R. H. (1980) Maximum likelihood fitting of ARMA models for time series with missing observations, *Technometrics* **22**, 389-395.
[11] Mélard, G. (2013) Forecasting daily and high-frequency data, presentation at WIP-FOR'13, EDF, Paris, June 5-7.
http://homepages.ulb.ac.be/ gmelard/rech/Forecasting_daily_and_high_frequency_data_v9.pdf
[12] Proietti, T. (2008) Missing data in time series: a note on the equivalence of the dummy variable and the skipping approaches, *Statistics and Probability Letters* **78**, 257-264.
[13] Van Bellegem, S., von Sachs, R. (2004) Forecasting economic time series with unconditional time-varying variance, *International Journal of Forecasting* **20**, 611-627.