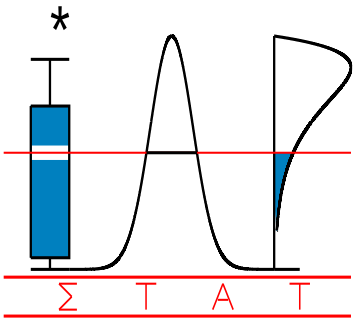
T E C H N I C A L

R E P O R T

0445

FORECASTING IN THE ANALYSIS OF MOBILE TELECOMMUNICATION DATA- CORRECTION FOR OUTLIERS AND REPLACEMENT OF MISSING OBSERVATIONS

AZRAK, R., MELARD. G. and H. NJIMI



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

FORECASTING IN THE ANALYSIS OF MOBILE TELECOMMUNICATION DATA – CORRECTION FOR OUTLIERS AND REPLACEMENT OF MISSING OBSERVATIONS

Rajae Azrak
Université Mohammed V, Faculté des sciences juridiques, économiques et sociales, Salé, Maroc
razrak96@yahoo.fr

Guy Mélard
ECARES, Université Libre de Bruxelles CP114, Av. Franklin Roosevelt, 50, B-1050 Bruxelles, Belgique
gmelard@ulb.ac.be

Hassane Njimi
Faculté des sciences sociales, politiques et économique & ISRO, ULB, Bruxelles, Belgique
hnjimi@ulb.ac.be

RÉSUMÉ - Nous discutons de séries chronologiques qui sont typiques dans un réseau de télécommunication mobile. Fréquemment, des observations manquantes et données aberrantes empêchent l'utilisation de la plupart des logiciels statistiques. Dans cet article, nous envisageons plusieurs approches pratiques pour traiter de telles séries, et nous montrons que l'analyse d'intervention permet d'obtenir des prévisions adéquates, et même d'une manière automatique.
MOTS-CLÉS
Télécommunication mobile, analyse des séries chronologiques, série temporelle, données manquantes, données aberrantes, prévision.

ABSTRACT - We discuss time series data which are typical in a mobile telecommunication network. They often show missing observations and outliers which prevent the use of most statistical software. In this paper, we discuss several practical approaches for dealing with such series and show that intervention analysis, preferably in an unattended form can help to obtain adequate forecasts.
KEYWORDS
Mobile telecommunication, time series analysis, missing data, outliers, forecasting.

1. INTRODUCTION

Since about thirty years, telecommunications has known an evolution both of quantitative and qualitative nature. From an economic point of view, it is of fundamental importance to forecast its evolution. As a proof of the interest of forecasting in telecommunications, let us mention the recent review of Fildes and Kumar [2002] who report more than 130 papers, most of them published during the last ten years. Data are generally collected in an automated way but, as is often the case in automatic data acquisition, recordings are also subject to equipment failure. This should be taken into account in any statistical treatment.

In this paper, we discuss a relatively simple example of traffic measurements in a cell of a Global System for Mobile (GSM) network. It will be shown that ignoring the nature of missing and abnormal observations leads to suboptimal forecasts.

2. FORECASTING IN THE FIELD OF TELECOMMUNICATION NETWORKS

As in many fields, telecommunications network operators need forecasts in order to dimension their network and make investments in due time. Data are collected in an automated way at several levels of disaggregation and at several frequencies. Examples are traffic in a cellular network, telephone calls to a call centre, subscriptions to a given service. For a cellular network, data are available at each cell of the network and are used to decide on cell subdivisions to better manage the workflow. Statistical data are collected by additional equipment, which is less redundant than the more fundamental telecommunication equipment and is therefore more subject to failures. Equipment failures will imply abnormal (generally underestimated) observations, which are called outliers in the usual statistical terminology. If the effect of an equipment failure is long enough, data may be completely missing for a time interval and is sometimes simply recorded as a zero.
Data are generally available at several locations (e.g. cells for a cellular network) and either at intervals of an hour or even sometimes as small as five minutes. That means that the number of observations is huge but the observations are also more sensitive to exogenous or endogenous events. When data are aggregated by weeks or months, the effect of outliers may be reduced but it is bigger at the day level, mainly during week ends, when maintenance effort is reduced, or a fortiori at smaller time intervals.
The complex nature of the networks may require more sophisticate methods than those in actual general practice of forecasting. As an example, Tych et al. [2002] have built an unobserved component model for hourly telephone call demand which includes an enhanced version of the dynamic harmonic regression model, recursive Kalman filter and fixed interval smoothing algorithms that are capable of automatically handling missing data and outliers. In this paper, we describe a much simpler approach but which has the same capability as far as missing data and outliers are concerned.

3. THE NEED FOR TREATMENT OF MISSING DATA AND OUTLIERS

In order to illustrate the need for treatment of missing data and outliers, we make use of a relatively simple example of traffic measurements, the total number of minutes in a cell of a Global System for Mobile (GSM) network, measured at the daily level. Data are presented in Figures 1 and 2, with a number of time points $T = 185$. The methods discussed here can be used for short term forecasting. To assess their usefulness we will use the first 171 observations and reserve the last two weeks for ex-post validation.
We should note that there are several missing data. They appear better during the week starting 24/11/00. As expected, some of these appear during weeks ends ($t =$

79, 80 or December 2 and 3, 2000; t = 128, 129 or January 20 and 21, 2001; t = 149, 150 or February 10 and 11, 2001) but there is another one during the week (t = 82 or Tuesday December 5, 2000). The plot in Figure 2 treats missing data as zeros, like they were recorded and would be treated by most statistical software.

Figure 1. Daily traffic data in a network cell, from September 15, 2000, to March 18, 2001

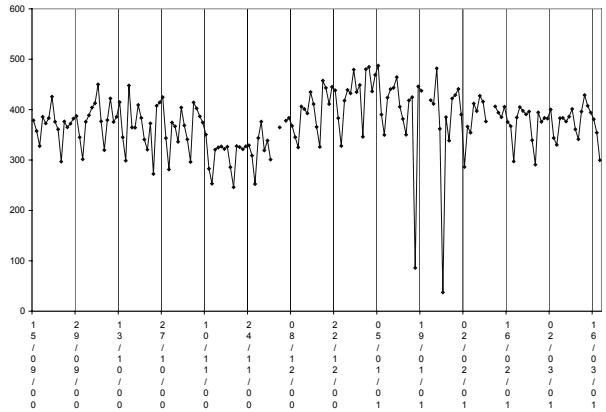
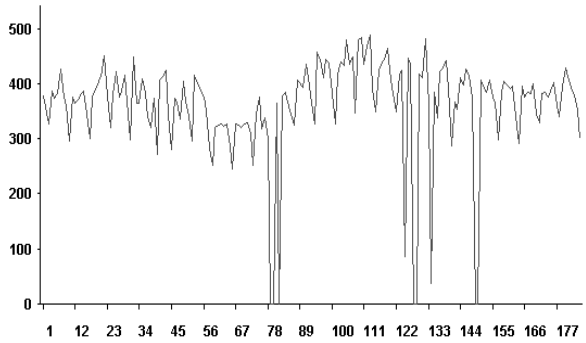


Figure 2. Same data as in Figure 1 but with missing values replaced by 0, as they were in the original file

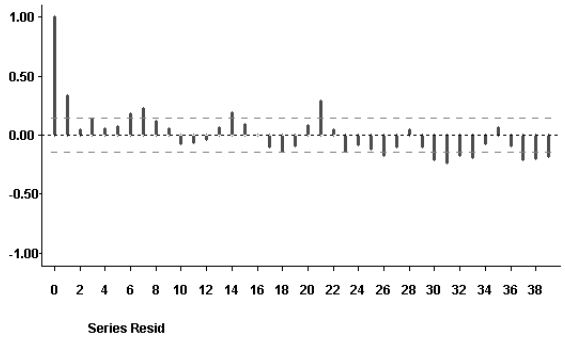


The usual way to forecast a time series y_t, t = 1,…, T, is to build an ARIMA (autoregressive integrated moving average) model. Apart from an error term, often called an innovation and denoted by e_t, that means an expression relating the observation at any time t in function of past observations and past errors. According to the Box-Jenkins [1994] methodology, it is recommended to first make the series stationary. This implies removing a trend and/or a seasonal component by means of, respectively, an ordinary difference ($y_t - y_{t-1}$), and a seasonal difference ($y_t - y_{t-s}$), where s is the seasonal period.

Let us consider the traffic in a network cell with missing observations replaced by zeros and denote them by y_t. We consider the autocorrelations of the series for lags 1 to 39, the autocorrelation meaning the correlation of the series with itself lagged by some delay. If we look at the correlogram, the graph in which the autocorrelations

are summarised, shown in Figure 3, we see just a few statistically significant autocorrelations (at the 5 % level), those outside of the band. Statistical analysis of these autocorrelations are only justified if the series is stationary, in practice if there is no level change, no trend, no periodic variations of any type. This is surely not the case of our series because there is a strong weekly pattern with Sunday traffic much lower than during the other days of the week. Lack of stationarity is often seen on the correlogram because of large autocorrelations that don't converge quickly (in an exponential way) to 0.

Figure 3. Correlogram of the raw data (with missing values replaced by 0)



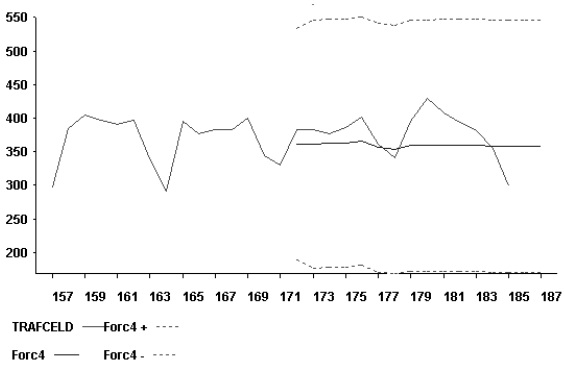
If we don't look at the plot of the data with respect to time, it seems plausible to consider a stationary model for the series. From now on, we have omitted the last two weeks from the analysis in order to check the forecasting performance. Using the model building methodology, as slightly simplified by Mélard [1990], and implemented in Time Series Expert, Mélard and Pasteels [1997], starting with the correlogram and also another device known as the partial correlogram, we are lead to a model described by the equation

$$y_t - 0.16\, y_{t-7} = e_t + 0.38\, e_{t-1}, \qquad (1)$$

with an innovation standard deviation of 88. The output of the residual analysis, i. e. the analysis of the estimated errors derived from the model, reveals a large number of outliers and some, not too big, residual autocorrelations. However, as Figure 4 shows, the forecasts over the last two weeks are not very close from the true observations, and don't picture a very strong weekly seasonal effect. Moreover, there is much variability as illustrated by the broad forecasting intervals (at the 95 % coverage probability).

This is an illustration of the lack of robustness of the statistical methods used, either the autocorrelation analysis, or model estimation using a nonlinear least-squares method. Robust methods in this area are just emerging (e.g. see the recent book by Lucas et al. [2003]) so it may be simpler to use the standard methodology but in a much clever way. This is what will be shown in the next section.

Figure 4. Data and forecasts for the last two weeks obtained from the model (1) built on the data with missing values replaced by zeros



4. USE OF INTERVENTION ANALYSIS

The procedure that will be used is a special case of what is called intervention analysis. Explanatory variables, such as advertising expenses, can be introduced in a time series model, giving a regression model with autocorrelated errors.

Intervention analysis consists in using binary variables that correspond to events that are supposed to have an effect on the variable being studied. We will present here a simplified description. The general idea is to replace the missing observations, and possibly corrections of the outlying observations, as additional parameters in the model. Looking at the data without the missing data (e.g. in Figure 1) we see a clear weekly seasonal component and a less strong trend. This would justify using only a seasonal difference. We denote $Y_t = y_t - y_{t-7}$ (with zeros for the y_t's at the missing dates).

We first try to correct the missing values by replacing them by the average of the whole series. This is equivalent to running an intervention analysis for the series Y_t and with one binary variable at each missing date, seven binary variables in all. Since one parameter is associated with each of these binary variables, that makes a total of 7 parameters. Figure 5 shows the residuals with respect to time and Figure 6 shows the correlogram of the residual series. Significant autocorrelations appear at least lags 1, 2, 3 and 7. There is also the need to take care of the two outliers which appear at time points t = 125 (Wednesday January 17, 2001) and t = 134 (Friday February 26, 2001).

After a few steps, a final model is obtained in Table 7. The equation is not shown here for reasons of space. Its residuals are shown in Figure 8 and their correlogram is displayed in Figure 9. Note that the residual standard deviation is equal to 0.029 so that a few of the residuals are still outlying (with respect to the normal distribution). However their magnitude is much smaller and their dates (t = 40 or Tuesday October 24, 2000; t = 133 or Thursday January 25, 2001; and t = 141 or Friday February 2, 2001) don't correspond to holidays, so we have stopped at this point.

Figure 5. The residuals from the first model with interventions, with just a constant term

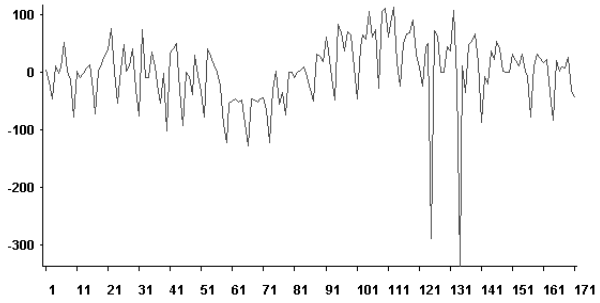
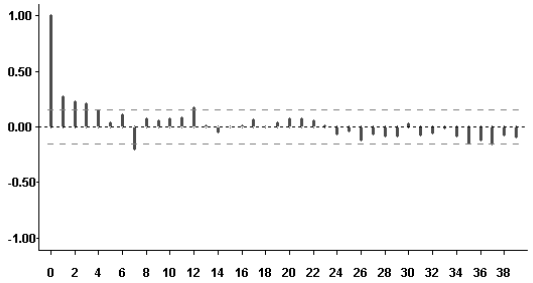


Figure 6. Correlogram of the residual series shown in Figure 5



No significant autocorrelation appears anymore. Forecasts are shown in Figure 10 and 11. They are much more convincing and the forecast intervals at 95 % are smaller than in Section 3.
In the next section, we will show that it is possible to obtain a similar result in an automated way.

5. AUTOMATIC MODEL BUILDING USING TSE-AX

In this section we give the result of the analysis of the daily traffic data set using an expert system called TSE-AX which is slightly improved from the version which is included in Time Series Expert 2.3, as described by Mélard and Pasteels [1997, 2000]. Njimi et al. [2003] give an early presentation of the improvements. One of the recent features is the possibility to handle other series than monthly and quarterly ones. The objective of TSE-AX is to build ARIMA models in an automated way, with and without an intervention analysis, but so that the user should be informed of the steps, receive the intermediate and final results, and be informed of the quality of the final model. Briefly, TSE-AX covers everything from the specification stage to the forecasting stage, given that the latter is immediate when a final model has been found. The user can specify his or her model building preferences (perform an intervention analysis or not, choose a specification strategy, etc.).
To analyse the daily traffic data set we based the choice of the differences in order to make the series stationary on the autocorrelation function.

Table 7. Output from Time Series Expert for the final model

```
ANSECH-PC 2.3c, AUTHOR:G.MELARD 03/10/03 15:47:45. PROBLEM(     1):  TRAFCELD
SERIES READ FROM DISK,         NAME IS TRAFCELD.DB             ,LENGTH  185
/\/\/\/\/\/\/\/\/\
WARNING *** MODEL FITTING IS PERFORMED WITH ONLY 171 DATA, ENDING
AT TIME     171. 14 FRESH DATA ARE RESERVED FOR EX-POST VALIDATION
===KNOWLEDGE ABOUT INTERVENTIONS (BOX-TIAO)
DIRECTIVE      TYPE          DATE        STEP NATURE      PARAM/VALUE COMMENTS
I  79: 0.100  BOX-TIAO           79          VALUE      KI  79: 0.100
. . .
I 125: 0.100  BOX-TIAO          125          VALUE      KI 125: 0.100
I 134: 0.100  BOX-TIAO          134          VALUE      KI 134: 0.100
9 DIRECTIVE(S),      9 PARAMETER(S),      0 CONSTANT(S).
13 PARAMETERS WITH STARTING VALUES :
    1     AR   1  .00000
    2     AR   2  .00000
    3     AR   3  .00000
    4     SMA  1  .00000
    5     KI  79  .10000
   . . .
   13     KI 134  .10000
=== ESTIMATION BY MAXIMIZATION OF THE EXACT (LOG)LIKELIHOOD
=== MODEL DESCRIPTION                  FORM      DEGREE/ORD PARAMETERS NUMBER
- SEASONAL PERIOD                                    7
- BOX-TIAO INTERVENTION              SEE ABOVE            9     KIdddd      9
- DIFFERENCE                         SEASONAL             1
- ADDITIVE CONSTANT                  AUTOMATIC
- ARMA MODEL
      AUTOREGRESSIVE POLYNOMIAL      REGULAR              3     AR  nn      3
      MOVING AVERAGE POLYNOMIAL      SEASONAL             1     SMA nn      1

NON LINEAR ESTIMATION:
. . .
=ITERATION STOPS - RELATIVE CHANGE IN EACH COEFFICIENT LESS THAN  1.00000E-03
FINAL VALUES OF THE PARAMETERS                          WITH 95% CONFIDENCE LIMITS
        NAME      VALUE          STD ERROR   T-VALUE  LOWER          UPPER
 1      AR   1    .38107       8.14008E-02     4.7     .22          .54
 2      AR   2    .14622       8.92730E-02     1.6   -2.88E-02      .32
 3      AR   3    .31196       7.96760E-02     3.9     .16          .47
 4      SMA  1    .88243       5.81513E-02    15.2     .77         1.0
 5      KI  79   -297.21        26.867        -11.1   -3.50E+02   -2.45E+02
 6      KI  80   -268.04        26.044        -10.3   -3.19E+02   -2.17E+02
 7      KI  82   -362.30        26.071        -13.9   -4.13E+02   -3.11E+02
 8      KI 128   -391.56        26.890        -14.6   -4.44E+02   -3.39E+02
 9      KI 129   -357.81        26.088        -13.7   -4.09E+02   -3.07E+02
10      KI 149   -361.69        26.197        -13.8   -4.13E+02   -3.10E+02
11      KI 150   -316.18        26.230        -12.1   -3.68E+02   -2.65E+02
12      KI 125   -349.40        26.219        -13.3   -4.01E+02   -2.98E+02
13      KI 134   -355.10        26.987        -13.2   -4.08E+02   -3.02E+02
ESTIMATION HAS TAKEN   .2 SEC. FOR 169 EVALUATIONS OF S.S. (MEAN TIME=,  .001)
THE FOLLOWING PARAMETERS WERE ESTIMATED SEPARATELY
        MEAN     -.12652
=== SUMMARY MEASURES
SUM OF SQUARES :     COMPUTED =  134378.      ADJUSTED =  126148.
VARIANCE ESTIMATES :   BIASED =  769.196      UNBIASED =  840.987
TOTAL  NUMBER  OF  PARAMETERS = 14  STANDARD DEVIATION =  28.9998
INFORMATION CRITERIA :    AIC =  1663.72          SBIC =  1712.86
=== RESIDUAL ANALYSIS WITH 164 RESIDUALS, BEGINNING AT TIME       8===
MEAN =  .873917     ,T-STATISTIC =    .39        (FOR TESTING ZERO MEAN)
=OUTLIERS
 < .01 % 141: -119.5
.01-.2 %  40: -93.98     133: -95.60
 1 - 5 %  32:  66.99      41:  66.13     132:  65.65     143:  57.17
=SIGNIFICANT AUTOCORRELATIONS (USING BARTLETT LIMITS)
=SIGNIFICANT PARTIAL AUTOCORRELATIONS
=LJUNG-BOX PORTMANTEAU TEST STATISTICS ON RESIDUAL AUTOCORRELATIONS
 ORDER D.F. STATISTIC  SIGNIFICANCE
 15     11   13.69        .251
 23     19   17.25        .573
 === FORECASTING FROM      171  WITH FRESH DATA <F>
     DATE   OBSERVATION   FORECAST      ERROR  % ERROR  95% FORECAST INTERVAL
      172        383.04     410.16     -27.12              353.32     467.00
      173        383.34     399.51     -16.17              338.68     460.33
      174        376.47     409.50     -33.03              346.46     472.54
      175        385.87     408.05     -22.18              339.38     476.71
      176        401.04     389.50      11.54              318.11     460.89
      177        360.88     362.28      -1.40              288.99     435.58
      178        341.28     320.47      20.81              245.08     395.87
      179        396.00     405.85      -9.85              327.33     484.37
      180        428.92     402.27      26.65              322.11     482.43
      181        407.55     406.91        .64              325.38     488.44
      182        394.33     406.10     -11.77              323.23     488.97
      183        380.70     389.22      -8.52              305.35     473.08
      184        354.20     361.06      -6.86              276.39     445.74
      185        299.71     319.34     -19.63              233.95     404.73
 CUMULATED ERROR          :    -96.89 (=     %); MEAN ERROR:      -6.92
 MEAN ABSOLUTE ERROR (MAE):     15.44 (=     %);
 ROOT MEAN SQUARE ERROR   :       .00 (=     %); MEAN SQUARE ERROR:  329.
```

Figure 8. Residuals from the final model

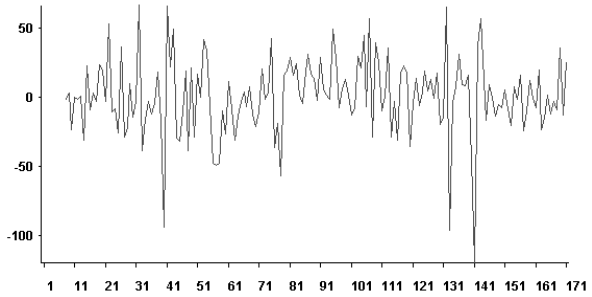


Figure 9. Residual correlogram of the final model

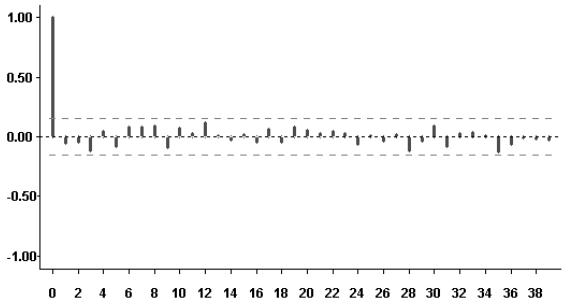
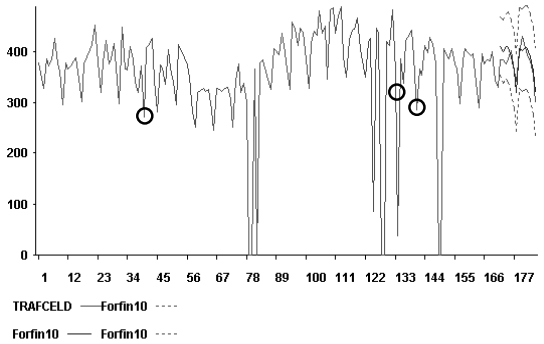
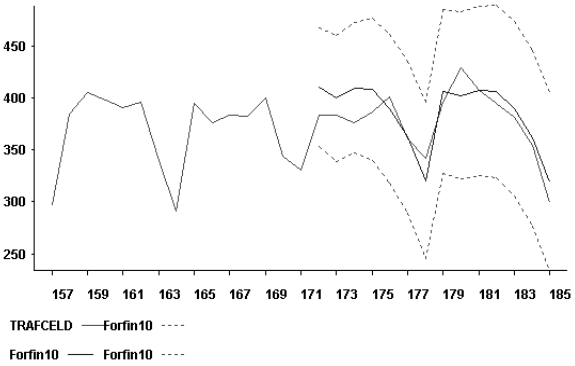


Figure 10. Plot of the data, with circles around the data points that correspond to the remaining largest residuals and the forecasts for the last two weeks



We requested the use of intervention analysis to avoid missing and extreme observations that would influence the analysis. We chose the "Mixed strategy" because this is the most complete strategy. The result of this analysis is obtained in Figure 11. The details are omitted for lack of space but the model is nearly identical to the one obtained by the manual approach and described in Section 4. For the final model, the value of the SymMAPE criterion is 4.07%. This shows closeness of the forecasts from the true observations over the last two weeks.

Figure 11. Zoom of Figure 10 over the last four weeks.



6. CONCLUSION

We have shown that forecasting mobile communication data should take outliers and missing observations into account and that a simple procedure can be used to reach that goal. Furthermore, we have shown that the problem can even be solved in an automated way. Some software packages offer similar capabilities.

ACKNOWLEDGEMENT

This paper has benefited from an IAP-network in Statistics grant, contract P5/24, Belgian Federal Office for Scientific, Technical and Cultural Affairs. We are grateful to an anonymous referee for his/her suggestions.

REFERENCES

Box G. E. P., Jenkins, G. M., Reinsel G. C. [1994]. *Time Series Analysis, Forecasting and Control*, Prentice-Hall Press, Upper Saddle River, NJ, 3rd edition.

Fildes R., Kumar V. [2002]. Telecommunications demand forecasting — a review, *International Journal of Forecasting* 18, pp. 489-522.

Lucas A., Franses P. H., Van Dijk D. [2003]. *Outlier robust analysis of economic time series*, Oxford University Press, Oxford.

Mélard G. [1990]. *Méthodes de prévision à court terme*. Editions de l'Université de Bruxelles, Bruxelles, and Editions Ellipses, Paris.

Mélard G., Pasteels J.-M. [1997]. Manuel d'utilisateur de Time Series Expert (TSE version 2.3), 3e édition, Institut de Statistique et de Recherche Opérationnelle, Université Libre de Bruxelles, Bruxelles.

Mélard G., Pasteels J.-M. [2000]. Automatic ARIMA modeling including interventions, using time series expert software, *International Journal of Forecasting* 16, pp. 497-508.

Njimi H., Mélard G., Pasteels J.-M. [2003]. Modélisation SARIMA assistée, *XXXVèmes Journées de Statistique, Société Française de Statistique*, Lyon, France, pp. 731-734.

Tych W., Pedregal D. J., Young P. C., Davies J. [2002]. An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system, *International Journal of Forecasting* 18, pp. 673-695.