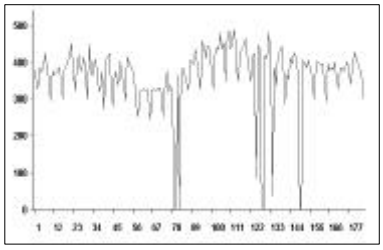
Forecasting Daily and High-frequency Data



by Guy Mélard

gmelard@ulb.ac.be

ECARES
Université Libre
de Bruxelles

(parts of this talk with R. Azrak, H. Njimi, A. Ouakasse
and contributions from M. Ngoko Njipkap and N. Molowayi)

Abstract

Forecasting daily and high-frequency data is often considered for market data but its use for other economic data leads to several problems. This talk will discuss some of them and also present some solutions.

The discussion will be introduced by some examples of forecasting: wind speed at the top of a windmill, traffic in a cell of a GSM network, daily sales of a product in a supermarket, treasury financial flows in order to avoid liquidity problems, energy consumption in plants and offices.

Although the series may be very long, ARIMA modelling is still useful provided some extensions are considered: intervention analysis (to handle failures and equipment breakdown, or promotions in sales), data reallocation (to cover most calendar effects), time rescaling (to take care of different month lengths), allowance for deterministic components (for seasonality in a wide sense), combination with a model for count data (when small integers are observed), ...

For such an important subject at the frontier of engineering, econometrics, statistics and operational research, the literature is surprisingly scarce. This talk will show some partial solutions which are parts of an ongoing research project.

Objectives of the talk

Forecasting daily and high-frequency data is often considered for market data, not for other economic data

Some examples to illustrate the various problems with these data: wind speed at a windmill, traffic in a cell of a GSM network, daily sales of a product, treasury financial flows, energy consumption in plants/offices

ARIMA modelling is still useful with some extensions: intervention analysis (sales promotions), data reallocation (most calendar effects), time rescaling (month lengths), allowance for deterministic components (seasonality), combination with a model for count data (small integers), ...

At the frontier of engineering, econometrics, statistics & OR. Literature is surprisingly scarce. This talk will show some partial solutions, parts of ongoing research

Plan of the talk

Plan of the talk

Introduction

- ⇨Motivation: more and more daily and high-frequency time series
- ⇨e.g. financial market data but other data are more challenging
- ⇨We will analyse some of the new problems for daily and high-frequency data
- ⇨Let us start with examples

Plan of the talk

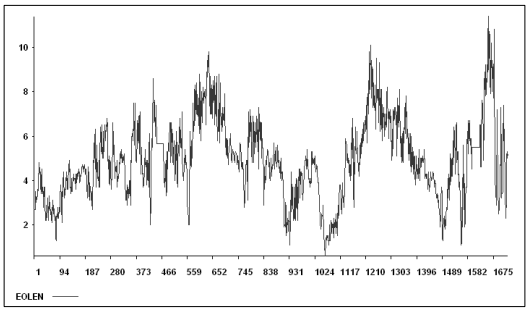
Example 1

- ⇨Production of electricity by windmills
- ⇨It is however irregular because of wind irregularity
- ⇨In order to maintain the offer of electricity
  at the level of demand, necessity to adapt production from traditional power stations
  in function of the amount of electricity produced by a park of windmills

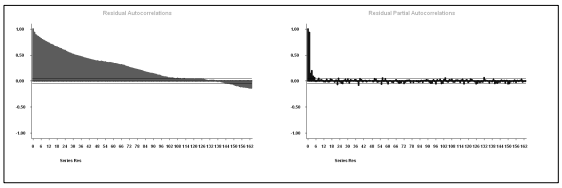
Example 1 (cont'd)

- ⇨Response time of a power station can go from a few minutes to several hours according to the technology being used
- ⇨It is therefore useful to forecast wind speed a few hours in advance
- ⇨The data come from speed of wind measurements at the top of a windmill
- ⇨They are available every ten minutes, hence 144 observations per day
- ⇨We have used about twelve days of measurements, more precisely 1728 observations

Data over a 12-day period

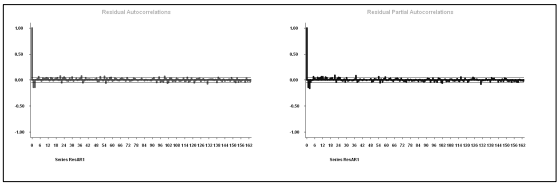


Analysis



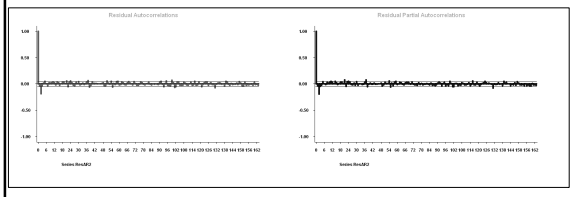
- ⇨ACF (left) and PACF (right)
- ⇨A unit root can be considered but we didn't
- ⇨Autoregressive specification procedure

Residuals from AR(1) model



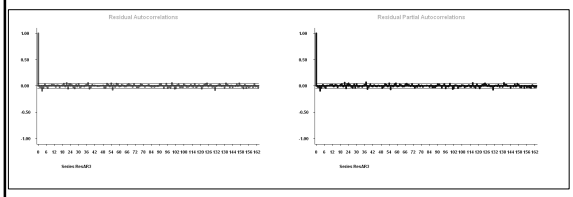
- ⇨Residual ACF reveals autocorrelation at lags 1 & 2

Residuals from AR(2) model



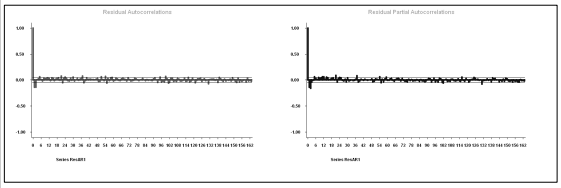
⇨Not simpler

Residuals from AR(3) model



⇨Better but not perfect

Residuals from AR(1) model



⇨Residual ACF reveals autocorrelation at lags 1 & 2

⇨Hence residuals can be modelled by MA(2) process

ARMA(1,2) fit

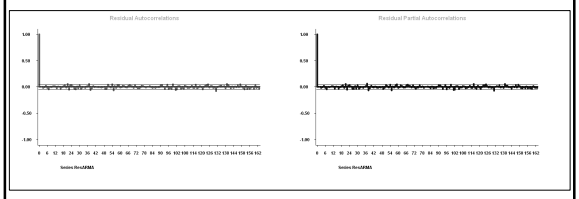
```
NON LINEAR ESTIMATION:
ITER SUM OF SQ AR   1    MA   1    MA   2
 0    5319.      .000     .000     .000
 1    1291.      .389    -.389    -.162
 2    777.9      .872    -.179    -.144
 3    663.2      .974     .121     .158
 4    652.5      .978     .241     .208
 5    652.4      .979     .243     .198
 6    652.4      .979     .243     .198
=ITERATION STOPS - RELATIVE CHANGE IN SUM OF SQUARES LESS THAN  1.00000E-06
FINAL VALUES OF THE PARAMETERS                        WITH 95% CONFIDENCE LIMITS
          NAME      VALUE          STD ERROR   T-VALUE  LOWER          UPPER
     1     AR   1   .97911        5.34907E-03   183.0     .97          .99
     2     MA   1   .24273        2.46997E-02     9.8     .19          .29
     3     MA   2   .19820        2.47181E-02     8.0     .15          .25
THE FOLLOWING PARAMETERS WERE ESTIMATED SEPARATELY
       MEAN       4.9550
=== SUMMARY MEASURES  <V>
TOTAL  NUMBER  OF  PARAMETERS =  4  STANDARD DEVIATION =  .621294
```

$$(1-0.98L)(\mathrm{EOLEN}_t - 4.95) = (1-0.24L-0.20L^2)e_t, \qquad \hat{s} = 0.621$$

ARMA(1,2) residual analysis

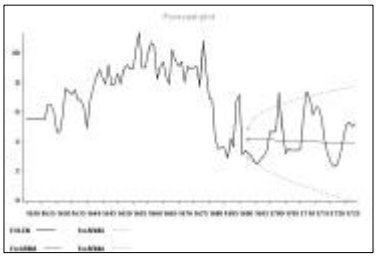
```
=== RESIDUAL ANALYSIS WITH 1692 RESIDUALS, BEGINNING AT TIME       1===
MEAN =  .944975E-03 ,T-STATISTIC =    .06        (FOR TESTING ZERO MEAN)
=OUTLIERS <R(OR)S>
  < .01 %  424: -2.550    604:  2.518    820: -2.454   1562:  2.451   1565:  2.729
          1678:  2.917   1682: -3.136   1689:  3.102   1691: -3.155
  . . .
=SIGNIFICANT AUTOCORRELATIONS (   BARTLETT LIMITS  ) <A(OR)S>
 .2 - 1 % 131:-.7363E-01
  1 - 5 %  21: .5625E-01  22:-.5050E-01  37: .5371E-01 38:-.5453E-01 55:-.6466E-01
           99: .5131E-01 101:-.5575E-01 139: .5488E-01 430: .6333E-01
=SIGNIFICANT PARTIAL AUTOCORRELATIONS <P(OR)S>
 .2 - 1 %  37: .6296E-01  38: -.6372E-01
  1 - 5 %  21: .5969E-01  22: -.5144E-01
=LJUNG-BOX PORTMANTEAU TEST STATISTICS ON RESIDUAL AUTOCORRELATIONS <L>
 ORDER D.F. STATISTIC  SIGNIFICANCE
 72     69   78.77        .197
 144    141  158.15       .153
```


Residuals from ARMA(1,2) model



⇨Conclusion: no improvement according to Mélard (2007) (for an ARMA model at least)

Forecasts from ARMA(1,2) model



⇨Conclusion: forecasts are not very good

Conclusions of Example 1

⇨Long series do not necessarily lead to more complex models

⇨Here information of wind direction and/or wind speed on other windmills can possibly be useful

⇨In Ouakasse and Mélard (2008) we have used on-line estimation for that series (using a new approach that makes use of the Fisher information matrix evaluated at each time)

⇨Given that off-line estimation is fast (a few seconds at most), both methods are competitive

Plan of the talk

Example 2

⇨Automatic data collection

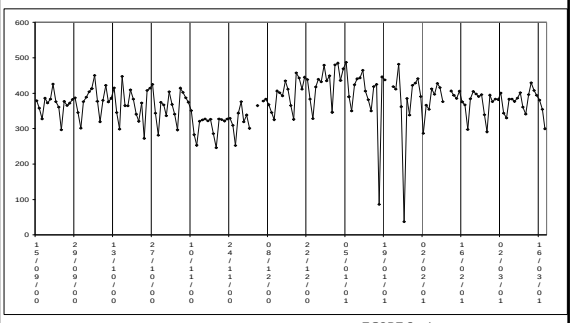
⇨e.g. mobile telecommunication network

⇨First manual time series modelling, then automatic analysis (Azrak, Mélard and Njimi, 2004)

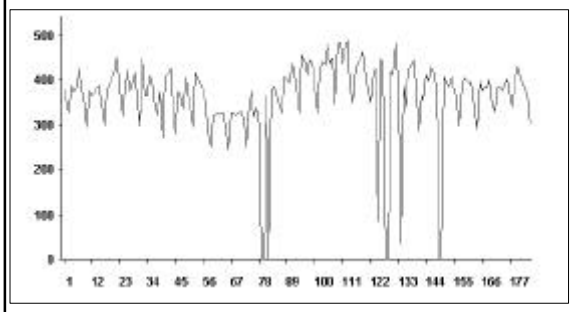
⇨More precisely, the effect of missing data and outliers on forecasting performance

⇨Data are daily total number of minutes in a cell a a GSM network for the period from September 15, 2000 to March 18, 2001 (185 observations)

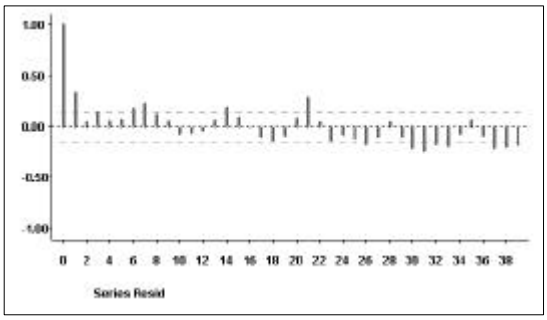
Daily traffic measurements in a cell of a GSM (Global System for Mobile) network



Same plot but with 0's instead of missing data (due to equipment failures, often during week ends)



A standard tool for studying such data:
Correlogram of the raw data
(with missing values replaced by 0)



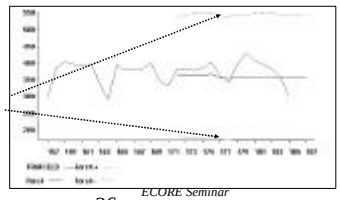
ARIMA model building technology (Box-Jenkins)

⇨Model obtained is defined by the equation

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

where y_t is the variable (traffic) and e_t is the forecast error

⇨Poor forecasts are obtained (shown with 95 % forecast intervals)



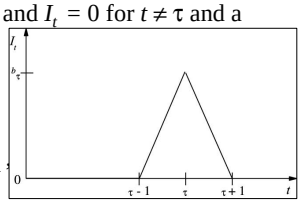
But the residual analysis reveals ...

```
=== SUMMARY MEASURES  <V>
SUM OF SQUARES :     COMPUTED =  .130599E+07  ADJUSTED =  .130322E+07
VARIANCE ESTIMATES :   BIASED =  7621.18      UNBIASED =  7757.28
TOTAL  NUMBER  OF  PARAMETERS =  3  STANDARD DEVIATION =  88.0754
INFORMATION CRITERIA :    AIC =  2022.16          SBIC =  2034.72

=== RESIDUAL ANALYSIS WITH 171 RESIDUALS, BEGINNING AT TIME       1===
MEAN =  .463096E-01 ,T-STATISTIC =    .01        (FOR TESTING ZERO MEAN)
 OUTLIERS <R(OR)S>
  < .01 %      82: -396.8         128: -362.0         149: -366.5
 .01-.2 %      79: -329.6         125: -298.8         134: -304.1
  1 - 5 %      80: -213.0          83:  179.7         126:  185.5
              129: -216.9         132:  172.9         135:  204.5
              150: -215.8
. . .
=== FORECASTING FROM      171  WITH FRESH DATA <F>
   DATE   OBSERVATION   FORECAST      ERROR  % ERROR  95% FORECAST INTERVAL
    172        383.04     361.49      21.55                188.86     534.11
 . . .
    185        299.71     357.60     -57.89                169.90     545.30
 MEAN ABSOLUTE ERROR (MAE):     29.10
```


AO effect (*additive outlier*)
Pointwise intervention

⇨A pointwise intervention or *pulse* at time τ is implemented by a regression term with a binary variable $I_t = 1$ for $t = \tau$ and $I_t = 0$ for $t \neq \tau$ and a parameter b_τ



⇨In the simplest case:
$y_t = b_t I_t + e_t$

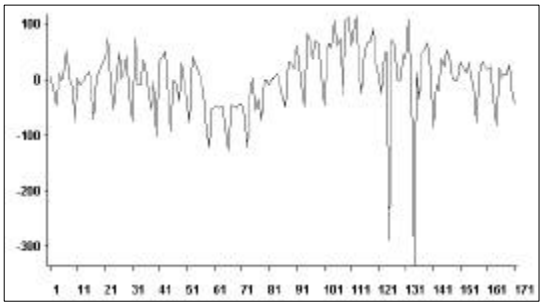
⇨In the example
$y_t - 0{,}9\, y_{t-1} = e_t - 0{,}5 e_{t-1}$
we write

$$(y_t - b_t I_t) - 0{,}9\,(y_{t-1} - b_{t-1} I_{t-1}) = e_t - 0{,}5 e_{t-1}$$

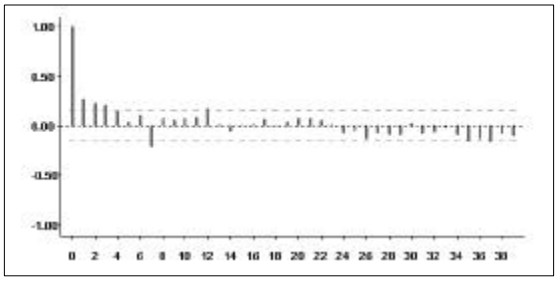
or

$$y_t = b_t I_t + \frac{1-0.5B}{1-0.9B} e_t$$

The residuals from the first model with 7 interventions, with just a constant term



Correlogram of the residual series



Output from Time Series Expert

```
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).
```

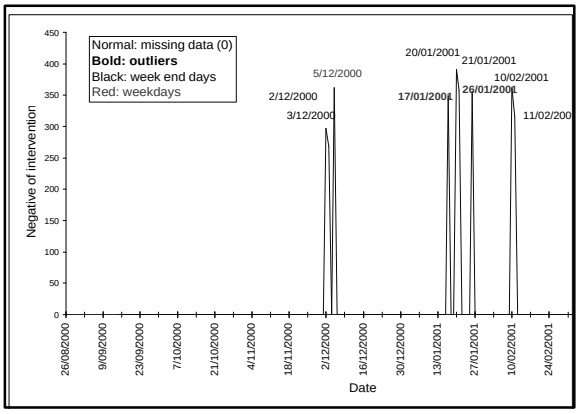
For the intervention analysis in TSE, see Azrak (1992)

Fitted model

```
=== 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
```


Estimates

```
=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
```

Model equation

Missing data:
all during week ends except
82: Tue Dec 5, 2000

Outliers
125: Wed Jan 17, 2001
134: Fri Jan 26, 2001

$$y_t = -0.12 - 297 I_t^{79} - 268 I_t^{80} - 362 I_t^{82}$$
$$-349 I_t^{125} - 392 I_t^{128} - 358 I_t^{129}$$
$$-355 I_t^{134} - 362 I_t^{149} - 316 I_t^{150}$$
$$+ \frac{1 - 0.88 L^7}{(1 - L^7)(1 - 0.38L - 0.15L^2 - 0.31L^3)} e_t$$

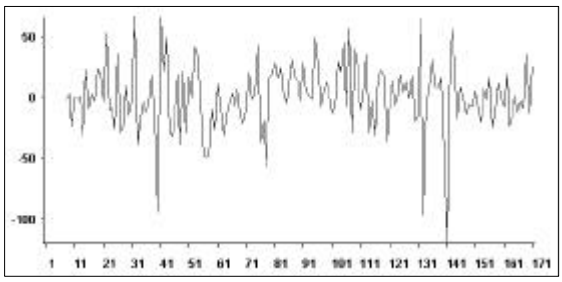
with $\hat{s} = 29{,}0$

(instead of 88,1 for the previous model)

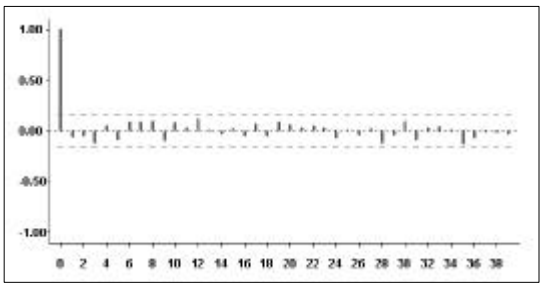
Residual analysis

```
=== 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
```

Residuals from the final model



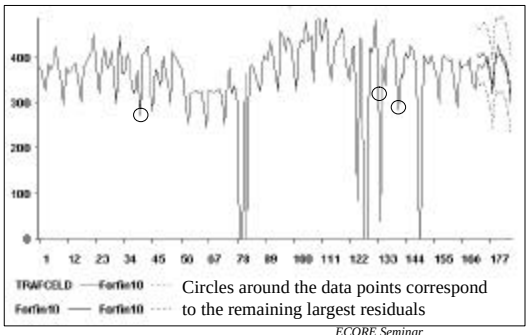
Residual correlogram of the final model



Forecasts (shown later)

```
---> WRITTEN TO FILE : Resfin8.DB , LENGTH =  164
=== 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
 MEAN ABSOLUTE ERROR (MAE):     15.44
```


Plot of the data and the forecasts for the last two weeks



Circles around the data points correspond to the remaining largest residuals

Zoom over the last two weeks (unused in model building)

TSE/AX expert system

- ⇨The same data were run using TSE/AX (Mélard-Pasteels, 2000, extended to daily and other series by Njimi et al., 2003)
- ⇨. . . but in a purely automated way (just specifying that periodicity is 7 days)
- ⇨The software tries to identify outliers in a series of modelling steps and build an ARIMA model with interventions
- ⇨In the present case, the final model was **identical**

Conclusion of Example 2

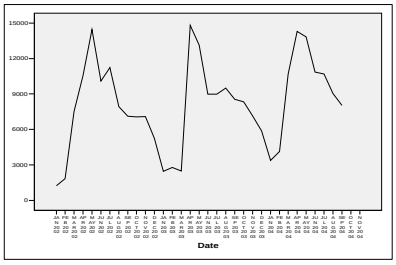
- ⇨Telecommunication data are often affected by missing data or outliers
- ⇨Statistical automated treatment may fail to take care of that
- ⇨Short term forecasts can be improved using the appropriate methodology, described here, even in an automated way

Plan of the talk

Example 3

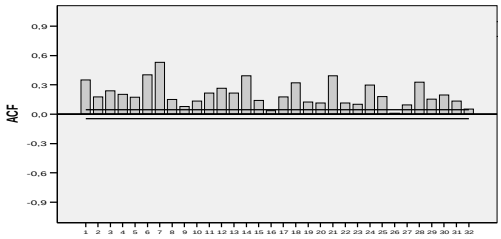
- ⇨Sales of lettuce (in units) in a given supermarket
- ⇨Daily data but most (not all!) Sundays, the shop is closed as well as on certain weekdays (holidays)
- ⇨First look at monthly data
- ⇨Then examine the crude time series
- ⇨Then propose an alignment procedure

Analysis of monthly data



- ⇨Clear seasonality. Modelling is easy but short series

Analysis of crude series



- ⇨In ACF, weekly (i.e. lag 7) seasonality is not very apparent

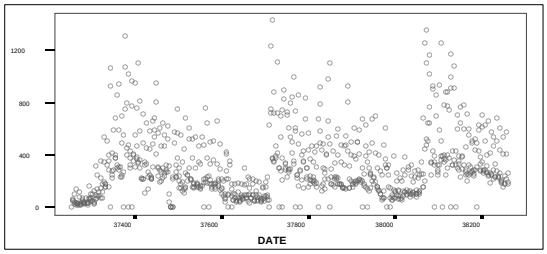
Preliminary treatment

- ⇨Alignment: sales during working Sundays are cumulated with those of the previous Saturday
- ⇨Holidays during the week are identified by a binary variable HOLIDAY = 1
- ⇨Working Sundays are identified by a binary variable SAT_SUN = 1
- ⇨Promotions are identified by a binary variable PROMO = 1
  (in this case it was defined *a posteriori*)

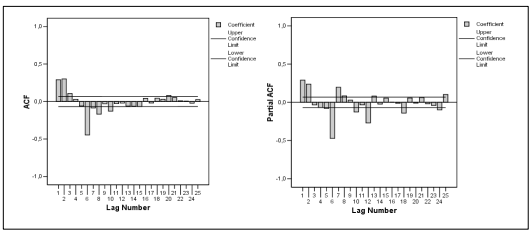
Example of coding

DAY	LETTUCE	HOLIDAY	SAT_SUN	PROMO	Comment
20020101	0	1	0	0	Holiday during the week
…					
20021221	215	0	0	0	
20021221	425	0	1	0	Working Sunday
20021223	409	0	0	0	
20021224	279	0	0	0	
20021225	0	1	0	0	Holiday during the week
20021226	93	0	0	0	
20021227	104	0	0	0	
20021228	220	0	0	0	
20021230	361	0	0	0	
20021231	329	0	0	0	
20030101	0	1	0	0	Holiday during the week
…					
20030404	1238	0	0	1	Promotion
20030405	1432	0	0	1	Promotion
…					

Plot of treated data



ACF and PACF after treatment and weekly seasonal difference



⇨Add AR(2) and SMA(1) terms

Model fitting

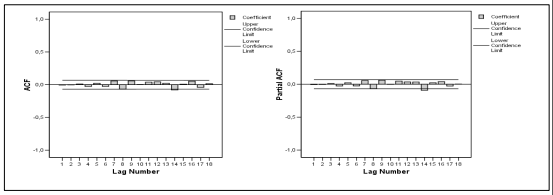
⇨Model equation is:

$$f(L)\nabla_6(\text{LETTUCE}_t - b_0 - b_1\text{HOLIDAY}_t - b_2\text{SAT\_SUN}_t - b_3\text{PROMO}_t) = q(L)e_t.$$

⇨Output from SPSS:

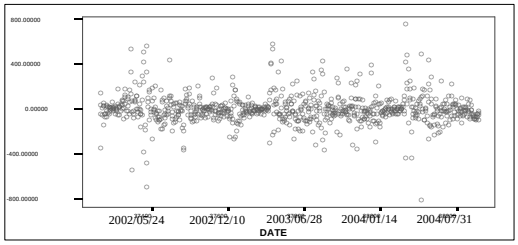
```
FINAL PARAMETERS:
Number of residuals  839
Standard error       133,35064
          Variables in the Model:
                    B          SEB       T-RATIO   APPROX. PROB.
AR1           0,39409    0,033945     11,609642      0,00000000
AR2           0,23857    0,033892      7,039183      0,00000000
SMA1          0,66832    0,026289     25,421937      0,00000000
HOLIDAY     -380,25777   24,384661    -15,594138      0,00000000
SAT_SUN      194,37259   51,045451      3,807834      0,00015049
PROMO        188,32126   24,253753      7,764624      0,00000000
CONSTANT       1,77421    4,206028      0,421826      0,67326106
```


Model adequacy



⇨Model is accepted up to a point …

Residual series



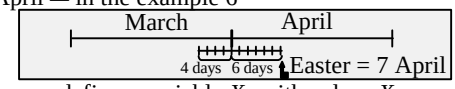
⇨The analysis should be improved at least by price effect

Easter effect (1/2)

- ⇨Easter is an important mobile holiday in western countries, less than Christmas and New Year, but …
- ⇨It is peculiar in the sense that it moves between March and April, hence between quarters I and II
- ⇨It is therefore not surprising that its position within a year has en impact on seasonality
- ⇨Corrections are usually considered on the basis of a period of n days before Easter, where n should be chosen, and counting the number of days of that period falling in March and in April, for example $n = 10$

Easter effect (2/2)

- ⇨More concretely, we can proceed as follows:
  - ✓let $n_{\text{March},t}$, the number among these n days falling in March — in the example 4
  - ✓let $n_{\text{Aprrl},t}$, the number among these n days falling in April — in the example 6



- ⇨Then we define a variable X_P with values X_{Pt}
  - ✓For t in March, we define $X_{Pt} = n_{\text{March}}/n$ — here 0,4
  - ✓For t in April, we define $X_{Pt} = n_{\text{April}}/n$ — here 0,6
  - ✓For t in another month, we define $X_{Pt} = 0$

Conclusion of Example 3

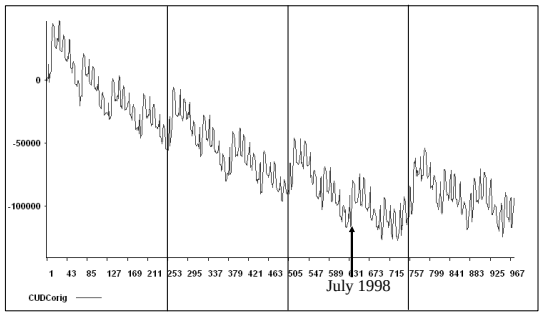
- ⇨Holidays can be taken into account by at least one binary variable (preferably more)
- ⇨Other calendar effects (working Sundays) are essential
- ⇨Weekly seasonal effect is also needed
- ⇨Preferably yearly seasonal effect should be included (not possible here because the series is too short)
- ⇨But then yearly alignment should also be considered
- ⇨And take care of Easter (and other moving holidays) like in Census X-12 seasonal adjustment program for monthly data

Plan of the talk

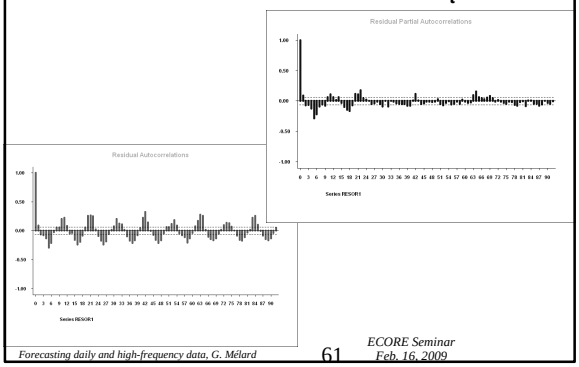
Example 4

- ⇨Analysis of treasury financial flows
- ⇨Objective: be able to predict them in order to avoid liquidity problems
- ⇨Case of the French social security in France (Laurent, 2006): ACOSS ("Agence centrale des organismes de sécurité sociale") covers 350 SS organisations, with transfers of 330 billions EUR in 2007
- ⇨We analyse the treasury balance (CUDC = "compte unique de disponibilités courantes"), starting with a approach like in the original work but simplified

Original data (1996-1999) in MFFR



ACF and PACF of $\tilde{N}CUDC_t$

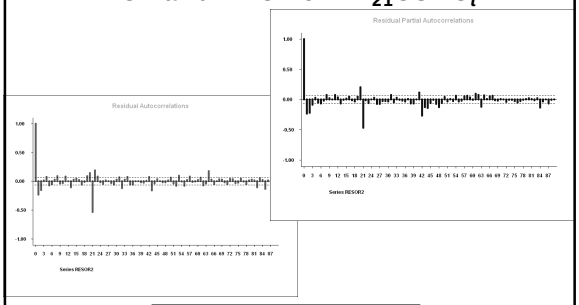


Problem of length of month

- ⇨Because different length of months, calendar effects and also that there is no bank transfer on Saturdays, Sundays and holidays, the number of observations per month is quite variable: between 18 and 23
- ⇨More specifically for 1996-1999:

  1×18, 7×19, 1×20, 16×21, 13×22, 4×23
- ⇨The most obvious "period": 21

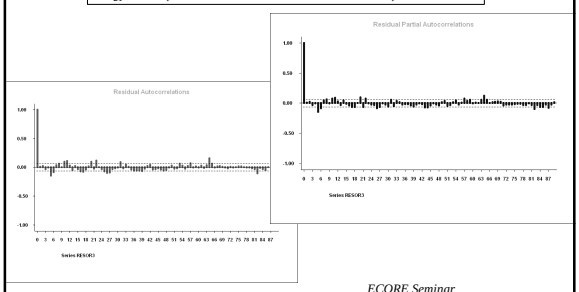
ACF and PACF of $\tilde{N}\tilde{N}_{21}CUDC_t$



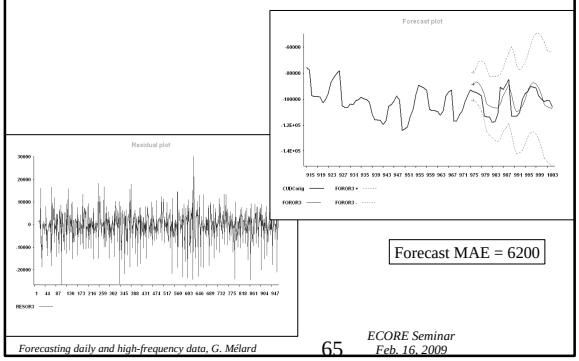
⇨Simple fit: $\nabla\nabla_{21}\mathrm{CUDC}_t = (1-q_1L-q_2L^2)(1-\Theta_1L^{21})e_t$

Residual ACF and PACF

$\nabla\nabla_{21}\mathrm{CUDC}_t = (1-0.21L-0.18L^2)(1-0.82L^{21})e_t, \quad \hat{s} = 6211$



Residuals and forecasts of CUDC



Forecast MAE = 6200

Alignment / rescaling

- ⇨In order to solve the length-of-month problem, some alignment or rescaling can be considered
- ⇨Here months alignment over 22 days (21 days later)
- ⇨Adding or deleting days in a <u>quiet period of the month</u>

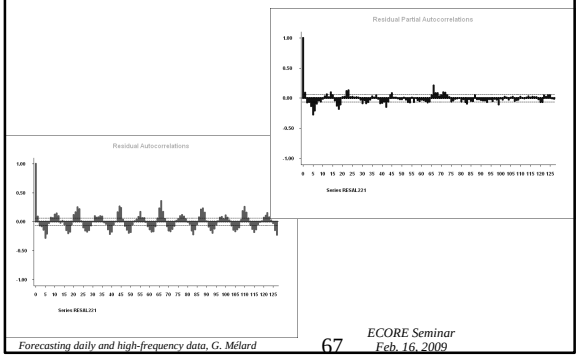
20	1996/02/28	3	16052636211
21	1996/02/29	4	15652013744
22			15251391277
1	1996/03/01	5	16604483595

20	1996/10/28	1	-29616554764
21	1996/10/29	2	-29003399625
22	1996/10/31	4	-29789879242
1	1996/11/04	1	-27134608765

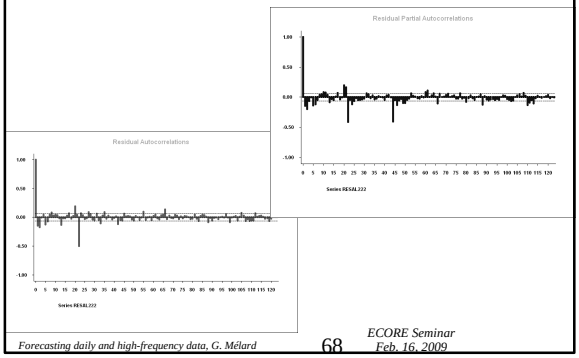
19	1997/08/28	4	-66758652660
20	1997/08/29	5	-66901395567
21			-67044138474
22			-67186881381
1	1997/09/01	1	-60142905134

18	1996/05/30	4	-6202145402
19	1996/05/31	5	-6668773881
20			-7135402359
21			-7602030838
22			-8068659316
1	1996/06/03	1	-5924878435

ACF and PACF of ÑCUDC22*t*

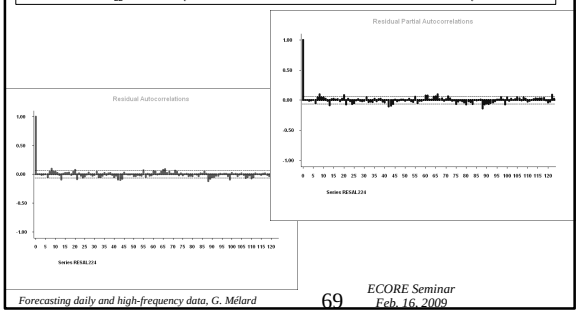


ACF and PACF of $\tilde{N}\tilde{N}_{22}$CUDC22$_t$

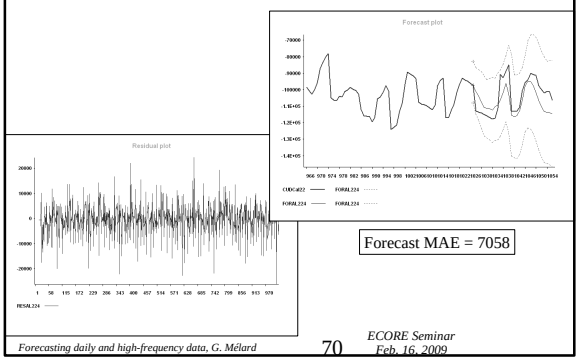


Residual ACF and PACF

$$(1+0.10L^5)\nabla\nabla_{22}\text{CUDC22}_t = (1-0.26L-0.17L^2)(1-1.22L^{22}+0.14L^{44})e_t, \quad \hat{s} = 5491$$

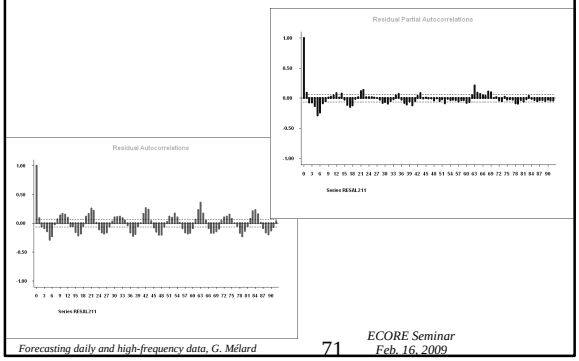


Residuals and forecasts of CUDC22

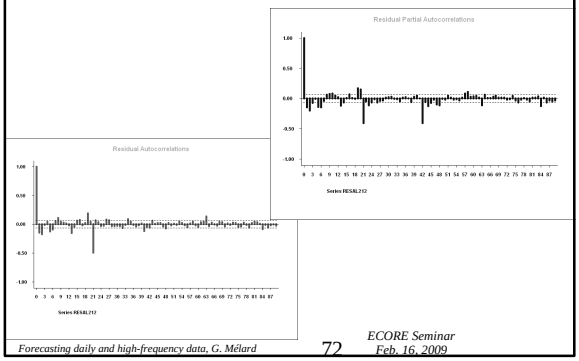


Forecast MAE = 7058

ACF and PACF of ÑCUDC21$_t$

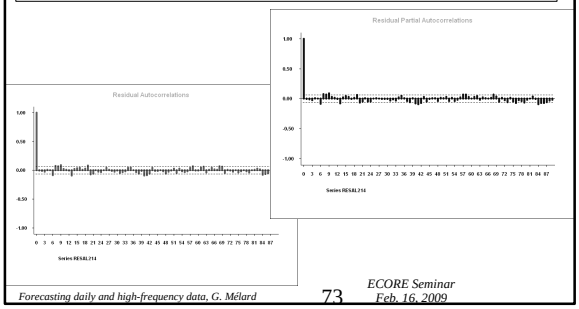


ACF and PACF of $\tilde{N}\tilde{N}_{21}$CUDC21$_t$

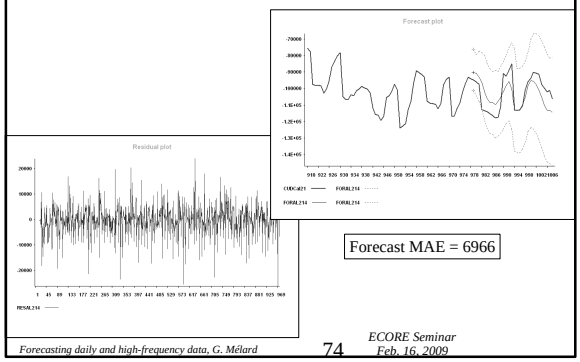


Residual ACF and PACF

$$(1+0.13L^5)(1-0.15L^{63})\nabla\nabla_{21}\text{CUDC21}_t = (1-0.26L-0.17L^2)(1-0.99L^{21})e_t,\ \hat{s} = 5550$$



Residuals and forecasts of CUDC21



Forecast MAE = 6966

Conclusion of Example 4

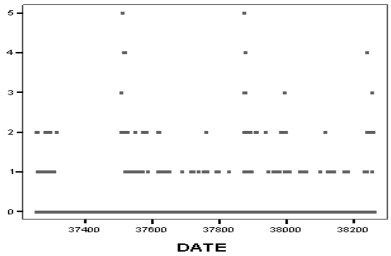
- ⇨Even if the fit is better (smaller $\hat{s}$), here the alignment doesn't improve forecasting over the test period but the models are more intellectually satisfactory
- ⇨The weekly seasonal lag polynomial (lag 5) is included in both models (ACOSS horizon = 5 days)
- ⇨The last one even includes a quarterly seasonality effect (lag 63 = 3×21)
- ⇨There should be a yearly seasonality (but only 4 years)
- ⇨There are outliers, some related to the changes after July 1998

Plan of the talk

Example 5

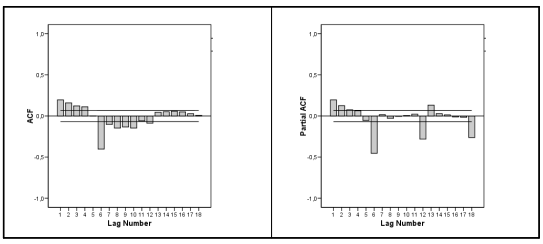
- ⇨Daily sales of four-season continental quilts (duvets) in a given supermarket (same as Example 3)
- ⇨Same remarks (holidays, working Sundays)
- ⇨Hence same treatment:
  - ✓alignment procedure
  - ✓binary variables
- ⇨Big difference: more costly than lettuces, hence daily sales are often 0 and sometimes 1, 2, …

Data



Specification

⇨We proceed as with the lettuce data: weekly seasonal difference, with same model specification



Model fitting

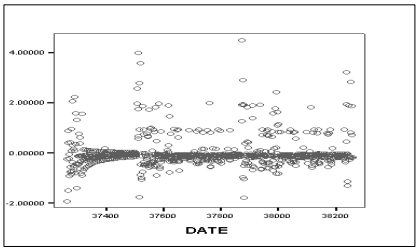
⇨Model equation is:

$$f(L)\nabla_6(\mathrm{DUVET}_t - b_0 - b_1\mathrm{HOLIDAY}_t - b_2\mathrm{SAT\_SUN}_t) = q(L)e_t.$$

⇨Output from SPSS:

```
FINAL PARAMETERS:
Number of residuals  839
Standard error       .57062622
Variables in the Model:
                        B          SEB       T-RATIO   APPROX. PROB.
AR1           .23827211   .03338504     7.137091       .00000000
AR2           .23707488   .03351499     7.073698       .00000000
SMA1          .99094317   .02374883    41.725983       .00000000
HOLIDAY      -.25356405   .11609014    -2.184200       .02922575
SAT_SUN      -.23298911   .24544895     -.949237       .34277562
CONSTANT     -.00022978   .00098416     -.233472       .81545193
```


Residual series



⇨The analysis is poor because of the integer data

Methods for time series of counts (1)

⇨There is some literature on regression for count data

⇨But dealing with time series is more challenging, see Kedem & Fokianos (2002)

⇨Zeger (1988) has a regression method which assumes a Poisson distribution and is based on quasi-likelihood

⇨INAR (INteger valued AR) processes, by Al-Osh & Alzaid (1987), McKenzie (1988), here INAR(1)

$$y_t = f \circ y_{t-1} + e_t, \qquad \text{where } f \in [0,1],\ y_0 \sim \text{Poisson}(l)$$

$$f \circ y = \sum_{i=1}^{y} \text{Bernoulli}_i(f),\ e_t \sim \text{Poisson}((1-f)\,l)$$

e_t and y_{t-1} are independent, and $\text{Bernoulli}_i(f)$ are i.i.d.

Methods for time series of counts (2)

⇨The INAR approach can be used for simple models only. No provision for seasonality, interventions

⇨Other promising approach by Davis, Dunsmuir, Streett (2003, 2005): GLARMA (Generalized linear ARMA) processes

$$y_t \mid \mathbf{F}_{t-1} \sim \text{Poisson}(l_t),\ \mathbf{F}_{t-1} = s\,(y_s, s \le t-1)$$

$$\log(l_t) = q_0 + q(L)e_t \quad \text{where } e_t = y_t - l_t$$

⇨Here AR with a constant but can be extended to regression with ARMA errors

⇨ML estimation does exist and justified by asymptotics

Methods for time series of counts (3)

⇨Difficult to allow for seasonality otherwise than in a deterministic (regression) way and other transformations than logs

⇨We are experimenting a simpler approach where

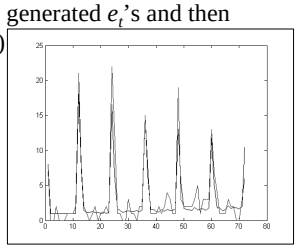
$$y_t \mid l_t, l_{t-1}, l_{t-2}, \ldots \sim \text{Poisson}(l_t) \text{ i.i.d.},$$

$$f(L)[\,f(l_t) - f(l)] = q(L)e_t, \quad \text{where } f \text{ is e.g. log or sqrt}$$

⇨This is similar to GLARMA processes except that e_t's are no longer related to observations y_t's

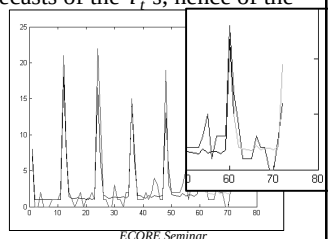
Methods for time series of counts (4)

- ⇨Example: let $(1-\Phi L^{12})[\sqrt{l_t}-\sqrt{l}\,]=e_t$
- ⇨Starting from a sequence (l_t; t = 1…12) = (8,10×1,18) l = 3, and Φ = 0.9, we generated e_t's and then obtained l_t's (in black) and observations y_t's (in red)
- ⇨Then we estimate l and f



Methods for time series of counts (5)

- ⇨For that simulation: $\hat{l}=3.35$ (instead of l = 3) and $\hat{\Phi}=0.79$ (instead of Φ = 0.9)
- ⇨and we deduce the forecasts of the l_t's, hence of the y_t's (in green)

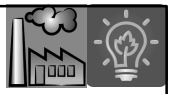


Conclusion of Example 5

- ⇨There remains to obtain more evidence (by Monte Carlo simulations) and to establish asymptotic properties
- ⇨Besides being able to model time series counts, we have to combine the method with the other features to be able to handle daily and high-frequency data
- ⇨In particular: adding regressors and/or interventions
- ⇨Work is still in progress

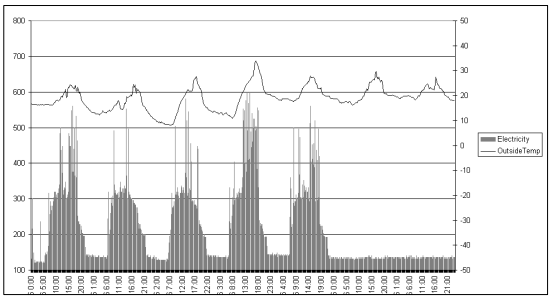
Plan of the talk

Example 6



- ⇨Electricity consumption of a company (but it can be fuel, water, …)
- ⇨Measurements every 15 minutes = forecast horizon
- ⇨Also outside temperature (or other measurements)
- ⇨Objective: issue warnings but only when needed!
- ⇨Here a few months of data ⇒ no yearly seasonality
- ⇨But working hours effect during the week (except on holidays)

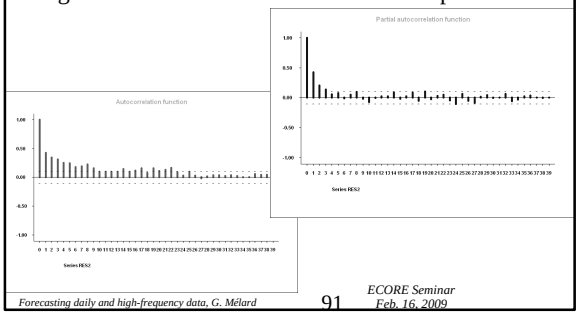
Data of some week



- ⇨We have used 4,5 days to fit an ARIMA model

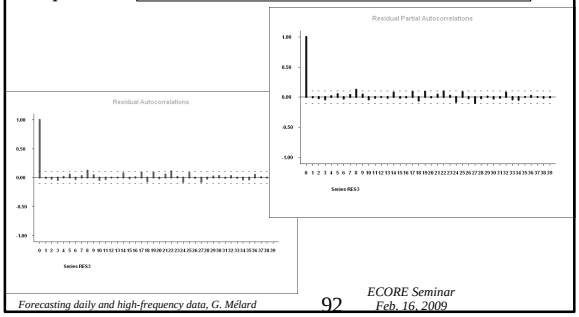
ARIMA approach: specification

⇨Log transform and seasonal difference of period 96

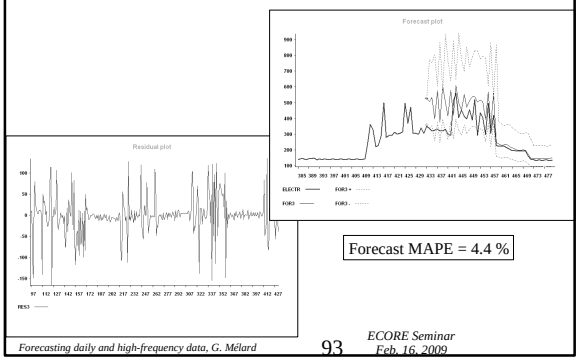


Model fit

⇨Equation: $(1-0.31L-0.16L^2-0.14L^3)\nabla_{96}\text{ELECTR}_t = e_t, \qquad \hat{s} = 44.0$



Residuals and forecasts for ELECTR



Forecast MAPE = 4.4 %

A regression approach using temperature

⇨The plot shows a working hour effect and a temperature effect

$$\text{Work}_{jt} = \begin{cases} 1 & \text{if} \quad j \le 5 \text{ and } t \ge 09{:}15 \text{ and } t \le 18{:}15 \\ \dfrac{t-07{:}00}{02{:}15} & \text{if} \quad j \le 5 \text{ and } t \ge 07{:}00 \text{ and } t < 09{:}15 \\ \dfrac{20{:}00-t}{01{:}45} & \text{if} \quad j \le 5 \text{ and } t > 18{:}15 \text{ and } t \le 20{:}00 \\ 0 & \text{if} \quad j=6 \text{ or } 7 \text{ or } (j \le 5 \text{ and } t \text{ elsewhere}) \end{cases}$$

⇨We also use temperature and electricity lagged by one

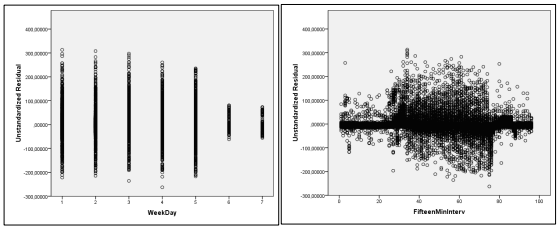
⇨. . . and the product Work*LaggedTemperature

Regression result for Electricity

Explanatory variables	Coefficients			
	Value	Std error	t-stat	P-value
Intercept	41.378	1.240	33.360	0.000
OutsideTemp_1	0.176	0.057	3.112	0.002
Electricity_1	0.696	0.006	119.666	0.000
Work	46.501	2.192	21.210	0.000
Work*OutsideTemp_1	1.343	0.104	12.898	0.000

$R^2 = 0.891$, $R^2_{adj} = 0.891$, DW = 2.38, $\hat{}$ = 41.9

Examination of residuals



⇨Scatter is clearly not constant w.r.t. day of week and time of the day

Conclusion of Example 6

- ⇨Residual standard deviation is similar in the two approaches (44.0 and 41.9), with probable small gain in combining them
- ⇨Clearly there is (marginal) heteroscedasticity in the two approaches
- ⇨Interventions on the standard deviation should help
- ⇨Here we have used a linear model, for simplicity, but a nonlinear effect of temperature should be studied
- ⇨Work is still in progress

Plan of the talk

References (1)

- ⇨ AL-OSH, M. A., ALZAID, A. A. "First order integer-valued autoregressive (INAR(1)) processes", *Journal of Time Series Analysis* **8**, 1987, 261-275
- ⇨ AZRAK, Rajae, MELARD, Guy, NJIMI, Hassane. "Forecasting in the analysis of mobile telecommunication data – Correction for outliers and replacement of missing observations", Journal Marocain d'Automatique, d'Informatique et de Traitement du Signal, numéro spécial CoPSTIC'03, novembre 2004, 1-14.
- ⇨ DAVIS , Richard A., DUNSMUIR, William T. M., STREETT, Sarah B. Observation-driven models for Poisson counts, *Biometrika* **90**, 2003, 777-790.
- ⇨ DAVIS, Richard A., DUNSMUIR, William T. M., STREETT, Sarah B. Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology and Computing in Applied Probability* **7**, 2005, 149-159.
- ⇨ KEDEM, Benjamin, FOKIANOS Konstantinos. *Regression Models for Time Series Analysis*, Wiley, New York, 2002.

References (2)

- ⇨ LAURENT, Emmanuel. *Optimiser la gestion de trésorerie par la modélisation économétrique des flux financiers*, Publibook, Paris, 2006.
- ⇨ McKENZIE, Ed. Some ARMA models for dependent sequences of Poisson cou,ts, *Advances in Applied Probability* **20**, 1988, 822-835.
- ⇨ MELARD, Guy. *Méthodes de prévision à court terme*, Editions Ellipses, Paris et Editions de l'Université de Bruxelles, Bruxelles, 2007.
- ⇨ MELARD, Guy, PASTEELS, Jean-Michel, "Automatic ARIMA modeling including interventions, using time series expert software", *International Journal of Forecasting* **16**, 2000, 497-508.
- ⇨ NJIMI, Hassane, MELARD, Guy, PASTEELS, Jean-Michel, "Modé-lisation SARIMA assistée", XXXVèmes Journées de Statistique, Société Française de Statistique, Lyon, France, 2003, pp. 731-734.
- ⇨ OUAKASSE, Abdelhamid, MELARD, Guy. "On-line estimation of the parameters of a time series model, with an application", submitted to Annals of Statistics, 2008.

Conclusion: Forecasting Daily and High-frequency Data

⇨ In principle, nothing new except that more parameters can be expected	⇨ *In practice, not really more parameters are needed*
⇨ Problems with missing data and outliers are more serious	⇨ *And models are not more complex*
⇨ Details are needed (calendar, events, …)	⇨ *Interventions can cope with them but at a price*
⇨ Explanatory variables can improve forecasts	⇨ *Regression with ARIMA errors*
⇨ Multiple "seasonal" periods	⇨ *But adequate explanatory variables may be needed*
⇨ But irregular length of months can raise problems	⇨ *Computing time not really a problem*
⇨ Modelling scatter is needed	⇨ *Alignment/rescaling can help*
	⇨ *Interventions should also help*
	⇨ *But automatic treatment is still a problem . . .*