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
0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a given supermarket
4 Example 4: Analysis of treasury financial flows
5 Example 5: Daily sales of four-season continental quilts
6 Example 6: Electricity consumption in a company
7 Conclusion
Plan of the talk

<table>
<thead>
<tr>
<th>8</th>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example 1: Production of electricity by windmills</td>
</tr>
<tr>
<td>2</td>
<td>Example 2: Network traffic in a GSM cell</td>
</tr>
<tr>
<td>3</td>
<td>Example 3: Sales of lettuce in a given supermarket</td>
</tr>
<tr>
<td>4</td>
<td>Example 4: Analysis of treasury financial flows</td>
</tr>
<tr>
<td>5</td>
<td>Example 5: Daily sales of four-season continental quilts</td>
</tr>
<tr>
<td>6</td>
<td>Example 6: Electricity consumption in a company</td>
</tr>
</tbody>
</table>

7 Conclusion

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

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

ARMA(1,2) residual analysis

Conclusion: no improvement according to Mélard (2007) (for an ARMA model at least)
Forecasting daily and high-frequency data, G. Mélard

**Plan of the talk**

0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a given supermarket
4 Example 4: Analysis of treasury financial flows
5 Example 5: Daily sales of four-season continental quilts
6 Example 6: Electricity consumption in a company
7 Conclusion

**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 of a GSM network for the period from September 15, 2000 to March 18, 2001 (185 observations)

**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

**Forecasts from ARMA(1,2) model**
- Conclusion: forecasts are not very good

**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 ...

AO effect (additive outlier) Pointwise intervention

A pointwise intervention or pulse at time \( \tau \) 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 \( b_\tau \)

In the simplest case:

\[ y_t = b_\tau I_t + e_t \]

In the example

\[ y_t - 0.9 y_{t-1} = e_t - 0.5 e_{t-1} \]

we write

\[ (y_t - b_\tau I_t) - 0.9 (y_{t-1} - b_{\tau-1} I_{t-1}) = e_t - 0.5 e_{t-1} \]

or

\[ y_t = b_\tau I_t + \frac{1 - 0.5 B}{1 - 0.9 B} e_t \]

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

Correlogram of the residual series
THE FOLLOWING PARAMETERS WERE ESTIMATED SEPARATELY

ESTIMATION HAS TAKEN .2 SEC. FOR 169 EVALUATIONS OF S.S.(MEAN TIME=,.001)

ITERATION STOPS - RELATIVE CHANGE IN EACH COEFFICIENT LESS THAN 1.00000E-03

AT TIME 171.14 FRESH DATA ARE RESERVED FOR EX-POST VALIDATION

WARNING *** MODEL FITTING IS PERFORMED WITH ONLY 171 DATA, ENDING

ANSECH-PC 2.3c, AUTHOR:G.MELARD 03/10/03 15:47:45. PROBLEM(    1): TRAF CELD

FINAL VALUES OF THE PARAMETERS

NAME IS TRAFCELD.DB LENGTH 185

SERIES READ FROM DISK, NAME IS TRAFCELD.DB LENGTH 185

DIRECTIVE     TYPE          DATE        STEP NATURE    PARAM/V ALUE COMMENTS
I 134: 0.100  BOX-TIAO          134          VALUE     KI 134: 0.100
I 125: 0.100  BOX-TIAO          125          VALUE     KI 125: 0.100
. . .
I  79: 0.100  BOX-TIAO           79          VALUE     KI  79: 0.100

===KNOWLEDGE ABOUT INTERVENTIONS (BOX-TIAO)

\[
\begin{align*}
\end{align*}
\]

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

Estimates

\[
\begin{align*}
\text{NAME} & \quad \text{VALUE} & \quad \text{STD ERROR} & \quad \text{T-VALUE} & \quad \text{LOWER} & \quad \text{UPPER} \\
\text{KI 134} & \quad 0.100 & \quad 0.12297268362 & \quad 1.03 & \quad -0.73 & \quad 2.33
\end{align*}
\]

Model equation

\[
y_t = -0.12 - 297.17T^7 - 268.04T^{12} - 362.14T^{12}
- 349.215 - 392.128 - 358.129 - 355.154 - 362.149 - 316.150
+ 1 - 0.88L^t
\]

\[
(1 - 0.38L - 0.15L^2 - 0.31L^3) e_t
\]

with \( \sigma = 29.0 \)

(Instead of \( \sigma = 88.1 \) for the previous model)

Residual analysis

\[
\text{SUM OF SQUARES : COMPUTED = 126778. ADJUSTED = 124169.}
\]

\[
nen = 114, \text{df} = 14, \text{MS} = 769.196, \text{F} = 28.45 \]

\[
\text{TOTAL NUMBER OF PARAMETERS = 14 STANDARD DEVIATION = 28.9948}
\]

\[
\text{SUMMARY MEASURES}
\]

\[
\text{OUTLETS: 123, WEEL-JAN 17, 2001 134, FRI-JAN 26, 2001}
\]

\[
\text{OUTLIERS: 134, FRI-JAN 26, 2001}
\]

\[
\text{SUMMARY MEASURES: COMBINED - 126778. ADJUSTED - 124169.}
\]

\[
\text{TOTAL NUMBER OF PARAMETERS = 14 STANDARD DEVIATION = 28.9948}
\]

\[
\text{SUMMARY MEASURES: COMBINED - 126778. ADJUSTED - 124169.}
\]

\[
\text{TOTAL NUMBER OF PARAMETERS = 14 STANDARD DEVIATION = 28.9948}
\]

\[
\text{SUMMARY MEASURES: COMBINED - 126778. ADJUSTED - 124169.}
\]

\[
\text{TOTAL NUMBER OF PARAMETERS = 14 STANDARD DEVIATION = 28.9948}
\]
Residuals from the final model

Residual correlogram of the final model

Forecasts (shown later)

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

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
Forecasting daily and high-frequency data, G. Mélard

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

0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a given supermarket
4 Example 4: Analysis of treasury financial flows
5 Example 5: Daily sales of four-season continental quilts
6 Example 6: Electricity consumption in a company
7 Conclusion

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.)
Forecasting daily and high-frequency data, G. Mélard

**Example of coding**

<table>
<thead>
<tr>
<th>DATE</th>
<th>LATITUDE</th>
<th>HOLIDAY</th>
<th>SAT</th>
<th>SUN</th>
<th>PROMO</th>
<th>HOLIDAY</th>
<th>SMA</th>
<th>AR</th>
<th>12</th>
<th>1432</th>
</tr>
</thead>
<tbody>
<tr>
<td>20020101</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020102</td>
<td>215</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020103</td>
<td>435</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020104</td>
<td>490</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020105</td>
<td>279</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020106</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020107</td>
<td>186</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020108</td>
<td>220</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020109</td>
<td>241</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020110</td>
<td>320</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020112</td>
<td>1228</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20020113</td>
<td>1432</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ACF and PACF after treatment and weekly seasonal difference**

⇒ Add AR(2) and SMA(1) terms

**Model fitting**

⇒ Model equation is:

\[ \text{P(LV, LETTUCE, -6 -12 HOURLY, SAT, SUN, -6 FRI, -6 PROMO) = 0 (LV)} \]

⇒ Output from SPSS:

<table>
<thead>
<tr>
<th>Variables in the Model:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>SE</td>
<td>T-RATIO</td>
<td>APPROX. PROB.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR1</td>
<td>0.39409</td>
<td>0.03345</td>
<td>11.659642</td>
<td>0.00000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR2</td>
<td>0.34857</td>
<td>0.03362</td>
<td>7.059683</td>
<td>0.00000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMA</td>
<td>0.64832</td>
<td>0.10245</td>
<td>25.32337</td>
<td>0.00000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOLIDAY</td>
<td>-383.25777</td>
<td>24.38461</td>
<td>-15.594138</td>
<td>0.00000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT, SUN</td>
<td>546.27259</td>
<td>54.45511</td>
<td>3.207334</td>
<td>0.0002049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROMO</td>
<td>198.32126</td>
<td>24.25735</td>
<td>7.764424</td>
<td>0.00000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSTANT</td>
<td>-1.7742</td>
<td>4.204038</td>
<td>0.422646</td>
<td>0.67330194</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 an 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 \( n_{\text{March} \cap t} \), the number among these \( n \) days falling in March — in the example 4
  ✓ let \( n_{\text{April} \cap t} \), the number among these \( n \) days falling in April — in the example 6

\[
\begin{array}{|c|c|}
\hline
\text{March} & \text{April} \\
\hline
4 & 5 \quad \cdots \quad 7 \quad \cdots \quad 9 \quad \cdots \quad 12 \\
\hline
\end{array}
\]

⇒ Then we define a variable \( \chi_t \) with values
  ✓ For \( t \) in March, we define \( \chi_t = \frac{n_{\text{March} \cap t}}{n} \) — here 0,4
  ✓ For \( t \) in April, we define \( \chi_t = \frac{n_{\text{April} \cap t}}{n} \) — here 0,6
  ✓ For \( t \) in another month, we define \( \chi_t = 0 \)

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

Original data (1996-1999) in MFFR

⇒ 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
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:
1x18, 7x19, 1x20, 16x21, 13x22, 4x23

The most obvious "period": 21

ACF and PACF of $\nu_CUDC_t$

Simple fit: $\nu_CUDC_t(1) = (1-0.1\ell_t-0.2\ell_{t-1})x_t$

Residual ACF and PACF

Residuals and forecasts of CUDC

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 quiet period of the month
Residual ACF and PACF

\[
1 + 0.10z^{-1} + 0.26z^{-2} + 0.17z^{-3} + 1.23z^{-4} + 0.14z^{-5}, \quad d = 540
\]

Residuals and forecasts of CUDC22

Forecast MAE = 7058
Residual ACF and PACF

\[ 1 = 0.13L^2y_t - 0.15L^3y_t \text{VAR}, \text{CUUC21}_t = (1 - 0.26L^2y_t - 0.17L^3y_t - 0.99L^4y_t) u_t, \phi = 550 \]

Residuals and forecasts of CUDC21

Forecast MAE = 6966

Conclusion of Example 4

- Even if the fit is better (smaller \( \hat{\sigma} \)), 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 = 3x21)
- 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

0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a given supermarket
4 Example 4: Analysis of treasury financial flows
   Straightforward analysis
   Alignment
5 Example 5: Daily sales of four-season continental quilts
   Treatment of count time series
6 Example 6: Electricity consumption in a company
7 Conclusion

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

Model fitting:
The model equation is:

$$\Delta L_{UV, ETH, HOLIDAYS AT_SUN} = \beta_1 + \beta_2 HOLIDAY + \beta_3 SAT_SUN + \epsilon$$

Output from SPSS:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AR1</td>
<td>0.238272</td>
<td>7.137091</td>
<td>0.000000</td>
</tr>
<tr>
<td>AR2</td>
<td>0.237075</td>
<td>7.073486</td>
<td>0.000000</td>
</tr>
<tr>
<td>SMA1</td>
<td>0.990943</td>
<td>41.72598</td>
<td>0.000000</td>
</tr>
<tr>
<td>HOLIDAY</td>
<td>-0.253564</td>
<td>-2.18420</td>
<td>0.029225</td>
</tr>
<tr>
<td>SAT_SUN</td>
<td>-0.232989</td>
<td>-0.94923</td>
<td>0.342776</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>-0.000229</td>
<td>-0.23347</td>
<td>0.815452</td>
</tr>
</tbody>
</table>

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)

$$Y = a + X + \epsilon$$

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

$$Y_i | F_{t-1} \sim \text{Poisson}(\lambda_i), \quad F_{t-1} = \sigma(y_s, s \leq t-1)$$

$$\log(\lambda_i) = \theta_0 + \theta(L)e_t$$

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

$$Y_i | \lambda_i, \ldots \lambda_i \sim \text{Poisson}(\lambda_i)$$

$$\phi(L)(f(\lambda_i) - f(\lambda)) = \theta(L)e_t$$

This is similar to GLARMA processes except that $e_t$'s are not any related to observations $y_t$'s.
Methods for time series of counts (4)

Example: let \((1-\Phi L)(\sqrt{\lambda}-\sqrt{\lambda})=\epsilon\)

\(\lambda = 3, \text{ and } \Phi = 0.9,\) we generated \(\epsilon_t's\) and then obtained \(\lambda_t's\) (in black) and observations \(y_t's\) (in red).

Then we estimate \(\lambda\) and \(\Phi\).

Methods for time series of counts (5)

For that simulation: \(\lambda = 3.35\) (instead of \(\lambda = 3\)) and \(\Phi = 0.79\) (instead of \(\Phi = 0.9\)).

And we deduce the forecasts of the \(\lambda_t's\), 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

0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a GSM cell
4 Example 4: Analysis of treasury financial flows
5 Example 5: Daily sales of four-season continental quilts
6 Example 6: Electricity consumption in a company
7 Conclusion

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: $y = 0.31x - 0.16z - 0.14w_{ELECTR} + e, \quad d = 44.6$

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

We also use temperature and electricity lagged by one and the product Work*LaggedTemperature

Regression result for Electricity

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficients</th>
<th>Std error</th>
<th>t-stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>41.378</td>
<td>1.240</td>
<td>33.360</td>
<td>0.000</td>
</tr>
<tr>
<td>OutsideTemp_1</td>
<td>0.176</td>
<td>0.057</td>
<td>3.112</td>
<td>0.002</td>
</tr>
<tr>
<td>Electricity_1</td>
<td>0.696</td>
<td>0.006</td>
<td>119.666</td>
<td>0.000</td>
</tr>
<tr>
<td>Work</td>
<td>46.501</td>
<td>2.192</td>
<td>21.210</td>
<td>0.000</td>
</tr>
<tr>
<td>Work*OutsideTemp_1</td>
<td>1.343</td>
<td>0.104</td>
<td>12.889</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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

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

0 Introduction
1 Example 1: Production of electricity by windmills
2 Example 2: Network traffic in a GSM cell
3 Example 3: Sales of lettuce in a given supermarket
4 Example 4: Analysis of treasury financial flows
   Straightforward analysis
   Alignment
5 Example 5: Daily sales of four-season continental quilts
   Example
   Treatment of count time series
6 Example 6: Electricity consumption in a company
7 Conclusion

References (1)
- AZRAK, Rajae; MELARD, Guy; NJIMI, Hassane. "Forecasting in the analysis of mobile telecommunication data – Correction for outliers and replacement of missing observations", Journal Muscatin d’Automatique, d’Informatique et de Traitement du Signal, numéro spécial COPSTIC’03, November 2004, 1-14

References (2)

Conclusion: Forecasting Daily and High-frequency Data

- In principle, nothing new except that more parameters can be expected.
- Problems with missing data and outliers are more serious.
- Details are needed (calendar, events, …).
- Explanatory variables can improve forecasts.
- Multiple “seasonal” periods.
- But irregular length of months can raise problems.
- Modelling scatter is needed.

In practice, not really more parameters are needed.
And models are not more complex.
Interventions can cope with them but at a price.
Regression with ARIMA errors.
But adequate explanatory variables may be needed.
Computing time not really a problem.
Alignment/rescaling can help.
Interventions should also help.
But automatic treatment is still a problem . . .