Supplementary notes

Contents

14 An example introducing new features in PROC UCM

14.1 Introducing the dataset for hourly traffic counts

Copenhagen is a city with many bicycles. People of all ages use the bike to reach for work, studies or leisure time activities independent of weather and the time of day. Bicycling is safe and fast because of bicycle roads along almost all major streets.

In this example the traffic is counted at Fredensbro (Danish "bro" means "bridge") which connects Nørrebro where many people live with the inner city of Copenhagen where many people work or study and where many activities like theatres, bars, dancing places, etc. are situated. Other people use the bikes for longer trips and just pass Fredensbro on their way.

This example uses data on the number of passing bicycles every hour in a period from January 1'st the hour from 00 to 01 (that is the first hour of the new year) and to March 19'th a total of 1872 observations in the winter and early spring 2014. The number of passing bicycles are counted by some automatic device. Such datasets are available for many streets of Copenhagen.

The dataset is named "cycling".

14.2 Seasonals in the dataset

The series are plotted at Figure 1 for just two weeks in order to see the details. The period is from Monday February 24'th to Sunday March 9'th. where no special holidays are present. Figure 2 and 3 are more detailed plots of Thursday February 27'th and Saturday March 1'st. The two series on the plots are the number of passing bicycles "to the city" and "from the city". It seems that the number of cyclists to the city is largest in the morning hours while the number of cyclists away from the city are largest in the afternoon hours. This is easily understood as a traffic pattern of people pending to and from work or studies.

Figure 1: The entire time series

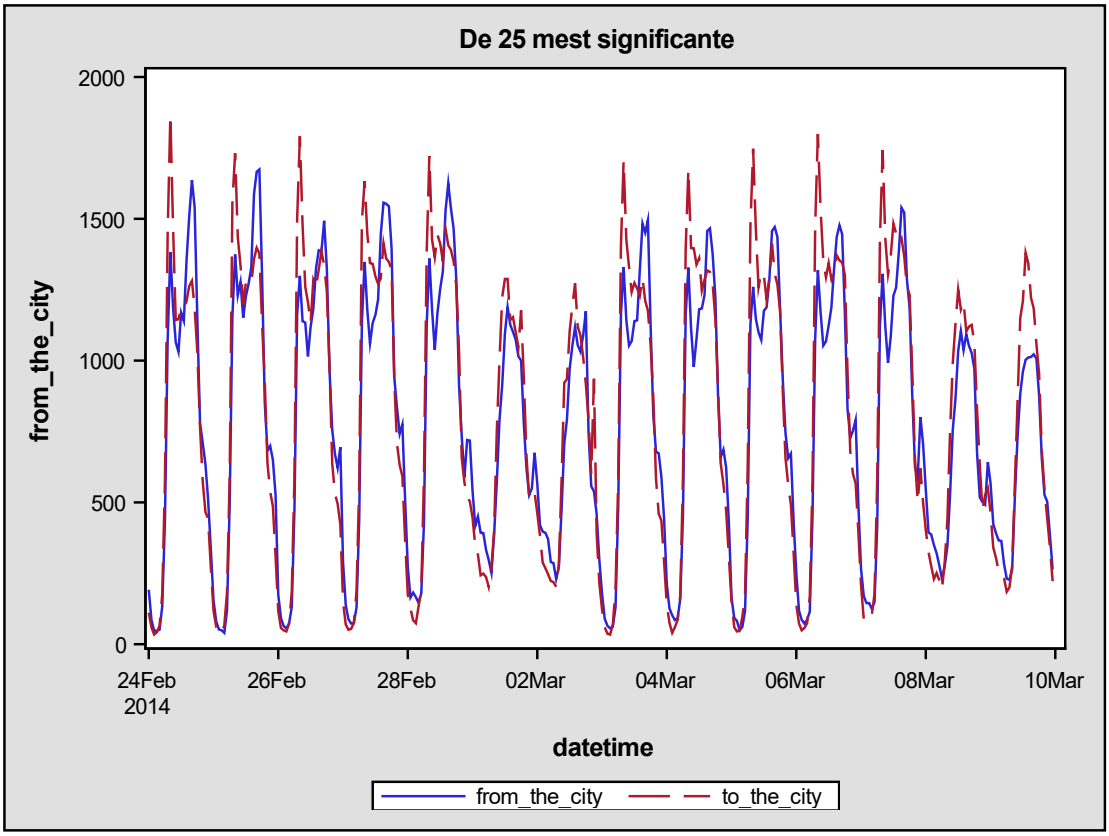


Figure 2: Thursday 27'th February

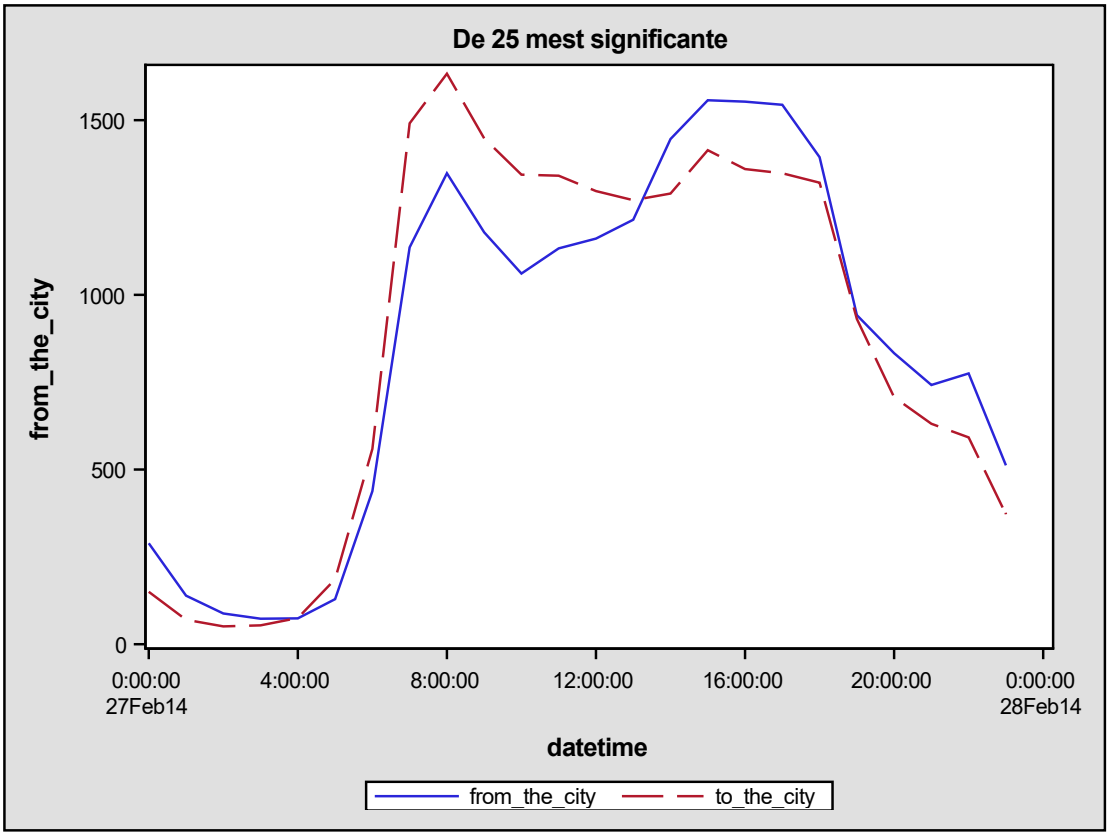
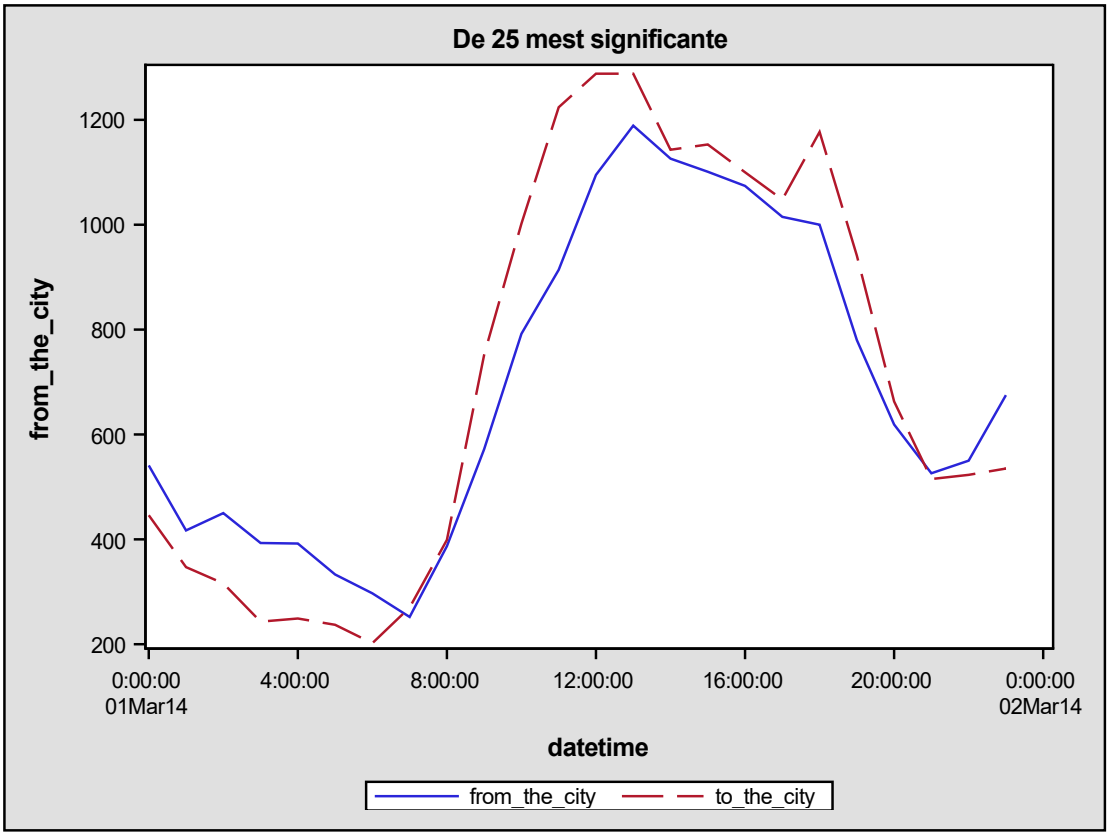


Figure 3: Saturday 1'st March



A weekday effect is of course present as most people are away from work and studies Saturday and Sunday - however many go for shopping, but a little later in the morning than on ordinary working days. Night life especially Friday night is also visible as the numbers of cyclists are large in the hours late Friday evening and early Saturday morning.

Looking at the plots of the series it is obvious that the level is rather constant and no trend is seen. When it comes to modelling by PROC UCM it is expected that a slope component is superfluous and that a level component of course is present but the level component variance is probably insignificant and could be excluded. The main problem is to model the seasonal structure. It is clear that a weekly component should be included in order to model the effect of working days and weekend days. Also an hourly effect is necessary as the traffic changes over the hours of the day.

Some autocorrelation is to be expected in this example using hourly observations. An argument for negative autocorrelation is that very few passes the bridge in the same direction in two consecutive clock hours; that is if you pass the bridge you probably will not go back soon after and then pass the bridge again. This means that if for some reason many people are one hour earlier than usual leading to high count one hour, they have crossed the bridge so not so many will pass the hour after. On the other hand it is possible that bad weather or traffic problems could increase or decrease the number of passing cyclists for many consecutive hours which could explain positive autocorrelation.

One possibility is to apply a seasonal component for all $24 \times 7 = 168$ hours a week. In principle this requires 168 seasonal dummies — which of course are much too many parameters. However, it is possible to reduce the number of parameters to a much smaller number by application of seasonal harmonics as in the Arctic ice example, Chapter 13. Another possibility is to apply a blockseason component which allows for a daily cycle of 24 hours to be embedded in a seven day weekly cycle. This possibility was introduced in PROC UCM by November 2018.

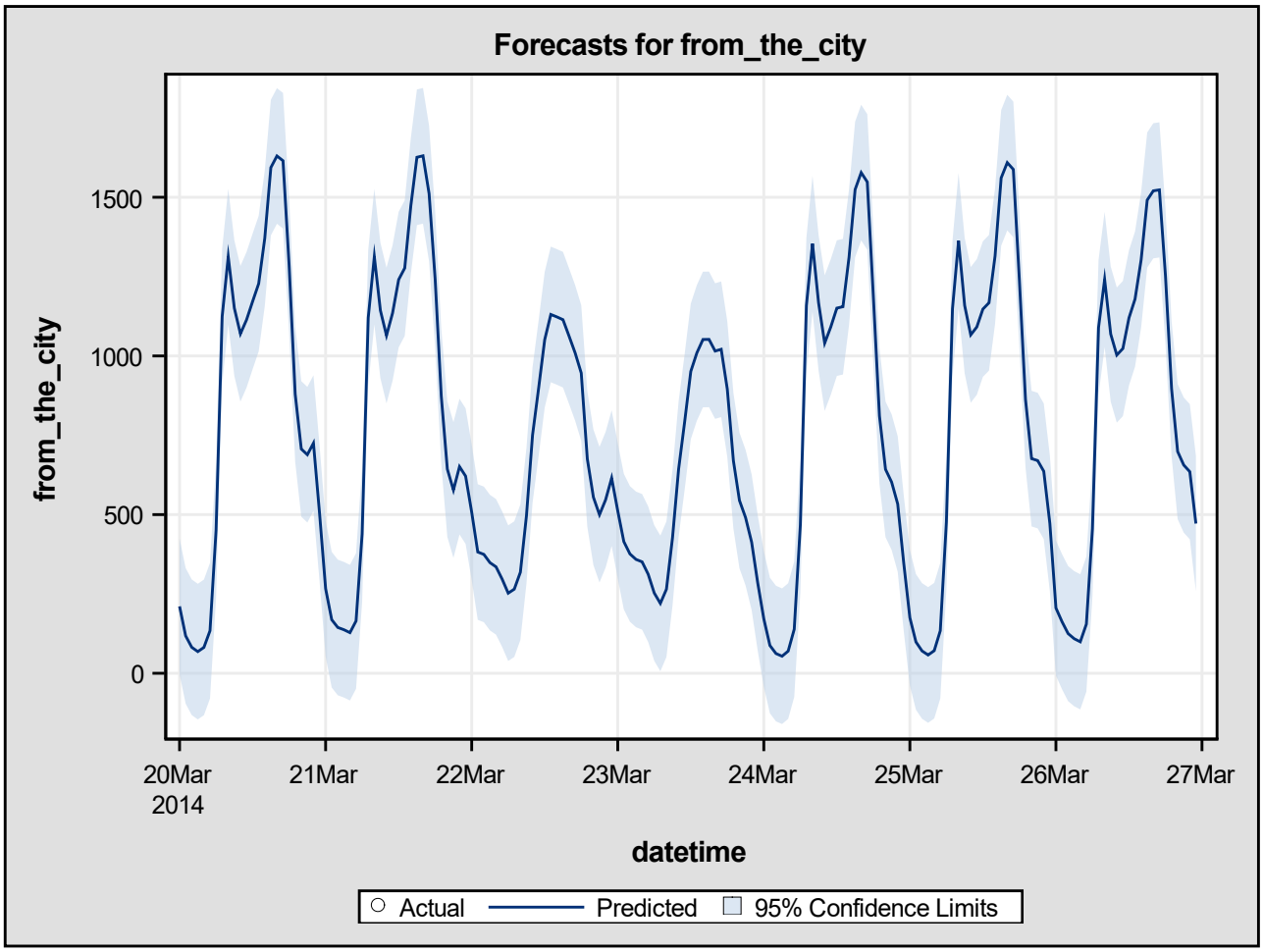
In the following all three possibilities are applied to the series of number of cyclists to the inner-city.

14.3 A length 168 seasonal component for each hour a week by dummies

As a benchmark model the following code just includes a constant level and dummies for the 168 hours a week. Both the level variance and the variance for the seasonal dummies are fixed as zero.

```
PROC UCM data=sasts.cycling;
id datetime interval=hour;
model from_the_city;
irregular;
level variance=0 noest;
season length=168 variance=0 noest;
estimate plot=(panel);
forecast lead=168 plot=forecasts;
```

Figure 4: Forecast a week ahead using all 168 hourly dummies. Note that March 22-23 are weekend days




```
run;
```

The output is of course too long to be printed here. The main point is that the huge amount of parameters gives a very precise picture of the hourly seasonal structure for a whole week. This is best seen by the forecast plot, Figure 4, where forecasts for the next 168 observations are plotted. All regular weekly patterns are seen, for instance it is obvious that the number of cyclists the nights after Friday and Saturday are larger than the nights before ordinary working days.

This model only includes one stochastic component, the irregular component,which has an estimated variance 10895. Of course it is possible, at least in principle, to include more stochastic components like a time varying level and a time varying slope. But the estimation and production of graphical output is rather time consuming, and it is often superfluous if the analysis does not point in the direction of specific extensions of the model.

In this example it is natural in some way to extend the model with an autoregressive term. This is done by extending the irregular statement by the p=1 option. The resulting estimated autoregressive parameter is $\varphi_1 = 0.85$ and the irregular component variance is reduced to 3125 from 10895 in the above model. But this code extends the computer time to more than a minute, so no further extensions is tried here.

14.4 A length 168 seasonal component for each hour a week using harmonics

Exactly the same model fit is obtained if the seasonal components instead of dummy variables are parametrized by harmonics. The fit is the same and the number of parameters to be estimated are also the same. The main advantage of this approach is the possibility to choose only the most important harmonics in a model and exclude the rest. In this way the number of important parameters could be reduced significantly.

The basic model is fitted by the code below with an autoregressive term included. The only difference is the seasonal statement where the method is changed from the default dummy parametrization to the parametrization by harmonics and the option print=harmonics gives a table showing the significance of all 168/2 = 84 harmonics. This table is saved as a data set named all_harmonics. This dataset is sorted by significance and printed in the last part of the code.

```
PROC UCM data=sasts.cycling;
id datetime interval=hour;
model from_the_city;
irregular p=1;
level plot=smooth var=0 noest ;
season length=168 type=trig print=harmonics var=0 noest;
estimate plot=(panel);
ods output SeasonHarmonics=all_harmonics;
run;
proc sort data=all_harmonics out=sort;
by descending chisq;
run;
proc print data=sort;
var harmonic period chisq;
run;
```

The first ten lines of table of the harmonics from the procedure output is shown in the Table.

Harmonic Analysis of Trigonometric Seasons (Based on the Final State)						
Name	**Season Length**	**Harmonic**	**Period**	**Chi-Square**	**DF**	**Pr > ChiSq**
Season	168	1	168.00000	92.40	2	<.0001
Season	168	2	84.00000	57.77	2	<.0001

Harmonic Analysis of Trigonometric Seasons (Based on the Final State)						
Name	**Season Length**	**Harmonic**	**Period**	**Chi-Square**	**DF**	**Pr > ChiSq**
Season	168	3	56.00000	1.75	2	0.4169
Season	168	4	42.00000	6.83	2	0.0328
Season	168	5	33.60000	111.51	2	<.0001
Season	168	6	28.00000	189.73	2	<.0001
Season	168	7	24.00000	8150.51	2	<.0001
Season	168	8	21.00000	87.99	2	<.0001
Season	168	9	18.66667	91.33	2	<.0001
Season	168	10	16.80000	8.79	2	0.0123

The 25 most important harmonics from the sorted table are shown in the Table.

Obs	Harmonic	period	ChiSq
1	7	24.00000	8150.51
2	21	8.00000	3543.76
3	35	4.80000	2479.17
4	22	7.63636	494.56
5	15	11.20000	460.98
6	49	3.42857	458.12
7	20	8.40000	439.54
8	34	4.94118	410.22
9	13	12.92308	366.59
10	42	4.00000	301.38
11	19	8.84211	289.79
12	16	10.50000	250.63
13	23	7.30435	240.32
14	63	2.66667	229.55
15	14	12.00000	225.08
16	27	6.22222	207.24
17	36	4.66667	205.37
18	6	28.00000	189.73
19	26	6.46154	169.34
20	37	4.54054	132.45
21	5	33.60000	111.51

Obs	Harmonic	period	ChiSq
22	33	5.09091	110.52
23	12	14.00000	97.55
24	28	6.00000	94.58
25	1	168.00000	92.40

The by far most significant harmonic is 7 which corresponds to period length 168/7 = 24, that is the variation within the 24 hours of a day. The next harmonics are either multiples of 7 or very close to multiples of 7, so they serve to modify the shape of the sinusoid as for instance the amplitude of the daily variation is less for Saturdays and Sundays than for ordinary working days. To study this behaviour using harmonics is called spectral analysis and is not treated any further in this note.

We use simply this table to pick out harmonics with high significance in order to reduce the number of parameters to estimate by excluding less significant harmonics.

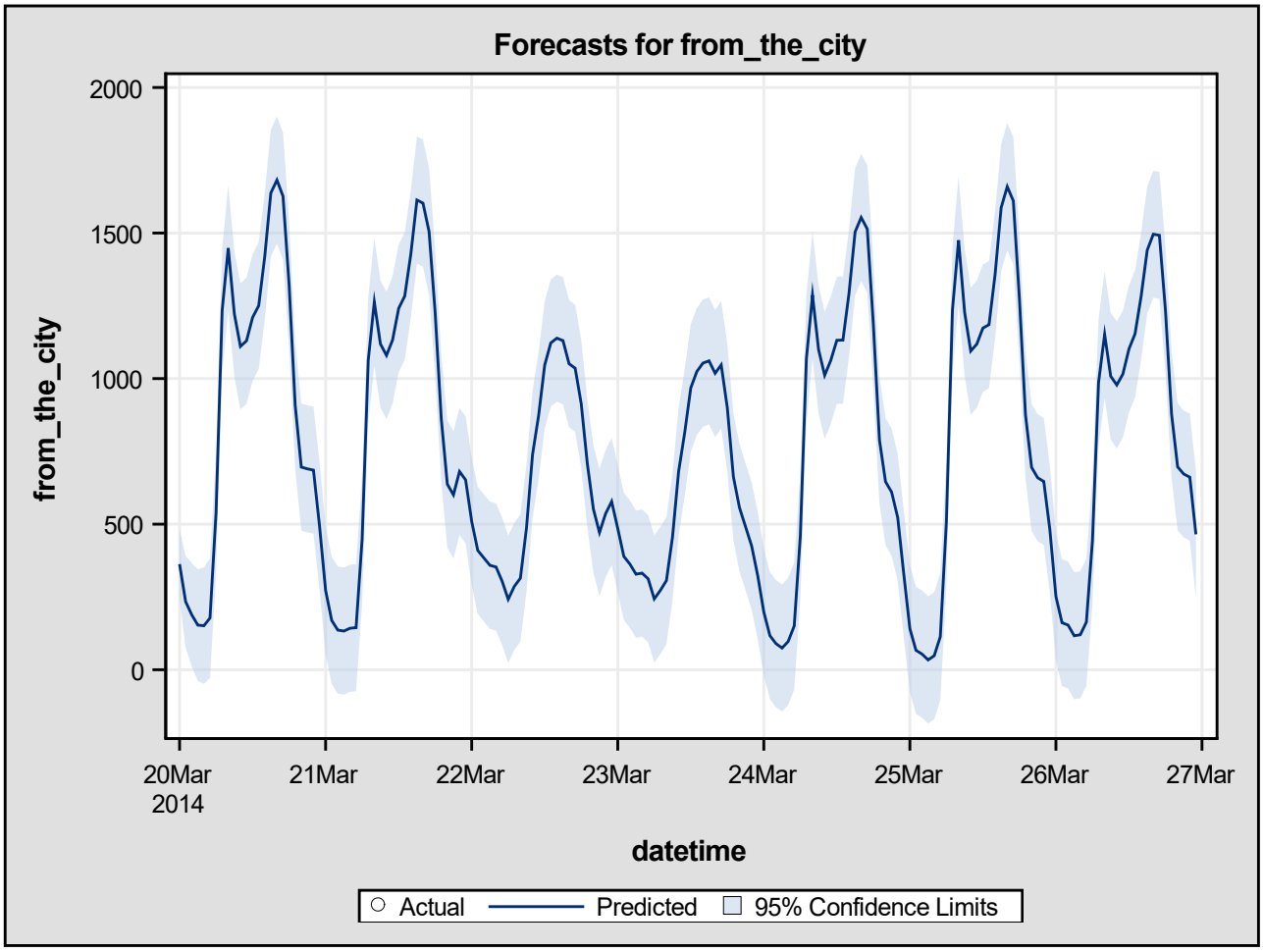
In the next code this is done by specification of which harmonics to keep by the option keeph=. In the code seasonal harmonics for 7, 14, . . . , 70 are applied and the nearby, plus-minus 1 and 2 harmonics for 5, 6, 8, 9, . . ., 45, 46, 48, 49 are also included. Note that all theses numbers are easily coded using notation as keeph=7 to 70 by 7. Moreover harmonics 1 and 2 are included to model long term behaviour. In total 41 harmonics out of the 84 possible harmonics are included.

These choices of which harmonics to include are of course rather arbitrary, as several highly significant harmonics are left out. But most of the significant harmonics from the table are included and the list only includes a few harmonics with small χ^2-square statistics.

```
PROC UCM data=sasts.cycling ;
id datetime interval=hour;
model from_the_city;
irregular p=1;
level var=0 noest ;
season length=168 type=trig print=harmonics var=0 noest
keeph=1 2 keeph=7 to 70 by 7
keeph=6 to 54 by 7 keeph=8 to 58 by 7
keeph=5 to 47 by 7 keeph=9 to 51 by 7;
estimate plot=(panel) extradiffuse=168;
forecast lead=168 extradiffuse=168 plot=forecasts;
run;
```

The option extradiffuse=168 tells that the first week of 168 observations is excluded from the calculations of forecasts, prediction errors and forecasts as the first predictions have large variances, which disturbs the overall picture.

Figure 5: Forecast a week ahead using 41 harmonics. Note that March 22-23 are weekend days



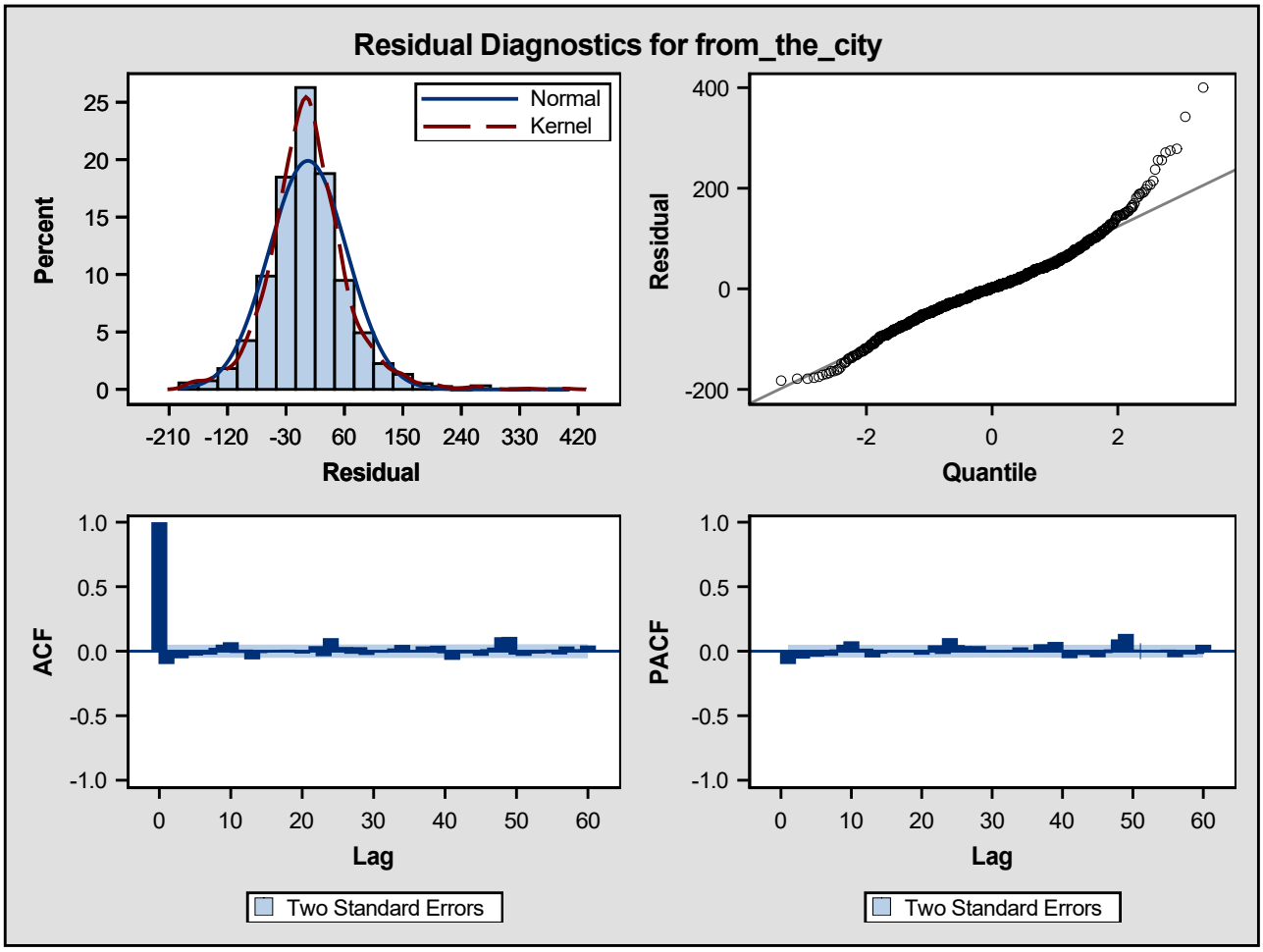
In the table of estimated parameters the variance of the irregular component is 3576 which of course is much larger than 3125 which we found if all 84 harmonics are included in the model. This tells that the 43 excluded harmonics have a statistically significant contribution to the model fit.

Final Estimates of the Free Parameters					
Component	**Parameter**	**Estimate**	**Approx Std Error**	**t Value**	**Approx Pr > |t|**
Irregular	**Error Variance**	3576.40560	119.58321	29.91	<.0001
Irregular	**AR_1**	0.83297	0.01325	62.86	<.0001

The forecast for the 168 hours of the following week is given in Figure 5. This graph is very close to the previous graph, Figure 4, so in practice the reduction of harmonics from 84 to 41 is unimportant. It is to believe that the number of harmonics could be reduced much further.

The model fit is acceptable as all autocorrelations are close to zero, even if some autocorrela-

Figure 6: Diagnostic plots



tions are significant because of the large number of observations, see Figure 6.

14.5 A blockseasonal component for seven days a week and 24 hours seasonal component for the hours a day

Another way of reducing the number of dummy variables from the 168 potential dummies is to exploit that the 24 hours a day are embedded in the seven days a week. This could be coded rather intuitively by a blockseasonal statement.

The following code includes hourly dummies for the 24 hours a day and daily dummies for seven days a week. These two sets of dummies are then combined in the final model. This gives a total of 31 dummies, which is 29 free parameters to be estimated as each set of dummies are restricted to sum to zero. All estimated components are saved in the new dataset, named predictions, by the option outfor=predictions in the forecast statement. The DATA step and the application of

PROC SGPLOT gives Figure 8 which will be discussed later.

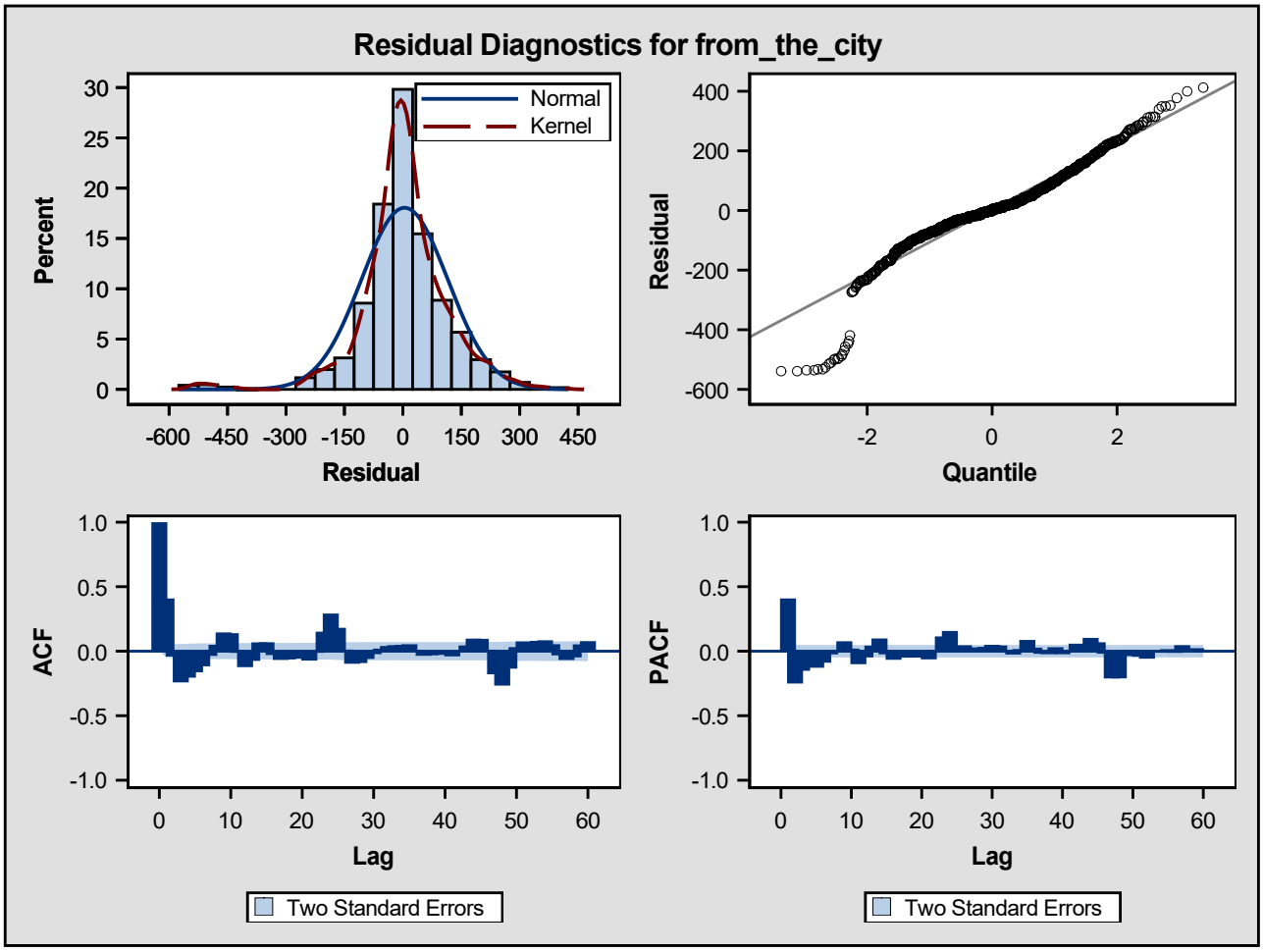
```
PROC UCM data=sasts.cycling;
id datetime interval=hour;
model from_the_city;
irregular p=1;
level  variance=0 noest;
season length=24 type=trig variance=0 noest plot=smooth;
blockseason nblocks=7 blocksize=24 variance=0 noest plot=smooth;
estimate plot=(panel) plot=residual;
forecast lead=168 plot=forecasts
outfor=prediction;
run;

data prediction_1;
set prediction;
if hour(timepart(datetime)) in (7,8)
then morning=from_the_city;
if hour(timepart(datetime)) in (0,22,23)
then night_life=from_the_city;
run;

proc sgplot data=prediction_1;
scatter x=datetime y=from_the_city/;
series x=datetime y=forecast;
scatter x=datetime y=morning/
  MARKERATTRS=(color=red symbol=CircleFilled);
scatter x=datetime y=night_life/
  MARKERATTRS=(color=green symbol=CircleFilled);
where datepart(datetime)>mdy(3,6,2014)
  and datepart(datetime)<mdy(3,11,2014);
run;
```

The model fit is poor as the variance of the irregular component is large, 11812, as seen in table of estimated parameters. Even if the autoregressive parameter is maintained in the model, this variance is much larger than the variance in the previous models.

Final Estimates of the Free Parameters					
Component	**Parameter**	**Estimate**	**Approx Std Error**	**t Value**	**Approx Pr > |t|**
Irregular	**Error Variance**	11812	389.22922	30.35	<.0001
Irregular	**AR_1**	0.85604	0.01379	62.09	<.0001

From a time series point of view the model fit is poor as the residuals are autocorrelated as seen

Figure 7: Diagnostic plots

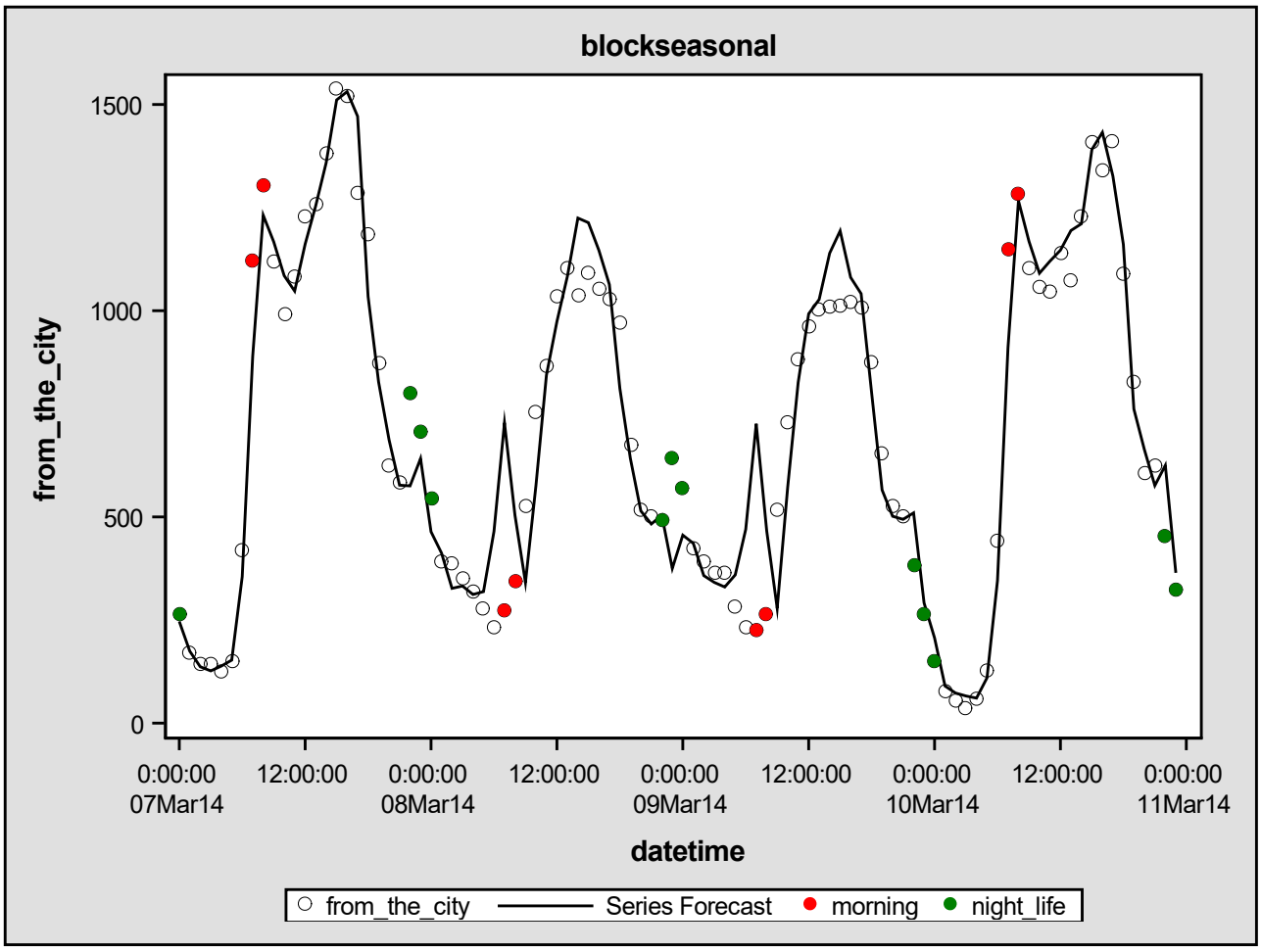


from the Figure 7 such that the first order autoregression is not sufficient and many more terms are needed. Moreover the histogram and the Q-Q plot tell that the distribution of the residuals is heavy left tailed as some residuals are less than -400 while the rest are larger than -275.

The reason for the poor fit is clearly seen from Figure 8 which is produced by the last PROC SGPLOT in the code above. Here forecasts and observations are plotted for four days, Friday March 7'th to Monday March 11'th, so the effect of a weekend is seen in detail. The red filled dots are the two morning hours 7 and 8 AM where mainly cyclists pass on ordinary work days while very few cyclists pass in these morning hours in the weekend. Moreover the green dots at 10 and 11 PM and also the first hour after midnight make clear that the traffic at night is different the nights after Friday and Saturday compared to the nights before working days.

The plot, Figure 8, spells out clearly that the seasonal structure of the 24 hours within a day is not constant for all seven days a week.

Figure 8: Forecasts and and model plot. Note that March 14. and 15. are weekend days



To conclude, the idea of overlaying two seasonals, one for the days a week and one for the variation within a day, is not working for the present data set. An application of 168 dummies or 84 harmonics is preferable as in Section 14.3. If one wants to reduce the number of estimated parameters the number of seasonals should be drastically reduced by deleting important harmonics.