Scientific Report of the STSM

Selini Hadjidimitriou

February 9, 2016

Estimation of missing temperatures

The objective of this STSM was to analyse natural gas flow data to develop forecast models of gas flow. First, for a one month dataset of the gas flow the missing temperatures have been estimated and similar exits and entries have been classified in terms of flow distribution.

For the estimation of missing temperatures, the nearest neighbour temperature has been assigned to the missing data based on time and day. Secondly, the analysis of correlation between the temperature and the gas flow has been computed for each entry/exit and time range. The aggregated plot shows that the majority of observations have negative correlation meaning that the higher is the temperature the lower is the gas flow (Figure 1). Figure 2 shows that the correlation also depends on the hourly time range. The classification of similar entry/exits was, therefore, based on the time range.

Clustering of similar nodes based on the flow distribution

The similarity between couple of distributions has been assessed using the Two-sample Kolmogorov-Smirnov. The dataset includes gas flows at hourly time range so that the clustering methodology needed to consider that variability over time. To tackle this issue, the pair comparison of distribution based on the Kolmogorov-Smirnov test, is performed for each hour.

The Kolmogorov-Smirnov test for distances between pair of distributions provides two possible results. If two distributions are similar, the result of the test is 0, otherwise 1. For each pair of nodes and for each hourly time range, the test is executed such that 24 matrices with the results of the Two samples Kolmogorov-Smirnov test are obtained. The 24 matrices are then summed such that the higher the value, the higher is the probability that the two exit points should be grouped together.

Finally, the hierarchical cluster analysis with the single linkage method is run to obtain groups of similar nodes in terms of distance between distributions

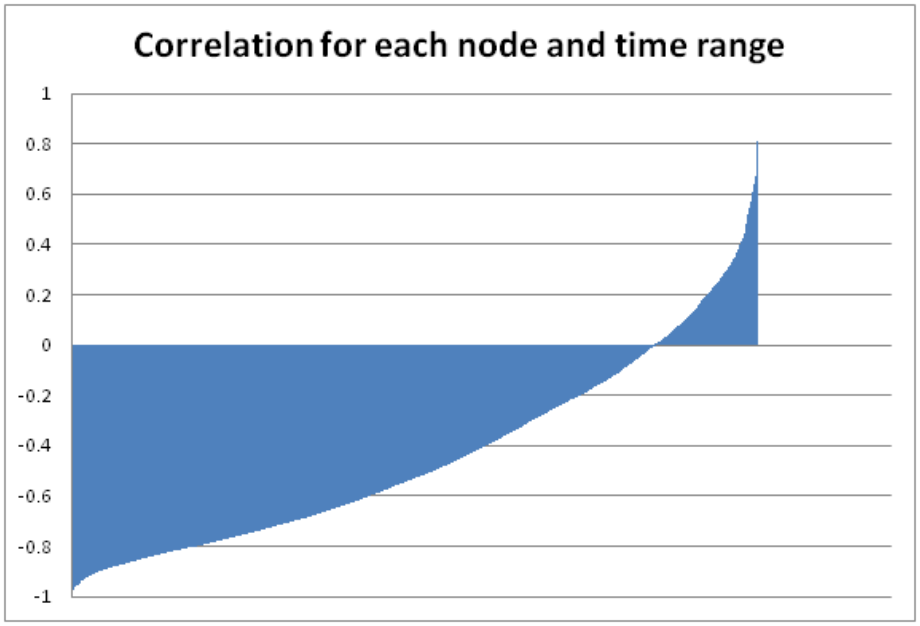

Figure 1: Correlation between temperature and gas flow by time and node.

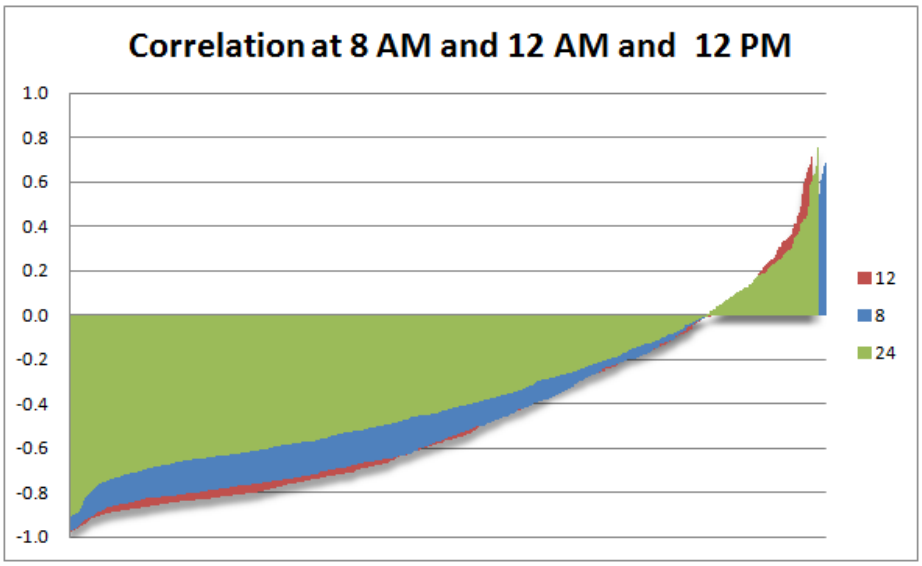

Figure 2: Correlation between temperature and gas flow by time and node (selected times).

and takes into account the hourly time range.

Results of the cluster analysis

The classification shows that the majority of clusters (73%) includes only one exit, while the 13,5% is formed by two elements. The kernel distributions of three samples of clusters are shown in Figures 3, 4 and 5.

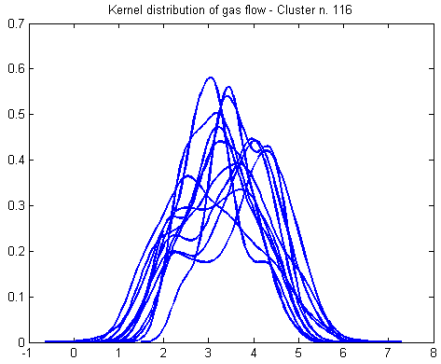

Figure 3: Kernel of flow cluster 116

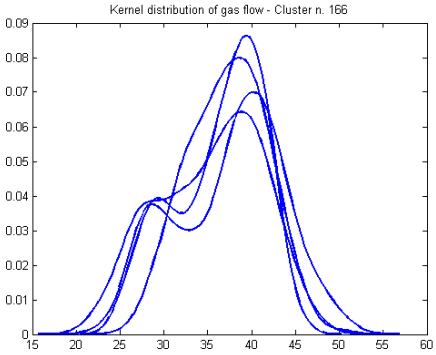

Figure 4: Kernel of flow cluster 166

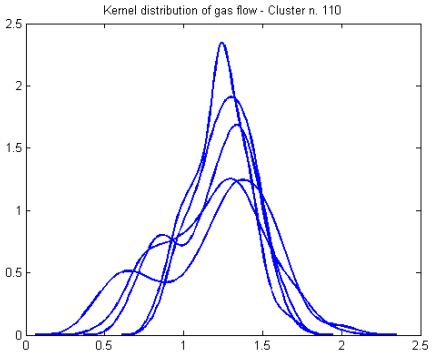

Figure 5: Kernel of flow cl. 110

Analysis of the new dataset

The second part of the analysis was focused on a new dataset consisting of an almost two years time horizon. The detailed analysis of the time series has allowed to identify some inconsistencies of the dataset dealing with:

- nodes that showed mixed behaviour (not clear if they were inflows or outflows nodes)

- missing data in correspondence of an entire month for all nodes clearly showing a measurement error
- in the last observed year, 32 nodes were eliminated by the data provider from the dataset accounting for the 16% of the total flow

First forecast results

Based on the descriptive statistics, the gas year which is considered for the analysis is the one included between the Oct 1st, 2013 and Sep 30, 2014. Furthermore, the entry node A635373 that account for the 9% of the total inflow is considered for the analysis.

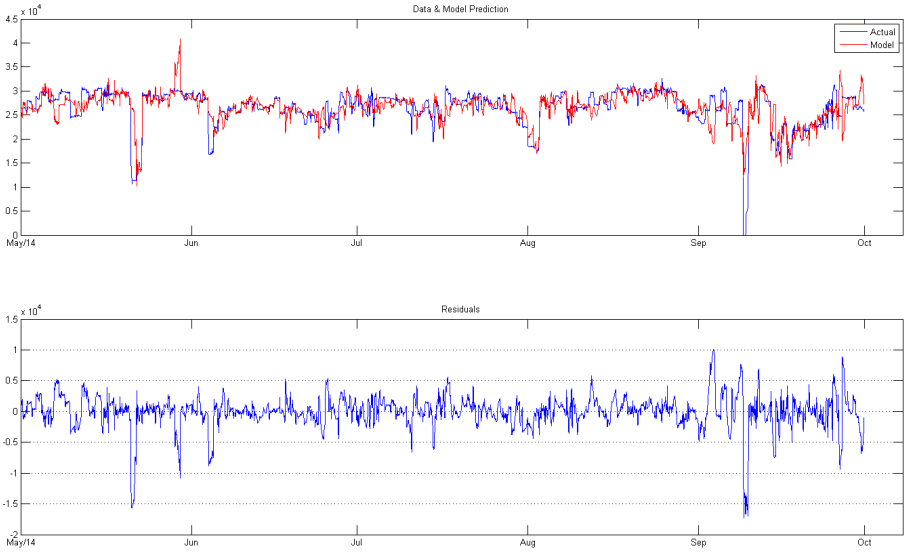

Figure 6: Forecast for node A635373

The forecast is obtained using the Neural Network model with 20 neurons. The holiday/weekend indicator, the value of flow in the previous day in the same hour, and the value of the flow in the previous week same day of the week and same hour are considered in the forecast model. Figure 6 shows a first result for entry node A635373. The training set is from Oct-2013 to Apr-2014 and the test set from May-2014 to Oct-2014. The model quite fit the actual data but there are some peaks (i.e. Jun-2013) which are clearly wrong.

The value of MAPE indicates a good level of fit between the model and the actual data.

- Mean Absolute Percent Error (MAPE): 8.60%

- Mean Absolute Error (MAE): 1687.10
- Daily Peak MAPE: 6.58%

The work initiated during this STSM is still ongoing. Current activities consists of improving the accuracy of the forecast model, comparing the results obtained using different forecast methodologies (i.e. autoregressive models) and performing the forecast on the aggregated inflows and outflows. For this reason, a new STSM is needed to complete the work and be able to meet with the working group which is based at ZIB.