Scientific report of the Short-Term Scientific Mission (STSM)

Andrea Taverna

May 16, 2016

1 Introduction

The objective of this STSM is to identify suitable feature extraction methods to support the analysis of gas flows time series in gas transmission networks. The main goal of the analysis is computing accurate forecasts for gas flow on the next 24 and 48 hours to support operational decisions. Further objectives include clustering and classification of nodes and time periods to better explain and synthesise the characteristics of nodes and the relationship among them.

The STSM brought the following results:

- identification of significant node clusters via cross-correlation
- identification of outliers as cause for high error rates for current forecasting methods
- proposal for an outlier detection method

2 Clustering of nodes via correlation

The goal of clustering for network nodes is to extract features from flow time series to succinctly represent relevant characteristics of the nodes and the differences between them. Similarity between nodes has been estimated via the correlation coefficient of gas flows. The analysis shows that a significant number of nodes (243 nodes out of 436) have high positive correlation ($\geq 80\%$) between them (tables 1, 2). Most of these nodes are exit nodes belonging to the DISTRIBUTION class, which is associated to distribution networks and public services. The high correlation is explained considering that gas demand at these nodes depends on human activities, which follow a similar pattern across the network for most of the year.

correlation c	0.8	0.825	0.85	0.875	0.9	0.925	0.95	0.975
# nodes couples with correlation above c	16703	13729	10887	8144	5449	2896	1307	265

Table 1: # node couples for positive correlation.

type	Distribution	Industrial	Cross-network
# of couples	14415	10	7

Table 2: # node couples with correlation above 80% and same class

Highly negative correlation (tab. 3) is less frequent and could also signal the presence of unreported "sibling nodes".

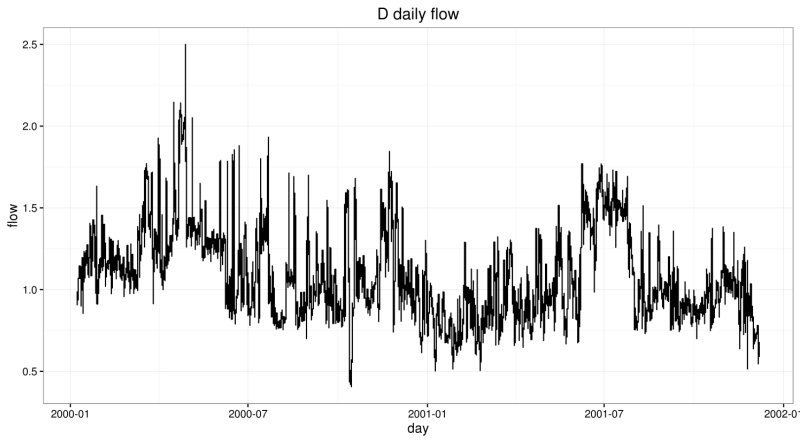
Correlation between time series seems to be unrelated to the distance between the nodes.

Nodes in classes other than DISTRIBUTION show low correlation between them. In general, these classes have fewer nodes with a more diverse behaviour. As a consequence, it is hard to extract common and meaningful features across all the nodes to obtain a useful clustering. This could explain the poor results of previous attempts at feature extraction using standard dimensionality reduction methods such as Principal Component Analysis (PCA) or specific time series dissimilarity measures such as Dynamic Time Warp (DTW) distance.

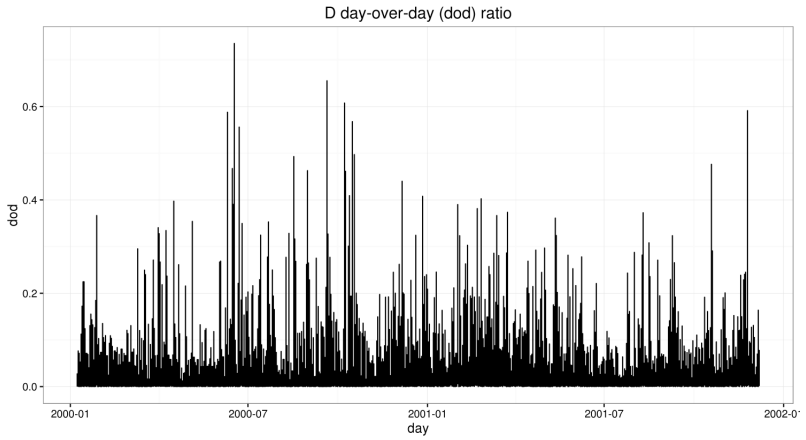
3 Explaining high-error rates for current forecasting methods

Among the forecasting methods being evaluated there are Functional Auto-Regression (FAR) and Support Vector Regression (SVR). These two methods have been applied on both the daily total flow and the hourly flow of some of the major entry nodes and result in an average error rate around 5% of the actual flow, which, due to the magnitude of the quantities involved, is considered to be excessively high.

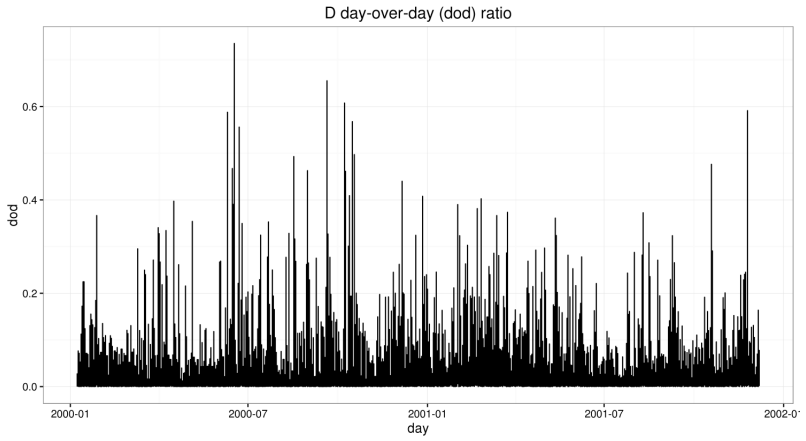
A comparison between forecast errors and real gas flow showed a correspondence between high errors and sudden level shifts in the flow. Outstanding level shifts suggest the presence of outliers due to data error, maintenance or other external factors, and should thus be filtered out from the forecasting procedure. In particular, maintenance periods have zero gas flow and can cause high error rates in the forecast. In figure 1 I report an example of forecast comparison.



(a)



(b)



(c)

Figure 1: Example of daily total flow, forecast errors for FAR method and day-over-day variations. APE is Absolute Percentage Error. The highest APE in chart (b) correspond to a sudden flow drop due to maintenance.

correlation c	-0.8	-0.825	-0.85	-0.875	-0.9	>-0.9
# nodes couples with correlation below c	189	139	95	48	2	0

Table 3: # node couples for negative correlation

4 Outlier detection

In order to detect outliers in a time series several methods from the literature have been considered, such as [1]. They are not practical for our case as they assume time series to be smoother and shorter than the ones being considered and, thus, tend to report more outliers than expected.

Let $X = \{x_t\}_{T=1..t_{\max}}$ be a time series with non-negative values. The following outlier detection method is proposed:

1. **Preprocessing:** Small flow values are likely to be outliers. They can be excluded when measuring the characteristics of the time series to obtain more accurate estimates.

2. Let $\epsilon \to 0^+$. Let $r = \mathrm{IQR}(X|X > \epsilon)$ be the interquantile range (IQR) of the time series, excluding small flow values. All the periods, including $t|x_t \leq \epsilon$, satisfying the following condition

$$x_t - x_{t-1} \geq \alpha r \qquad \alpha \geq 0$$

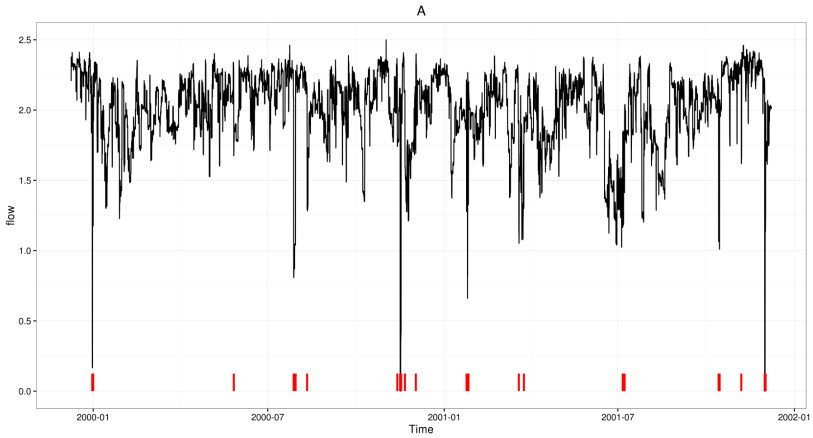
   are marked as outliers. I suggest $\alpha = 1.25$.

   Rationale: If the condition is true then the period-by-period variation is comparable to the IQR, which is the overall spread of the time series including 50% of the most likely values. Hence any lag-1 difference $(x_t - x_{t-1})$ above the threshold is likely to correspond to an outlier.

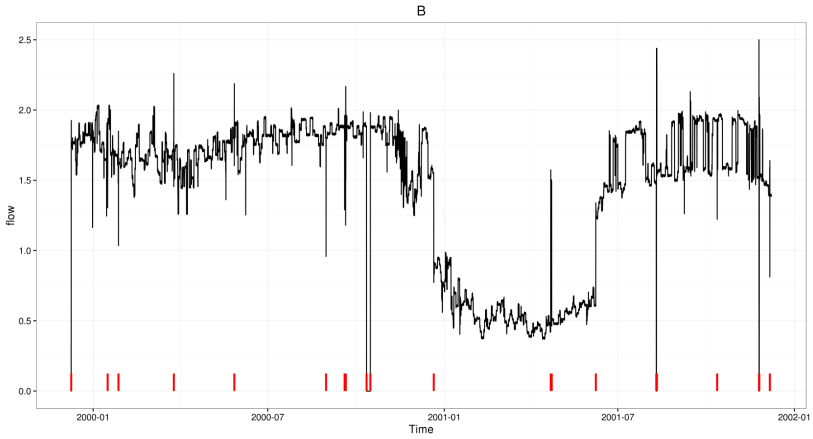
Examples (fig. 2) show the method can robustly identify outliers, especially the ones likely due to maintenance, for different nodes with different gas flow trends.

References

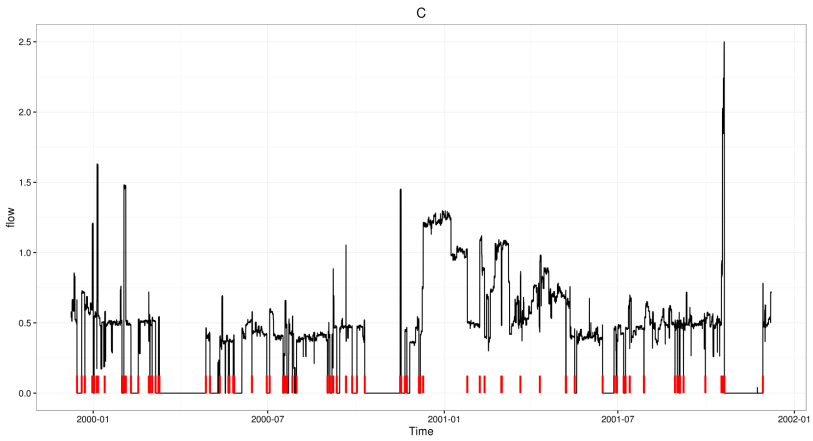
[1] Chen C. & Liu L. *"Joint Estimation of Model Parameters and Outlier Effects in Time Series"* – (1993)



(a)



(b)



(c)

Figure 2: Outlier detection for three entries (hourly flow) with $\alpha = 1.25$ and $\epsilon = 10^{-3} \max(X)$. Outliers are marked with red segments on the time axis.