Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

NAZGUL ZAKIYEVA[1], MILENA PETKOVIC[2]

Modeling and forecasting gas network flows with multivariate time series and mathematical programming approach

[1] 0000-0001-9106-9916
[2] 0000-0003-1632-4846

ZIB Report 21-23 (July 2021)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Modeling and forecasting gas network flows with multivariate time series and mathematical programming approach

Nazgul Zakiyeva[1] and Milena Petkovic[1]

Zuse Institute Berlin
zakiyeva@zib.de petkovic@zib.de

Abstract. With annual consumption of approx. 95 billion cubic meters and similar amounts of gas just transshipped through Germany to other EU states, Germany's gas transport system plays a vital role in European energy supply. The complex, more than 40,000 km long high-pressure transmission network is controlled by several transmission system operators (TSOs) whose main task is to provide security of supply in a cost-efficient way. Given the slow speed of gas flows through the gas transmission network pipelines, it has been an essential task for the gas network operators to enhance the forecast tools to build an accurate and effective gas flow prediction model for the whole network. By incorporating the recent progress in mathematical programming and time series modeling, we aim to model natural gas network and predict gas in- and out-flows at multiple supply and demand nodes for different forecasting horizons. Our model is able to describe the dynamics in the network by detecting the key nodes, which may help to build an optimal management strategy for transmission system operators.

Keywords: forecasting, mathematical programming, natural gas

1 Introduction

Germany is the largest market for natural gas in the European Union. Natural gas is the second most used energy source in Germany, with 25% share in primary energy consumption [1] and plays an important role in the German "Energiewende", especially in the transition from fossil fuels to renewables in the power sector. Recently, the natural gas market is becoming more and more competitive and is moving towards more short-term planning, e.g., day-ahead contracts, which makes the dispatch of natural gas even more challenging [4]. Despite these challenges, TSOs have to fulfill all transport demands in a cost effective way. Since the average gas pipe velocity is only 25km/h, a high-accuracy and high-frequency forecasting of supplies and demands in natural gas network is essential for efficient and safe control and operation of the complex natural gas transmission networks and distribution systems.
This work is part of a joint project with one of Germany's biggest transmission

system operators, Open Grid Europe [2]. In order to provide a comprehensive understanding of the network dynamic, together with OGE, we develop a Network Autoregressive Linear model with balance constraint (NAR-LB) model that detects influential nodes in the network (the nodes that demonstrate a strong effect on the future flows of other nodes and drive the movement of the network over time) and provides high-precision, multi-step ahead hourly forecasts for more than 200 nodes in the gas network. These nodes have very different statistical characteristics, as they represent a variety of different source and consumption points, ranging from connections to other gas networks or countries over industrial consumers to storage facilities. Integrating recent advances in optimization and statistical modeling models, NAR-LB models the hourly gas flows of all the nodes jointly to incorporate the temporal cross dependence in the network. As the size of the network grows, the number of parameters is controlled by regularizing the only significant dependencies, which highlights the influential nodes in the network.

2 Methodology

Let N denote the number of nodes in a large-scale complex gas transmission network. We denote with $q_{t;i}$ the gas flow time series at node i, where $t = 1;:::T+1$ and $i = 1;:::;N$. Folowing the model proposed in Chen et al. [3] we define Network Autoregressive Linear model with Balance constraint, to capture the network effect of different nodes and determine influential nodes for each node in the network, where the total gas in-flows and out-flows need to be balanced, as follows:

$$\begin{array}{ll} q_{t;i} = \sum_{j=1}^{N} q_{t-1;j}\, w_{j;i}; & i = 1;:::;N; t = 2;:::;T+1 \\ \text{s.t.} \sum_{i=1}^{N}\sum_{j=1}^{N} q_{t-1;j}\, w_{j;i} = 0; & t = 2;:::;T+1 \end{array} \tag{1}$$

The parameter $w_{i;i}$ model the autoregressive dependence for each node while parameter $w_{j;i}$ when $i \neq j$ model the influence of j-th node to the i-th node i.e. influence of the past value of the j-th node on the current value of the i-th node. Let us define $T \times N$ matrix $Q = (q_1;:::;q_N)$ containing T previous flow values of N network nodes $q_i = (q_{1;i};:::;q_{T;i})^\top$, matrix $X = (x_1;:::;x_N)$ containing previous values of Q and matrix W representing $N \times N$ matrix of unknown parameters $w_{i;j}$ that model mutually influence of the network nodes.

While autoregressive dependence is modeled by diagonal elements of W, the non-diagonal elements define the weighted adjacency matrix where row vector $W_j = (w_{j;1};:::;w_{j;N})$ represents the influence of the j-th node on the future values of the other nodes in the network. We assume that the weighted adjacency matrix is sparse but we have no prior knowledge of the sparse structure in terms of number of significant elements and their location. To detect the influential nodes in the dynamic network, we adopt the Lasso type regularization in estimation. Since in any point in time only a small number of nodes have a significant effect to the network dynamic and each of influential nodes only have

an influence on the future flow of small number of nodes, we impose 2 layers of sparsity: groupwise and inner group sparsity on the weighted adjacency matrix. We observe matrix W in N groups (W_j) and estimate the parameters by applying Sparse-Group Lasso penalty [3].

$$\begin{aligned} \min_{W} & jEj + \quad \sum_{j \quad g} \quad W_j \quad _2 + \quad \sum_{i \in j} \quad w_{i;j} \\ & \text{s.t. } E = Q \quad XW \\ & \quad XWI = 0 \\ lb \quad & w_{i;j} \quad ub; \quad i = 1;:::;N; j = 1;:::;N \end{aligned} \tag{2}$$

where I is is $N \quad N$ identity matrix, $0 \quad p\overline{N} \quad 1$ is a e Lasso parameter for the individual weight $w_{i;j}$ when $i \neq j$ and $_g = \sqrt{N}$ is a group Lasso parameter for W_j. In addition we set the lower and upper bounds for the weights to $lb = \quad 2$ and $ub = 2$, respectively.

We use the estimated weighted adjacency matrix to choose the most influential lagged flows for each node in the network as features for multi-step ahead forecast. For each node $i \; 2 \; 1;::;N$ we select F features with the highest absolute value of $w_{j;i}; j = 1;::::;N:$ Further, we approximate future gas flow with weighted sum of features:

$$\hat{q}_{t;i} = \sum_{j=1}^{F} q_{t \quad 1;j} \quad f_{j;i}; \quad i = 1;::N \tag{3}$$

where $\hat{q}_{t;i}$ is the approximated gas flow for node i at the time t, $q_{t \quad 1;j}; j = 1;::::;F$ are previous flows of F nodes with the highest influence to node i and $f_{j;i}$ are corresponding weights. If we define the error of approximation as:

$$e_{t;i} = q_{t;i} \quad \sum_{j=1}^{F} q_{t \quad 1;j} \quad f_{j;i}; \quad i = 1;:::;N \tag{4}$$

then the optimal weights are calculated by minimizing the sum of absolute errors:

$$\begin{aligned} & \min_{f} \quad \sum_{i=1}^{N} je_{t;i}j \\ \text{s.t. } & \sum_{j=1}^{F} q_{t;i} \quad \sum_{j=1}^{F} q_{t \quad 1;j} \quad f_{j;i} = e_{t;i}; \quad i = 1;:::;N; t = 2;:::;T+1 \\ & \sum_{i=1}^{N} \quad \sum_{j=1}^{F} q_{t \quad 1;j} \quad f_{j;i} = 0; \quad i = 1;:::;N; t = 2;:::;T+1 \\ & \quad lb \quad f_{i;j} \quad ub; i = 1;:::;N; j = 1;:::;F \end{aligned} \tag{5}$$

To optimize the forecasting performance we choose the tuning parameters using the rolling window technique with window size of 120h over the training set T_{train}. We minimize average mean square forecast error (MSE), $MSE = \sum_{t2T_{train}} \sum_{h=0}^{H} \frac{1}{(q_{t+h} \quad \hat{q}_{t+h})^2} = (H \quad jT_{train}j);$ for each node by performing the grid search for $\quad 2 \; f0;:::;1g$, where $q_{t;h}$ and $\hat{q}_{t;h}$ are the real and forecasted values of the natural gas flows at hour t and H is a forecasting horizon.

3 Experimental setup and results

In this paper we investigate the dynamic patterns of natural gas flows in the high-pressure gas pipeline network of OGE [2]. The dataset consists of demand and supply flows with an hourly time resolution for a period of 21 months (625 days). The network contains 1029 nodes in total. We consider 210 nodes with mean daily flow of more than 25MW and less than 95% of zero flows. In the observed data set we have 14 supply nodes (labeled S1-S14), 9 storages (can change a behavior and have both positive and negative flows over time, labeled ST1-ST9) and 187 demand nodes (labeled D1-D187). Furthermore, we add an artificial node to the network to represent the contribution of the nodes that are less important in terms of volume and active time.

All gas flow data are normalized with mean zero and unit variance. Figure 1 illustrates the temporal dependence among the observed 210 nodes. As it can be seen in the diagonal, there is a strong positive autocorrelation of each node with its own past values. Off the diagonal, the cross-correlations represent the dynamic flow of gas from one node to another. While all of the supply nodes and most of the demand nodes contain their own predictive information, the dynamic dependence in the network is sparse and is driven by small number of nodes. Among the supply nodes, only the 4 seem to be active, demonstrating a strong positive cross-correlation within the supply group and strong negative cross-correlation to some demand nodes and very limited dependence with the storage nodes. This is also the case for the demand nodes. The temporal dependence suggests that only a small number of nodes may have a significant effect on the future gas flows in the network.

We use training-validation sample of 260 days (T_{train}) to choose optimal hyperparameters for the sparsity estimation (see Section 2). Further, we use the chosen parameters to estimate the large-scale weighted adjacency matrix W at each point in the testing period T_{test} consisting of 365 days (T_{test}). With a rolling window size of 120h, we move forward one period at a time to update adjacency matrix and calculate the forecast for three different forecasting horizons (1h, 6h and 12h ahead), until we reach the end of the sample. We compare the performance of NAR-LB model with several well-known benchmarks: Baseline forecast (repeating last known value for the same hour in the day), ARIMA and LSTM. We determine the best ARIMA models for an univariate time series of 210 nodes according to a Akaike information criterion (AIC) using 28 days of rolling window. The LSTM is implemented using one LSTM hidden layer with an optimized number of hidden neurons for each node (32-128) and learning rate for each individual node (0.0001-0.001), a dropout of 0.1 followed by a fully connected output layer with the number of neurons equal to the forecast horizon and trained for 100 epochs. The parameters are optimized based on the performance on the training set (T_{train}).

The performance of NAR-LB model is measured and quantified by calculating the forecast accuracy for individual nodes, as well as the mean for the entire network. We use root mean squared error (RMSE), as well as normalized mean absolute percentage error (nMAPE) defined as: $RMSE = [\sum_{t2T_{test}} \sum_{h=0}^{H} \frac{1}{}(q_{t+h}$

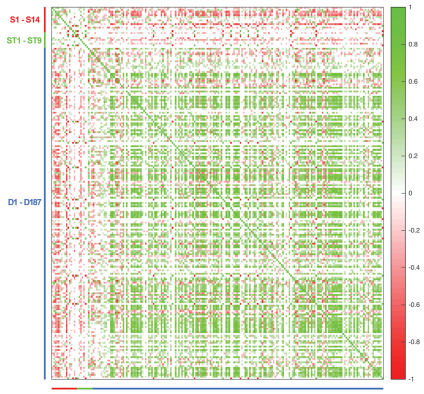

Fig. 1: Sample cross-correlation heatmap for 210 nodes in gas network.

$\hat{q}_{t+h})^2 = (H \; jT_{test}j)]^{1=2}$ and $nMAPE = (\sum_{t2T_{test}} \sum_{h=0}^{H} \frac{1}{j(q_{t+h} \; \hat{q}_{t+h})=\max(q)j)=(H \; jT_{test}j)}$; where $q_{t;h}$ and $\hat{q}_{t;h}$ are the real and forecasted values of the natural gas flows at hour t and H is a forecasting horizon.

	RMSE				nMAPE			
H	NAR-LB	BAS	ARIMA	LSTM	NAR-LB	BAS	ARIMA	LSTM
1	0.249	0.534	0.408	0.387	0.095	0.261	0.150	0.177
6	0.443	0.534	0.503	0.462	0.185	0.261	0.281	0.205
12	0.513	0.534	0.559	0.559	0.220	0.261	0.299	0.299

Table 1: Performance comparison

In Table 1 we report an average $RMSE$ and $nMAPE$ for three considered forecasting horizons and comparing to three alternative benchmark models. The results show that NAR-LB consistently outperforms all benchmark models. The difference is most significant for the shorter horizons where mean $nMAPE$ is improved for 37 % comparing to second best alternative model, while for the longer horizon (12h) NAR-LB perform similar to LSTM model with the improvement of 7.6 %. It is clear that the proposed model benefits from modeling temporal dependencies in the network. Figure 2 illustrates the estimated weighted adjacency matrix $\hat{W}$ in different periods of the year. It can be seen that during the winter time there is much more dynamic in the network while during the summer time, the number of influential nodes is significantly smaller.

4 Conclusion

In this paper we propose the Network Autoregression Linear model with Balance constraint for identifying the influential nodes in the large-scale complex