Modeling flow in gas transmission networks using shape-constrained expectile regression

Fabian Sobotka[1], Radoslava Mirkov[2], Benjamin Hofner[3], Paul H. C. Eilers[4], Thomas Kneib[5]

[1] Department of Mathematics, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany, fabian.sobotka@uni-oldenburg.de
[2] Department of Mathematics, Humboldt-University Berlin, 12489 Berlin, Germany, mirkov@math.hu-berlin.de
[3] Institute for Medical Informatics, Biometrics and Epidemiology, Friedrich Alexander University Erlangen-Nürnberg, 91054 Erlangen, Germany, benjamin.hofner@imbe.med.uni-erlangen.de
[4] Erasmus Medical Center, University Rotterdam, 3015 GE Rotterdam, Netherlands, p.eilers@erasmusmc.nl
[5] Department of Economics, Georg August University Göttingen, 37073 Göttingen, Germany, tkneib@uni-goettingen.de

Abstract: The flow of natural gas within a gas transmission network is studied with the aim to optimise such networks. The analysis of real data provides a deeper insight into the behavior of gas in- and outflow. A geoadditive model for describing the dependence between the maximum daily gas flow and the temperature on network exits is proposed. Semiparametric expectile regression provides the possibility to model the upper tail of the response distribution while accounting for the spatial correlation between different exits. The effect of the temperature is modeled with shape constraints to include knowledge about gas load profiles and to allow for a realistic prediction. Estimates based on least asymmetrically weighted squares (LAWS) and boosting are presented. The forecast of gas loads for very low temperatures is included and the application of the obtained results is discussed.

Keywords: Expectiles; P-splines; Semiparametric Regression; Gas Flow; Boosting.

1 Introduction

Expectiles are a great way to study trends *and* variation, skewness, etc. of observed data. In this paper we will illustrate that by modeling the flow of gas measured at the exits of a gas transmission network. We continue the work of Friedl et al (2011) who used semiparametric mean regression to analyse gas consumption subject to temperature changes for single exits of the gas network. However, the upper tail of the conditional distribution of the gas flow is of more importance. It helps to generate extreme

scenarios of gas consumption that are necessary to examine the capacities of the gas network. Expectile regression offers a flexible and easy way to model the tails of a distribution as it avoids a full specification of the error term distribution. The estimation of expectiles is based on minimising an asymmetrically weighted sum of squared residuals. A semiparametric least squares regression allows for a very flexbile regression structure by including nonlinear effects of continuous covariates, random effects or spatial effects. These extensions often rely on penalised least squares or penalised likelihood estimation with quadratic penalties and are therefore natural partners for least squares estimation.
Our geoadditive model comprises various parametric effects, a nonlinear shape-constrained function of the temperature and a continuous spatial effect based on the longitude and latitude of the exits.

2 Data Description

Data for this study were obtained from measuring stations within the German pipeline network operated by Open Grid Europe GmbH (OGE), one of the leading German gas transmission operators. It contains hourly gas flow for 238 network exits for the period between June 2009 and May 2010. Mean daily temperatures from the corresponding weather stations are also provided. Additionally, we distinguish several exit types. Typical exits in such networks are public utilities, industrial and areal consumers, as well as exits on border and market crossings. Continuous geographic coordinates, i.e., longitude and latitude for every node are also included.
We study the dependence of gas loads on air temperature, exit type and the geographic location of exits within the network, simultaneously on all exits along the pipelines of the gas transmission network. Since we want to maximise the transportation capacity through the pipelines, we concentrate on the daily maximum flows $y_{i,k}^{max}$, $i = 1, \dots, n$ ($n = 365$), at the exits $k = 1, \dots, 238$ in the network. We note here that in this study we observe the so-called H-network, which denotes the network with high Wobbe Index (a measure for the heating value of the gas).
In what follows, we study the daily maximum flows standardised separately for each exit

$$y_{i,k} = \frac{y_{i,k}^{max} - \bar{y}_k}{\hat{\sigma}(y_k)}$$

in order to obtain comparable response values. As we are interested in the upper tail of the conditional distribution of the response, a mean regression is not sufficient. However, a quantile or expectile regression can be a sensible estimate for the quantity of interest.

3 Expectile Regression

The results of an expectile regression can be acquired by computing the least asymmetrically weighted sum of the squared residuals (LAWS) analogue to a quantile regression that minimises the asymmetrically weighted absolute values of the residuals. LAWS minimises

$$S = \sum_{i=1}^{n} w_\tau(y_i)(y_i - \mu_i(\tau))^2$$

with weights

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \mu_i(\tau) \\ 1-\tau & \text{if } y_i < \mu_i(\tau) \end{cases}$$

where y_i is a continuous response and $\mu_i(\tau)$ is the estimated expectile for different values of the asymmetry parameter $\tau \in (0,1)$. Hence the computation of expectile regression is very easy, since it avoids the non-differentiable absolute value criterion that is used to estimate quantiles. In further comparison, expectiles lack the intuitive interpretation of quantiles. While the quantile of a random variable Z immediately depicts the amount of probability that lies below it, the τ-expectile $\mu(\tau)$ can only be defined implicitly:

$$\tau = \frac{\int_{-\infty}^{\mu(\tau)} |z-\mu(\tau)| f(z) dz}{\int_{-\infty}^{\infty} |z-\mu(\tau)| f(z) dz} = \frac{G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))}{2(G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))) + (\mu(\tau) - \mu(0.5))}$$

where $G(m) = \int\limits_{-\infty}^{m} zf(z)dz$ and $G(\infty) = \mu(0.5)$ is the expectation of Z. In addition to the computational advantages of expectiles, one can build additive models that contain different kinds of effects. We portray these effects by design matrices $B^{(j)}$ and assign a vector of regression coefficients β_j to each effect. We can then create the following additive expectile regression model:

$$\mu(\tau) = 1\beta_0 + X\beta_1 + B^{(2)}\beta_2 + \ldots + B^{(r)}\beta_r + \varepsilon_\tau.$$

For continuous univariate covariates, smooth expectile curves can be fitted using penalised splines (see Schnabel and Eilers, 2009). Additionally the model can include spatial effects based on either Markov random fields, tensor product splines or Kriging (see Sobotka and Kneib, 2010). The smoothing can be induced by a quadratic penalty on the regression coefficients:

$$pen(\beta_{j,\tau}) = \lambda_j \beta'_{j,\tau} K_j \beta_{j,\tau}$$

with adaptable smoothing parameter λ and penalty matrix K.
As in the case of modeling gas flow we are interested mainly in extreme expectiles describing the upper tail, we need to clarify the possible interpretations of a single expectile curve. An expectile can be related to the risk measure expected shortfall (ES) as for a random variable Z holds

$$
\begin{aligned}
ES_p(t) &= E(Z(t)|Z(t) > \tilde{z}_p(t)) \\
&= \left(1 + \frac{\tau}{(1-2\tau)p}\right)\mu_\tau(t) - \frac{\tau}{(1-2\tau)p}\mu_{0.5}(t)
\end{aligned}
$$

with the τ-expectile μ_τ and $p = F_t(\mu_\tau(t))$ (see Taylor, 2008). As shown the expected shortfall is the conditional mean above the p-quantile and provides more reliable information about extreme observations than the quantile. This allows us to explicitly compute a risk bound for the maximum daily gas flow.

4 Estimating and Forecasting Gas Flow

The model explaining the standardised maximum gas flows includes parametric effects X indicating a weekend day and the types of the exits. Further a P-spline basis is used to model the effect of the local daily mean temperature $\mathtt{t}$. Finally, in a spatial effect the longitude $\mathtt{u}$ and latitude $\mathtt{v}$ of each exit are included by a Kriging basis. The knots for this basis are chosen as a subset $k_1, \ldots, k_K$ from the covariate observations $(u, v)_1, \ldots, (u, v)_{238}$. The basis evaluation is defined by Matérn correlation functions like

$$B_k(r, \phi) = \exp(-|r/\phi|)(1 + |r/\phi|)$$

with the Euclidean distance $r = ||k - x||$ and a fixed $\phi \propto \max_{i,j}(||k_i - k_j||)$. The penalty matrix $K = (B_{k_i}(||k_i - k_j||))_{i,j}$ then comprises the evaluated distances between the knots.

The analysed model has the form

$$\mu_\tau(y) = 1\beta_{0,\tau} + X\beta_{1,\tau} + B(\mathtt{t})\beta_{2,\tau} + B(\mathtt{u}, \mathtt{v})\beta_{3,\tau} + \varepsilon_\tau.$$

Due to the physical properties of the gas and the observed behaviour of industries and private households regarding gas consumption, certain restraints can be made to the effect of the temperature on the gas flow. First of all, the demand for gas will generally decrease with higher temperatures. However, even on very warm days there might be a minimum consumption of gas. Even for very low temperatures the loads of gas will not exceed a certain capacity. Hence, we include an additional iteration in the estimation process, restricting the regression coefficients of the P-spline basis to the realistic behaviour. This is achieved similar as in Bollaerts (2006) by an additional iteration within the estimating. First, the shape-constraints are defined as difference penalties on the spline coefficients. After an estimation step those spline basis elements are identified that do not follow the restrictions. Here, the penalty is invoked and the estimation is repeated. Preliminary results are shown in Figure 1. In the results one can easily see the gain of information through expectile regression. The variance and skewness clearly changes with the temperature. Hence, the information about extreme observations is more accurate than with mean regression. Also the spatial effect is non-informative for the mean while a difference between the north-eastern exits and the south-western exits is clearly shown in the upper tail through the 0.99-expectile.

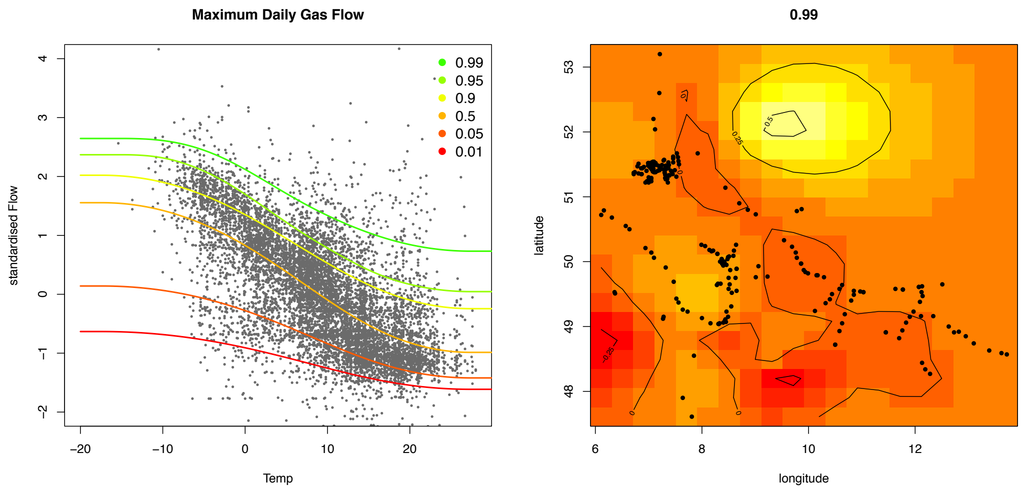

FIGURE 1. Left: Standardised flow for temperature interval. Shape-constrained effects for 6 asymmetries τ. Right: Spatial effect of the 0.99-expectile for 238 exits marked as black points.

The analyses are performed using the R–package “expectreg” (Sobotka, Schnabel, Schulze Waltrup, 2012).

Acknowledgments: The authors are grateful to Prof. Werner Römisch for the possibility to work on this interesting project. Thanks to OGE for supporting the project and providing real data as well as the relevant questions we investigate in this paper. Financial support from the German Research Foundation (DFG) grant KN 922/4-1 is gratefully acknowledged.

References

Bollaerts, K., P. H. C. Eilers and I. van Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology* , **59** , 451–469.

Friedl, H., R. Mirkov, and A. Steinkamp (2011). Modeling and Forecasting Gas Flow on Exits of Gas Transmission Networks. *To appear in: International Statistical Review Special Issue on Energy Statistics*.

Hofner, B. and J. Müller and T. Hothorn (2011). Monotonicity-Constrained Species Distribution Models. *Ecology*, **92**, 1895-1901.

Schnabel, S. and P. Eilers (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**, 4168-4177.

Sobotka, F. and T. Kneib (2010). Geoadditive expectile regression. *Computational Statistics and Data Analysis*, doi: 10.1016/j.csda.2010.11.015.

Sobotka, F., S. Schnabel and L. Schulze Waltrup (2012). *expectreg: Expectile and Quantile Regression*. R package version 0.30.

Taylor, J. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* **6**(2), 231–252.