Modeling and Forecasting Gas Flow on Exits of Gas Transmission Networks

Herwig Friedl[1], Radoslava Mirkov[2] and Ansgar Steinkamp[3]

[1]*Institute of Statistics, Graz University of Technology, Austria*
[2]*Department of Mathematics, Humboldt University Berlin, Germany*
[3]*Open Grid Europe GmbH, Essen, Germany*

E-mails: hfriedl@tugraz.at, mirkov@math.hu-berlin.de,

ansgar.steinkamp@open-grid-europe.com

Summary

The flow of natural gas within a gas transmission network is studied with the aim to optimize such networks. The analysis of real data provides a deeper insight into the behavior of gas in- and outflow. Several models for describing dependence between the maximal daily gas flow and the temperature on network exits are proposed. A modified sigmoidal regression is chosen from the class of parametric models. As an alternative, a semi-parametric regression model based on penalized splines is considered. The comparison of models and the forecast of gas loads for very low temperatures based on both approaches is included. The application of the obtained results is discussed.

Key words: sigmoidal regression; penalized splines; gas flow; gas transmission networks; prediction; optimization.

1 Introduction

Transportation and supply of natural gas is an important topic. We study historical data of the flow of gas transported in networks in order to support the optimization of such networks and thus improve the supply of gas. Statistical modeling techniques enable a reliable and realistic prediction of the future gas flow, and thus the reduction of operational costs. In particular, the cost of control energy necessary to satisfy peak demand at low temperatures can be minimized with a good forecast. Furthermore, gas transportation operators are obliged to sustain the supply of gas even during very cold days. Since there is not much data available for very low temperatures, a good prediction is crucial for reliable operation.

We fit a parametric as well as a semi-parametric non-linear logistic regression model and analyze the properties of the gas flow through the pipelines in dependence of the temperature.

The relationship between gas flow and air temperature is closely related to empirical models for growth data used for modeling trends in data, which are frequently employed in natural and environmental sciences, and sometimes in social sciences and economics. Some examples of those models and different applications are explored in, for example, Jones et al. [2009], Vitezica et al. [2010], Jarrow et al. [2004].

Hellwig [2003], Geiger and Hellwig [2002] and Wagner and Geiger [2005] suggest the use of sigmoidal growth models for description of typical gas load profiles in various economic sectors. An overview of methods useful for understanding the behavior of gas transport based on those models can be found in Cerbe [2008].

Theoretically, an empirical growth curve is a scatter plot of some measure of the size of an object against time x. The general assumption is that, apart from random fluctuation, the underlying growth follows a smooth curve. This theoretical growth curve is usually assumed to belong to a known parametric family of curves $f(x|\theta)$ and the aim is to estimate the parameters θ. It is important to reduce the complexity of the growth curve to a small number of parameters so that changing patterns of growth can be understood in terms of changes in these parameters. Moreover, each growth curve may be summarized by its parameter estimates as a single low-dimensional multivariate observation, which then may be subject to an analysis of variance or to a regression or even a correlation analysis.

The same type of models occur when the explanatory variable x is not time but increasing intensity of some other factor. We observe change, or more precisely, reduction of gas consumption with increased outdoor temperatures, and seek a model with a physical basis and physically interpretable and meaningful parameters. Detailed description of growth models can be found in Seber and Wild [2003].

As a more flexible alternative, semi-parametric models can be utilized to tackle the problem. We choose penalized splines (P-splines), which combine two ideas from curve fitting: a regression based on a basis of B-splines and a penalty on the regression coefficients, cf. Wegman and Wright [1983], Eilers and Marx [2010] and Eilers and Marx [1996]. This approach emphasizes modeling of underlying smooth regression relationship, and the penalty controls the amount of smoothing.

The paper is organized as follows. Section 2 describes the available data and motivates the choice of the studied models. The parametric approach is presented in Section 3, whereas Section 4 provides details about the application of the P-splines method. All models are then compared based on several criteria in Section 5. Section 6 explains several possible applications of the results, and Section 7 concludes the paper.

2 Data Description and Model Motivation

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 the period between January 2004 and June 2009. Mean daily temperatures from the corresponding weather stations are also provided.

We study the dependence of gas loads and air temperature on all exits along the pipelines. Typical exits in such networks are public utilities, industrial consumers and storages, as well as exits on border and regional crossings. Since we want to maximize the transportation capacity through the pipelines, we concentrate on the daily maximum flows y_i^{max}, $i = 1, \ldots, n$ ($n = 2005$), at each exit, for every exit in the network.

An important aspect of the study is the forecast of gas loads on exits of the network at the so-called design temperature. The design temperature is defined as the lowest temperature at which the gas operator is still obliged to supply gas without failure, and differs within Germany depending on the climate conditions in different regions. It usually lies between $-12°\mathrm{C}$ and $-16°\mathrm{C}$. Such low mean daily temperatures are very uncommon in Germany, and there is no observed gas flow data available at the design temperature. For this reason gas operators are forced to use predicted gas loads at the design temperature, and we investigate several possible models for the forecast.

Based on Cerbe [2008], Hellwig [2003], Geiger and Hellwig [2002] and Wagner and Geiger [2005], as well as on the German Energy Law, gas operators in Germany agreed to use a non-linear regression of sigmoidal type for the prediction of gas flow at the design temperature (cf. Agreement [2008]). The document describes the model given by the sigmoidal mean function with a weighted four-day mean temperature as an explanatory variable (see model equations (1) and (2) in Section 3), which should be used to predict maximum gas loads at the design temperature. The physical properties of gas imply the choice of the model. The weighted four-day mean temperature is motivated by the fact that typical buildings in Germany accumulate the heat up to 85 hours, and Cerbe [2008] suggests the weights as in (3). Agreement [2008] also contains starting values necessary for the non-linear regression. We remark that one of the parameters of the sigmoidal model as given in Agreement [2008] is the horizontal upper asymptote of gas loads, which occur at low temperatures. In this work we suggest the generalization of the upper asymptote, as we expect increased and not constant gas consumption as temperatures get lower and fall down to the design temperature, as well as a flexible weighting of the temperature of the last four days. Alternatively, we propose a local smoothing approach and utilize the P-splines models for the prediction.

In what follows, we study the standardized daily maximum flows $y_i = y_i^{max}/\bar{y}$, where $\bar{y}$ denotes the empirical mean of all maximal daily gas flows at a specific measuring station.

3 Parametric Models

In the parametric approach, the assumption is that the growth curve belongs to a known parametric family of curves. We observe change, or more precisely, reduction of gas consumption with increased outdoor temperatures, and seek a model with physically meaningful parameters, as the physical interpretability of parameters in the model motivates the choice of the growth curve.

In our study, we concentrate on the data observed at one specific station. Based on the Agreement [2008] between gas companies, we take the following sigmoidal growth model to describe the dependence of gas consumption on temperature:

$$y_i = S(t_i|\theta) + \varepsilon_i \,. \tag{1}$$

Here y_i denotes the standardized daily maximum flow and the corresponding mean function parameterized in $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ is given by

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\dfrac{\theta_2}{t_i - 40°\mathrm{C}}\right)^{\theta_3}} \,, \tag{2}$$

and also depends on the predictor t_i, which stands for the weighted four-day mean temperature with weights

$$w = (8, 4, 2, 1)/15 \,, \tag{3}$$

i.e.

$$t_i = \sum_{j=0}^{3} w_j t_{ji} \,, \tag{4}$$

where t_{0i}, t_{1i}, t_{2i} and t_{3i} are the temperatures in °C of the last four days. Finally, ε_i is an error term reflecting zero mean and constant variance σ^2.

Based on the physical properties of gas, Geiger and Hellwig [2002] and Cerbe [2008] introduce this kind of models for the description of typical gas loads in dependence of air temperature. According to the description of the log-logistic model provided by Ritz and Streibig [2008], the four parameters in model (2) have the following meaning: θ_1 and θ_4 are the upper and lower horizontal asymptotes, and the other two parameters describe the shape of the decrease of the (logistic like) curve. More precisely, θ_2 is the inflection point of the curve, i.e. the point around it is symmetric on logarithmic temperature axis, and the parameter θ_3 is proportional to the slope at $t_i - 40°\mathrm{C}$ equals to θ_2. This follows from the identity

$$\left(\frac{\theta_2}{t_i - 40°\mathrm{C}}\right)^{\theta_3} = \exp\Big(- \theta_3\big(\log(t_i - 40°\mathrm{C}) - \log\theta_2\big)\Big) \,.$$

Similarly, the properties of logistic growth models, as explained in Seber and Wild [2003], imply that θ_3 acts as a scale parameters on $t_i - 40°\mathrm{C}$, thus influencing the growth rate.

From the point of view of the energy industry, Geiger and Hellwig [2002] discuss the meaning of parameters in the following way: θ_4 describes the constant share of energy for warm water supply and process energy, while the difference $\theta_1 - \theta_4$ explains extreme daily gas consumption on cold days. θ_2 indicates the beginning of the heating period, i.e. the change point from the constant gas loads in summer to the increasing consumption in the heating period at approximately $15°\mathrm{C} - 18°\mathrm{C}$, and θ_3 measures flexibly the dependance in the heating period.

Geiger and Hellwig [2002] note that apart from the choice of the appropriate mean function $S(t_i|\theta)$, the adequate aggregation of mean daily temperatures to be included in the explanatory variable t_i is essential. Physical properties of buildings play an important role here. The four-day mean temperature is motivated by the fact that typical buildings in Germany accumulate the heat up to 85 hours, and the use of weights as in (3) is suggested. The weights given by (3) are obtained from the standardized geometric series with basis 2 applied to the temperature of the last four days, i.e.

$$t_i = \frac{t_{0i} + \frac{t_{1i}}{2} + \frac{t_{2i}}{4} + \frac{t_{3i}}{8}}{1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \sum_{j=0}^{3} w_j t_{ji} \, .$$

Based on these facts, German gas companies signed the Agreement [2008], which regulates different issues related to gas transmission within the German network, and agreed to use the sigmoidal function $S(t_i|\theta)$ defined in (2) with the explanatory variable t_i given by (3) and (4) to describe the dependance of gas loads on air temperature and to forecast the gas consumption at the design temperature.

The starting values for the iteration necessary to calculate the least squares estimates, as provided in the Agreement [2008], are given in the Table 1. As the parameters in the mean function of the non-linear regression model have physical interpretation, the starting values may be read from a plot of the data. According to Seber and Wild [2003], the crude initial estimates of θ_1 and θ_4 are calculated from the scatter plot, while θ_2 and θ_3 can be obtained using the linearization

$$y_i^* = \log\left(\frac{\theta_1 - \theta_4}{S(t_i|\theta) - \theta_4} - 1\right) = -\theta_3 \log(40°\mathrm{C} - t_i) + \theta_3 \log(-\theta_2) \, .$$

The substitution

$$\begin{aligned} \alpha &= \theta_3 \log(-\theta_2) \, , \\ \beta &= -\theta_3 \, , \end{aligned}$$

yields

$$y_i^* = \alpha + \beta \log(40°\mathrm{C} - t_i) \, ,$$

which enables an easy calculation of θ_2 and θ_3.

For repeated use of the same non-linear regression model some automated way of providing starting values is very important. One possibility is to construct a self-starter function. Self-starter functions substitute a manual search for starting values and are specific for a given mean function. They calculate the initial values for a given data set and make any further analysis based on the given model considerably easier.

We construct a self-starter function for the mean function (2) and calculate starting values for the given data set. For the resulting starting values, given in Table 1, the model fit is identical to the model fit obtained using the starting values supplied by another methods. We note here that the estimation algorithm based on the self-starter has a better convergence rate. The self-starter function is implemented in R (for the manual see R Development Core Team [2008]), based on the procedure suggested by Ritz and Streibig [2008]. We refer to Table 1 for both sets of starting values and for the estimated parameters and their standard errors when fitting model (2).

Table 1: Initial values (A, SS), least squares estimates (LSE) and standard errors (SE) of the parameters in the sigmoidal model (2).

Model (2)	θ_1	θ_2	θ_3	θ_4
Agreement (A)	2.509	-34.721	5.816	0.121
Self-Starter (SS)	2.243	-34.345	6.884	0.447
LSE	2.033	-32.647	6.664	0.447
SE	(0.032)	(0.214)	(0.218)	(0.009)

In Figure 1 we show the model fitted to data describing the typical gas outflow for public utilities, as well as the curves corresponding to both sets of starting values (denoted by A and SS) given in Table 1. The plot suggests that the considered sigmoidal model (2) (denoted by M) reproduces the gross characteristics of the gas flow well, though it obviously underestimates the mean flow for low temperatures.

The value of the Akaike Information Criterion (AIC) of this model is 31046.8. Here and in what follows we use

$$\text{AIC} = n \log \tfrac{1}{n}\text{SSE}(\hat{\theta}) + 2p\,,$$

with the minimized sum of squared errors

$$\text{SSE}(\hat{\theta}) = \sum_{i=1}^{n} (y_i - S(t_i|\hat{\theta}))^2\,,$$

and where p denotes the number of parameters in the mean model.

The next step is to extend model (2) with the aim to include an additional predictor, which describes the effect of weekend and holidays on the gas flow. To this end, we introduce an indicator variable d_i. It indicates whether the i-th gas load is observed on a working day or not. So we additionally include

$$d_i = \begin{cases} 1 & \text{if day } i \text{ is a working day,} \\ 0 & \text{if day } i \text{ is a holiday or at weekends,} \end{cases} \tag{5}$$

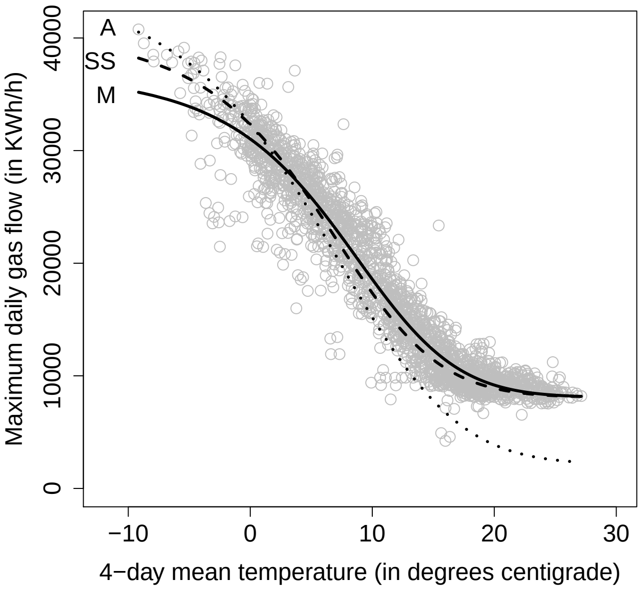

Figure 1: Fitted sigmoidal model (M) given by (2) (solid line) and curves based on starting values from the self-starter function (SS, dashed line) or from Agreement (A, dotted line). Data points represent the maximal daily outflows at a public utility depending on the four-day mean air temperatures.

in model (2) and obtain

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\dfrac{\theta_2}{t_i - 40^\circ\mathrm{C}} + d_i \theta_5 \right)^{\theta_3}} . \tag{6}$$

It is interesting to note that the AIC drops down to 30800.0 and that the estimate of this factor is $\hat{\theta}_5 = -0.054\ (0.003)$ generally indicating larger gas loads on working days.

To investigate the influence of the temperature more precisely, we extend the sigmoidal model further. Instead of using the four-day mean temperature with fixed weights w given by (3) as in the model defined by growth function (2), we allow for flexible temperature parameters. However, we stick to the temperature of the last four days, and consider two models. In terms of the sigmoidal function, we have

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\dfrac{\theta_2}{\frac{8}{15} t_{0i} + \theta_{21} t_{1i} + \theta_{22} t_{2i} + \theta_{23} t_{3i} - 40^\circ\mathrm{C}} + d_i \theta_5 \right)^{\theta_3}} , \tag{7}$$

or

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 - \theta_4}{1+\left(\frac{\theta_{20}}{t_{0i}-40°\mathrm{C}}+\frac{\theta_{21}}{t_{1i}-40°\mathrm{C}}+\frac{\theta_{22}}{t_{2i}-40°\mathrm{C}}+\frac{\theta_{23}}{t_{3i}-40°\mathrm{C}}+d_i\theta_5\right)^{\theta_3}} \,. \quad (8)$$

With this approach we want to find out what the coefficients of the optimal predictor for this data set are, i.e. what are the optimal weights for the four temperatures (today's, yesterday's, two and three days ago). As mentioned before, physical properties of gas imply that the gas flow depends not only on the today's temperature, but the temperature of a few previous days influences the flow as well. In these two models, the temperature coefficients are optimally estimated from data, and we analyze how the temperature effects the predicted flow. We emphasize here that the given weights w as in (3) imply that we already know the exact relationship between the temperatures and the gas flow, which is not true. So we try to estimate this relationship.

Model (7) can be seen as a direct extension and a flexible counterpart of model (2), with working day indicator and $\theta_{20} = w_0 = 8/15$. Model (8) enables an assessment of the importance of the flexible parametrization. This motivates the choice of the starting values for both models. In the case of model (7), we take

$$\theta_{2j} = w_j \,, \qquad j = 1, 2, 3,$$

as the initial values for the new parameters and for model (8) we use the starting value $\theta_2 = -34.345$ from model (2) and split it up into

$$\theta_{2j} = w_j\theta_2 \,, \qquad j = 0, 1, 2, 3 \,.$$

Initializing $\theta_5 = 0$ implies that weekdays do not influence the gas flow. However, as shown in Table 2 the effect of working days on gas consumption is significantly relevant under both models. Starting values, parameter estimates and their standard errors for models (7) and (8) are shown in Table 2.

The results indicate that the today's temperature and a sort of difference between the temperatures prevailing three and two days ago seem to be particularly relevant. The AIC values of 30687.6 for model (7) and 30688.4 for model (8) indicate an enormous improvement of the fit, but do not give a clear preference to model (7) or (8).

All models considered so far reproduce the gross characteristics of the gas flow well, though they obviously underestimate the mean responses for low temperatures, mainly because of the horizontal upper asymptote. This motivates the generalization of the upper asymptote by allowing an extra slope parameter and leads us to the so-called Brain-Cousens model (BC-model), convenient for describing the phenomenon called hormesis.

First, we examine the influence of the slope parameter in the case of the model

Table 2: Initial values (SS), least squares estimates (LSE) and standard errors (SE) of the parameters in models (7) and (8), both allowing for flexible weights.

Model (7)	θ_1	θ_2	θ_{21}	θ_{22}	θ_{23}	θ_3	θ_4	θ_5
SS	2.243	-34.345	0.266	0.133	0.067	6.884	0.448	0
LSE	2.025	-35.047	0.179	-0.267	0.345	8.674	0.433	-0.040
SE	(0.028)	(0.296)	(0.069)	(0.069)	(0.042)	(0.686)	(0.010)	(0.004)

Model (8)	θ_1	θ_{20}	θ_{21}	θ_{22}	θ_{23}	θ_3	θ_4	θ_5
SS	2.243	-18.316	-9.160	-4.578	-2.291	6.884	0.447	0
LSE	2.047	-24.671	-4.753	9.033	-13.502	6.397	0.445	-0.054
SE	(0.029)	(1.386)	(2.593)	(2.547)	(1.344)	(0.190)	(0.009)	(0.003)

with fixed temperature parameters. The simple BC-model is defined by

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 + \theta_6\left(\dfrac{\theta_2}{t_i - 40^\circ\mathrm{C}}\right) - \theta_4}{1 + \left(\dfrac{\theta_2}{t_i - 40^\circ\mathrm{C}} + d_i\theta_5\right)^{\theta_3}}\,. \tag{9}$$

As we already know that flexible temperature weights yield better results, we are interested in the BC-extension of the model given by (8), which has the form

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 + \theta_6\left(\dfrac{\theta_{20}}{t_{0i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{21}}{t_{1i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{22}}{t_{2i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{23}}{t_{3i} - 40^\circ\mathrm{C}}\right) - \theta_4}{1 + \left(\dfrac{\theta_{20}}{t_{0i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{21}}{t_{1i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{22}}{t_{2i} - 40^\circ\mathrm{C}} + \dfrac{\theta_{23}}{t_{3i} - 40^\circ\mathrm{C}} + d_i\theta_5\right)^{\theta_3}}\,. \tag{10}$$

The results of both fits are given in Table 3. Again, the initial value of 0 for θ_6 implies that the upper asymptote is a horizontal line. The parameters change somewhat, and the new slope parameter in the model is significant. Thus, the upper asymptote is a line with negative slope θ_6. For low temperatures, the modified upper asymptote in the BC-model implies the increase of the mean gas flow for approximately 2 times scaled $\bar{y}$ when the temperature decreases for $1^\circ\mathrm{C}$. Yesterday's temperature seems to be the least relevant of all observed temperatures and the difference between temperatures three and two days ago seems to be important again. As the AIC value of model (9) we get 30764.8 (compare with 30800.0, the AIC value of model (6) with horizontal asymptote). The lower AIC value of 30655.0 characterizes model (10), and indicates the most adequate parametric model.

All parametric models are fitted in R using the function `nls()`.

Table 3: Initial values (SS), least squares estimates (LSE) and standard errors (SE) of the parameters in the simple BC-model (9) and under its extended version allowing for flexible weights (10).

Model (9)	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
SS	2.243	-34.345	6.884	0.447	0	0
LSE	3.232	-29.603	6.071	0.514	-0.100	-1.914
SE	(0.107)	(0.537)	(0.302)	(0.018)	(0.013)	(0.189)

Model (10)	θ_1	θ_{20}	θ_{21}	θ_{22}	θ_{23}	θ_3	θ_4	θ_5	θ_6
SS	2.243	-18.316	-9.160	-4.578	-2.291	6.884	0.447	0	0
LSE	3.226	-21.540	-4.184	7.662	-11.640	6.074	0.511	-0.097	-1.883
SE	(0.112)	(1.255)	(2.253)	(2.215)	(1.185)	(0.296)	(0.017)	(0.013)	(0.197)

4 Semi-Paramteric Models

The nuances missed by the sigmoidal models as well as the numerical sensibility of the BC-models motivate the search for an alternative model. As one possibility, Mirkov et al. [2010] suggest the use of the FlexMix approach introduced in Leisch [2004], which offers a framework for flexible fitting of finite mixtures of (generalized) linear regression models. This approach improves the fit for low temperatures, but the approximation of our non-linear models by polynomials is unstable. In our opinion, the FlexMix approach would offer better results, if the method would be generalized to mixtures of sigmoidal models. Another possibility is to use semi-parametric models, such as locally weighted regression (LOWESS cf. Cleveland [1979]) or spline models. Unfortunately, LOWESS models are not suitable for prediction. More about general splines theory can be found e.g. in De Boor [1990], Hämmerlin and Hoffmann [1994] or Powell [1991]. Many authors propose some variant of spline regression for this kind of problems, see e.g. Jones et al. [2009], Vitezica et al. [2010], Jarrow et al. [2004], Mackenzie et al. [2005], Cadorso-Suárez et al. [2010], Riedel and Imre [1993].

We choose the penalized splines approach, based on Wegman and Wright [1983], Eilers and Marx [1996] and Eilers and Marx [2010]. Simplicity and flexibility of the method motivate the choice. The advantage of P-splines over B-splines is easy control of smoothness as well as the simple way to handle knots, i.e. their number and their positions. As Jarrow et al. [2004] emphasize, another advantage of the P-splines method is that knots can be chosen automatically. The number of knots should be sufficiently large (at least 8), to accommodate the non-linearity of the underlying data, but a larger number of knots does not cause over-fitting provided the smoothing parameter is suitably chosen.

We use the following model to describe the dependence of the standardized maximal daily gas loads y_i on temperature t_i:

$$y_i = S_\Delta(t_i) + \varepsilon_i \,, \tag{11}$$

where $y_i = y_i^{max}/\bar{y}$, and $\bar{y}$ is the empirical mean of all maximal daily gas flows at a particular node of the network and t_i stands for the weighted four-day mean temperature with the weights w given in (3). The function $S_\Delta(t)$ is the linear combination of basis functions $B_j, j = 1, \ldots, m$, on the mesh Δ, given by

$$S_\Delta(t_i) = \sum_{j=1}^{m} a_j B_j(t_i) \,, \tag{12}$$

and ε_i, $i = 1, \ldots, n$, are random noise terms reflecting zero mean and constant variance. B_j are basis functions of the B-spline of degree q, and the mesh Δ is an equidistant grid over $m - q$ segments, i.e. with $m - q + 1$ inner knots.

If we introduce a smoothing penalty λ, then instead of minimizing a least squares criterion like

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} a_j B_j(t_i) \right)^2 \,,$$

the objective function to be minimized is the penalized residual sum of squares

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} a_j B_j(t_i) \right)^2 + \lambda \sum_{j=k+1}^{m} (\delta^k a_j)^2 \,, \tag{13}$$

where δ^k denotes the k-th order finite differences of the coefficients of the corresponding B-splines. In fact, δ^k is a discrete approximation of the k-th derivative of S_Δ, usually used in the smoothing splines approach.

Minimizing (13) is equivalent to solving

$$B^\top y = (B^\top B + \lambda D_k^\top D_k) a \,,$$

where D_k is the matrix representation of the difference operator δ^k. The elements of the matrix B are $b_{ij} = B_j(t_i)$, and the vector y consists of the responses y_i, for $i = 1, \ldots, n$, and $j = 1, \ldots, m$. Obviously, a is the vector of the coefficients a_j.

Note, when $\lambda = 0$, we have the standard linear regression with a B-spline basis. In that case the fitted curve is over-fitting the data, giving a result with too many fluctuations. By increasing λ the smoothness is tuned. In the limit of a very large λ a linear or quadratic fit (depending on the degree of the basis function) is obtained.

Now that we can control the smoothness of the fitted curve with λ, we want to choose its value so that the AIC of the model is minimal. In the case of P-splines, p denotes the effective dimension of the vector of parameters, which can be approximated by the trace of the hat matrix H. The effective dimension provides an intuitive measure for the complexity of a P-spline fit that appropriately accounts for the effective dimensionality reduction induced by the penalty (cf. Cadorso-Suárez et al. [2010]). Since H, with the diagonal elements h_{ii}, can be expressed as

$$H = B(B^\top B + \lambda D_k^\top D_k)B^\top \,,$$

we can easily calculate

$$p = \mathrm{tr}(H) = \sum_{i=1}^{n} h_{ii} \,.$$

As an estimate $\hat{\sigma}^2$ of the variance for the optimal λ, Eilers and Marx [1996] suggest to use

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \Big(y_i - \hat{S}_\Delta(t_i) \Big)^2 \,.$$

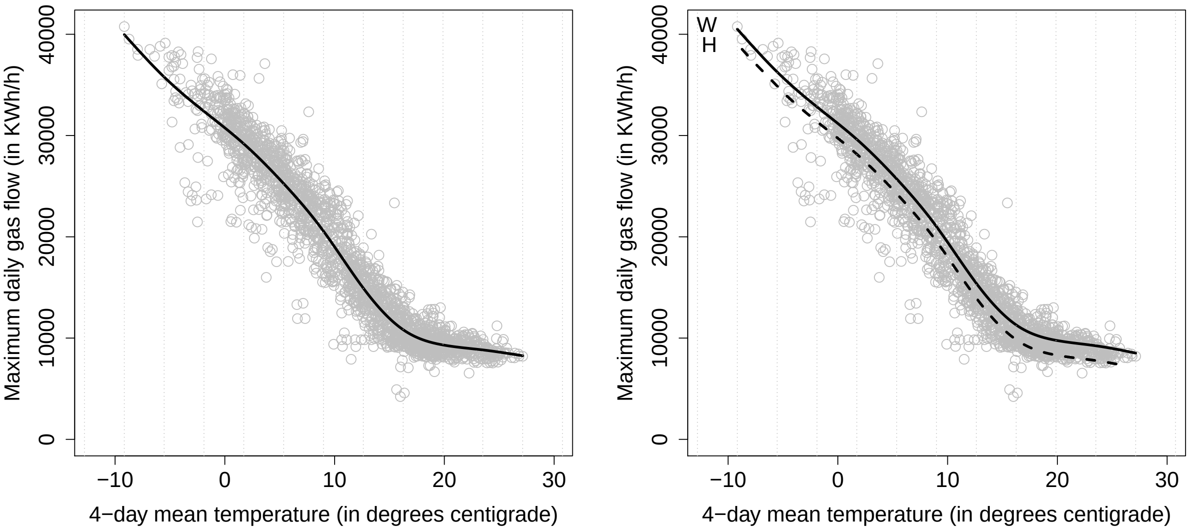

Figure 2: Penalized splines regression given by model (12) and (14), i.e. without (left) and with working day indicator (right), respectively. Data points illustrate the maximal daily outflow at a public utility in dependence of the four-day mean air temperature. The position of the inner knots, based on cubic B-splines on the mesh with 10 segments and the second order penalty $\lambda = 10^{0.4} = 2.51$ is also shown. The effect of working (W) days (solid line) and holidays (H) and weekends (dashed line) from the model defined by (14) is obvious.

If we further include the predictor d_i in our model, then model (12) becomes

$$S_\Delta(t_i) = \sum_{j=1}^{m} a_j B_j(t_i) + a_{m+1} d_i \,. \tag{14}$$

With the smoothing penalty λ we want to minimize

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} a_j B_j(t_i) - a_{m+1} d_i\right)^2 + \lambda \sum_{j=k+1}^{m} (\delta^k a_j)^2 \,. \tag{15}$$

In a matrix form, the optimality condition for (15) becomes

$$C^\top y = (C^\top C + \lambda D_k^\top D_k)\tilde{a}\,,$$

where D_k is the matrix representation of the difference operator δ^k and y are the standardized responses. The elements of the matrix C are c_{ij}, where

$$c_{ij} = \begin{cases} b_{ij} & \text{for } i = 1,\dots,n, \;\; j = 1,\dots,m, \\ d_i & \text{for } i = 1,\dots,n, \;\; j = m+1, \end{cases}$$

and $\tilde{a}$ is the vector of regression coefficients a_j, $j = 1,\dots,m+1$.

Figure 2 (left) shows the model fit and the position of the inner knots, based on cubic B-splines on the mesh with 10 segments and the second order penalty $\lambda = 10^{0.4} = 2.51$. The AIC value of 30993.5 implies much better fit than the corresponding parametric model (6). In the same setting, the model which includes the working day indicator, with the AIC value of 30808.9, is displayed in Figure 2 (right). The upper and lower curves show the fit for working days (W) and weekends and holidays (H), respectively.

We note here that P-splines allow straightforward interpolation and extrapolation. When extrapolating, the B-spline coefficients form a polynomial sequence of degree $k-1$. Thus, the forecast values depend critically on the order of penalty k.

Furthermore, since the forecast includes the level of uncertainty described by the confidence or prediction intervals, the assumption about the distribution of the error terms plays here an important role. P-splines can be applied both to normal and non-normal data, and we note that considerations included here assume the Gaussian distribution. In particular, the calculation of the AIC and the estimation of the model parameters is influenced by this assumption. Alternatively, the bootstrap approach for interval estimation and some other model selection criterion, e.g. Bayesian information criterion (BIC) could be used.

5 Comparison of Models

After presenting several different models appropriate for this kind of data, we would like to compare the goodness of fit of those models.

Figure 3 (left) illustrates the difference between the fit of the sigmoidal model (6) and the simple BC-model (9) in case of working days. For temperatures above $0°\mathrm{C}$ the fits are almost identical, but for temperatures below $0°\mathrm{C}$, the modified upper

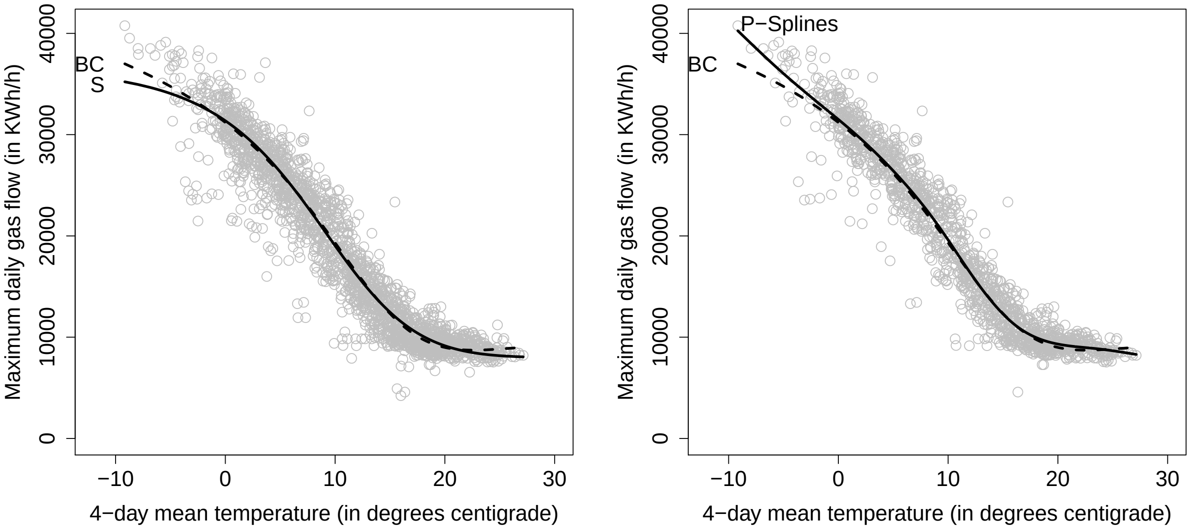

Figure 3: Fitted sigmoidal (S) model (solid line) vs. simple BC-model (dashed line) (left) and BC-model (dashed line) vs. P-splines (solid line) fit (right) for working days. Notice the different behavior of the models for cold temperatures. Data points illustrate the maximal daily outflow on working days at a public utility in dependence of the four-day mean air temperature.

asymptote of the BC-model implies a higher mean gas flow than the other sigmoidal models. Due to the local smoothing methodology, the P-splines model displays even higher consumption of gas for low temperatures. This behavior can be clearly seen in Figure 3 (right), and will be even more important for forecast of gas load at the design temperature.

As model selection criterion we use the value of the AIC. A difference of more than 10 is required in order to definitely prefer one model over another model (cf. Ritz and Streibig [2008]). Table 4 gives AIC values of all considered models.

Obviously, extensions to the original parametric model (2) improve the AIC value gradually. A further explanatory indicator d_i and estimated weights in the temperature effect both improve the model's AIC values for 246.8 and 359.5 or 358.4 units, respectively. Note that both models with estimated temperature weights have better AIC values than that based on the four-day mean temperature. We cannot prefer any of the model (7) or (8) since their AIC values are pretty close. Eventually, we quantify the effect of the slope of the upper asymptote. The AIC of the complex BC-model (10) is 33.4 units less than that of the model (8), indicating that the fit of the complex BC-model is the best among all parametric models.

In the case of semi-parametric models, the initial model based on P-splines given by (12) yields an AIC value which is 53.3 units less than that of the simplest sigmoidal model (2). Including the effect of weekdays in the model improves the fit also in the

Table 4: Comparison of AIC values of the models considered.

Model	AIC
(2)	31046.8
(6)	30800.0
(7)	30687.6
(8)	30688.4
(9)	30764.8
(10)	30655.0
(12)	30993.5
(14)	30808.9

semi-parametric case. The difference of 184.6 units between models (12) and (14) is somewhat less than in the parametric case, but the result is very promising especially keeping in mind the prediction of gas loads for low temperatures.

6 Application of Modeling Results

The main purpose of the presented models is a reliable forecast of gas loads at the design temperature, in order to sustain the supply of gas without failure. The models given by (10) and (14) in Sections 3 and 4 are utilized for the prediction of gas loads at the design temperature. Recall, the design temperature in Germany lies between $-12°\mathrm{C}$ and $-16°\mathrm{C}$, depending on the climate conditions in the region. As such low mean daily temperatures are very uncommon in Germany, there is no gas flow data available at the design temperature, and gas operators like OGE use the predicted gas loads at the design temperature for different purposes in their daily business.

The predicted values based on the BC-model are obtained using the `predict` method in R, as described in Ritz and Streibig [2008]. P-splines allow straightforward smooth extrapolation, and we exploit this property to forecast gas loads at the design temperature. The second order penalty implies the extrapolation by a linear sequence, implying that the gas consumption is increasing at the constant rate as the temperature decreases, cf. Currie et al. [2004] or Eilers and Marx [2010].

Since the design temperature is outside of the domain of the observed predictor variable, we create a new data frame to determine the predicted gas loads. To this end, we generate a regular grid of four-day mean temperatures $\tilde{t}_k$ starting from the lowest possible design temperature, i.e. from $-16°\mathrm{C}$, and go up to $30°\mathrm{C}$ with step size $1°\mathrm{C}$. Based on the new data and the fitted models (10), and (14), the predictions $\tilde{y}_k = \hat{S}(\tilde{t}_k)$ are calculated.

Figure 4 illustrates the prediction based on the BC-model and on the P-splines regression. At the design temperature of $-12°\mathrm{C}$ the predicted gas loads on working

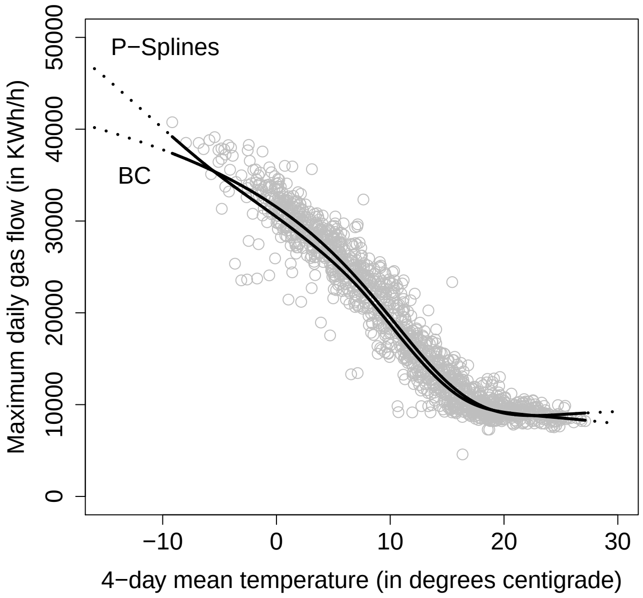

Figure 4: Fitted models (solid lines) and predictions (dotted lines) for working days based on the BC- and on a P-splines model. Data points show the maximal daily outflow on working days at a public utility in dependence of the four-day mean air temperature.

days based on models (9) and (14) are 39 and 43 MWh/h, respectively.

Figure 5 (left) represents the predicted values for working days based on the BC-model (9), and the corresponding (pointwise) 95% prediction intervals. The construction of the band is based on the additional assumption of Gaussian distributed responses for convenience and utilizes the delta method to obtain approximations of the variances of the raw residuals. Note here that it is the estimated variance of the responses and not the approximative variance of the fitted values that makes the variance of the residuals. Because of this, the prediction intervals have very similar widths over the entire range of four-day mean temperature values. Further methods for constructing prediction intervals for non-linear regression can be found in Gauchi et al. [2010]. In the case of P-splines, the Bayesian estimate of the covariance matrix is used to construct the prediction band. The fitted P-splines model (14) for working days together with the corresponding 95% prediction intervals is shown in Figure 5 (right). As already mentioned, in this particular case prediction bands are constructed under the additional assumption of normally distributed error terms but it could be done with other distributions also.

To ensure the reliable operation of gas transmission networks, gas operators like OGE use this kind of statistical models and the prediction based on them to determine technically available transportation capacities on entries, and to validate gas network

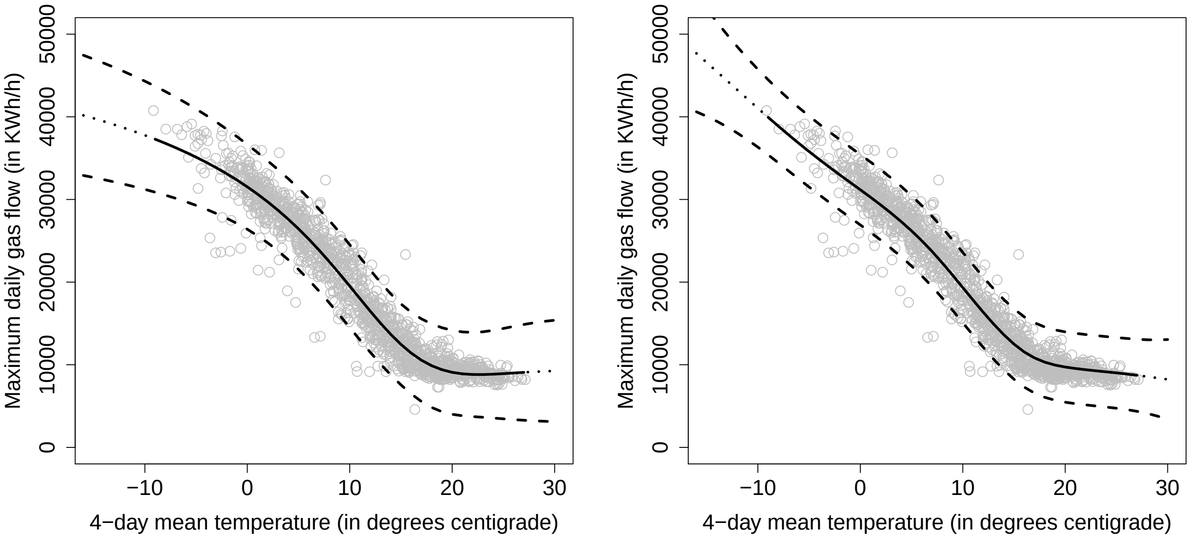

Figure 5: Fitted (solid line) and extrapolated (dotted line) gas consumption for working days based on the BC model (left) and on P-splines (right) with corresponding 95% prediction intervals (dashed lines). Data points correspond to the maximal daily outflow on working days at a public utility in dependence of the four-day mean air temperature.

nominations. A nomination describes the balanced in- and outflow of gas at entries and exits of the network, and needs to be feasible for every temperature, including the design temperature. A nomination is said to be feasible, or validated, if for the given inflow and outflow allocation, a flow of gas through the network exists, taking into account all technical limitations of the gas network and physical properties of gas. The mean maximal gas inflow on exits of the network over the whole temperature range necessary for the nomination validation can be estimated using the presented models. Detailed elaboration of the nomination validation can be found in Fügenschuh et al. [2011b]. The inflow on entries, expressed in terms of the technically available capacities, is assessed from the outflows of the neighboring exits, where additionally an unfavorable inflow allocation is considered. Here, for instance, one can also use projection intervals to determine the realistic or projected minimal and maximal flow at the given exit. These concepts are used for single exits and entries as well as for groups of exits and entries, which belong together regionally and share the same technical properties, the so called exit zones. The same type of models and the obtained results can be applied to the outflow of an exit zone.

Under an appropriate distributional assumption for the error term and the variance homogeneity, gas operators could also evaluate the flow limits for the given acceptable level of risk at the given temperature. The mean value as well as the prediction intervals of the models are the basis for the evaluation of some types of contracts, such as freely allocable capacity (FAC) contracts and contracts which arrange the cooperation on market crossings (CMC). The predicted gas loads at the design temperature are matched against the booked capacities from the existing contracts, and new contracts are evaluated according to the relation between the observed loads. Depending on the type of the exit we are dealing with (e.g. whether the exit is a public utility or an industrial consumer), a new FAC contract can be entered or a new latent booking is anticipated. The CMC contracts, and other temperature-dependent contracts can be also evaluated based on the results of this study.

Finally, we want to mention that the results of this study are used within the ForNe project supported by OGE. The aim of the project is the analysis and optimization of gas transmission networks. In particular, we investigate stochastic aspects of gas transport. Statistical models appear as a natural tool for describing the dependence of gas outflow and air temperature, and are especially convenient for the prediction. The predicted gas loads on exits of the network at the design temperature according to the P-splines model (14) are utilized as input data for the process of the network optimization, see for e.g. Koch et al. [2011], Martin et al. [2011], Fügenschuh et al. [2011a], Geissler et al. [2011]. Here, the P-splines approach is chosen because of its flexibility.

7 Conclusions

We study several models useful for forecasting gas flow on exits of gas transmission networks based on parametric and semi-parametric statistical modeling techniques.

Preliminary results show that a simple sigmoidal model enables a good starting approach to the observed problem. The shape of the sigmoidal function we used to model the dependence of the maximum gas flow from temperature is suitable, but there is room for improvement. We suggest the BC-model in its simple and extended form in the parametric setting as it reflects the behavior of gas flow for low temperatures in a more realistic way. The flexible temperature parameters are also important for the good model fit, leading us to the extended BC-model, which is the most appropriate parametric model.

In the case of the semi-parametric approach, we utilize the P-splines, which represent a very flexible semi-parametric alternative and model gas behavior in an adequate manner, especially for very low temperatures.

However, we note that non-linear regression models are generally more difficult to handle than the local smoothers like the P-splines, since we have to take care about many issues like starting values, (numerical) derivation and convergence properties. In particular, the extended BC-model is very sensitive both to the choice of initial values and to numerical derivation. Contrary to them, the P-splines methodology is simpler and numerically less problematic, but in order to exploit the advantage of flexible temperature effects one would have to use general additive models (cf. Eilers and Marx [2002]) or two-dimensional P-splines regression (cf. Eilers and Marx [2003]).

The forecast of mean gas loads based on parametric models, in our case the BC-model, is safer than the one relying on the P-splines, due to the numerical properties of the models.

Both approaches call for a careful choice of the error term distributions, as the errors of maxima are usually not Gaussian. This issue requires further study and is beyond the scope of this paper.

The practical use of the predicted mean gas consumption at the design temperature is highlighted.

Acknowledgements

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.

We are thankful to the co-editor-in-chief Prof. Vijay Nair and to an anonymous referee for all the insightful comments made, and we appreciate useful hints about the multiple regression techniques based on P-splines made by Prof. Brian Marx.

References

Coopertion Agreement. Vereinbarung über die Kooperation gemäß §20 Abs. 1b) EnWG zwischen den Betreibern von in Deutschland gelegenen Gasversorgungsnetzen. Bundesministerium der Justiz Deutschland, 2008.

C. Cadorso-Suárez, L. Meira-Machado, T. Kneib, and F. Gude. Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. *Statistical Modelling*, 10(3):291–314, 2010.

G. Cerbe. *Grundlagen der Gastechnik*. Technik. Hanser Verlag, Leipzig, Germany, 2008. ISBN 978-3-446-41352-8.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

I.D. Currie, M. Durban, and H.C. Eilers. Smoothing and Forecasting Mortality Rates. *Statistical Modelling*, 4:279–298, 2004.

C. De Boor. *Splinefunktionen*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1990. ISBN 3-7643-2514-3.

H.C. Eilers and B.D. Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121, 1996.

H.C. Eilers and B.D. Marx. Generalized Linear Additive Smooth Structures. *Journal of Compuational and Graphical Statistics*, 11(4):758–783, 2002.

H.C. Eilers and B.D. Marx. Multidimensional Calibration with Temperature Interaction Using Two-Dimensional Penalized Signal Regression. *Chemometrics and Intell Lab Sys*, 66:159–174, 2003.

H.C. Eilers and B.D. Marx. Splines, Knots, and Penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.

A. Fügenschuh, B. Hiller, J. Humpola, T. Koch, T. Lehman, R. Schwarz, J. Schweiger, and J. Szabó. Gas Network Topology Optimization for Upcoming Market Requirements. *IEEE Proceedings of the 8th International Conference on the European Energy Market (EEM), 2011*, pages 346–351, 2011a.

A. Fügenschuh, B. Hiller, J. Humpola, T. Koch, T. Lehman, R. Schwarz, J. Schweiger, J. Szabó, and S. Vigerske. A Model and an Algorithm for Gas Network Nomination Validation. *Working Paper*, 2011b.

J-P. Gauchi, J-P. Villa, and L. Coroller. New Prediction Interval Band in the Nonlinear Regression Model: Application to Predictive Modeling in Foods. *Communications in Statistics - Simulation and Computation*, 2(39):322–334, 2010.

B. Geiger and M. Hellwig. Enwicklung von Lastprofilen für die Gaswirtschaft Gewerbe, Handel und Dienstleistungen. Technical report, Technische Universität München, Department of Electrical Engineering and Information Technology, Institute for Power Engineering, 2002.

B. Geissler, A. Martin, A. Morsi, and L. Schewe. Solving Large-Scale Optimiziation Problems on Gas Networks with MIP Techniques. *Working Paper*, 2011.

G. Hämmerlin and K.H. Hoffmann. *Numerische Mathematik*. Grundwissen Mathematik. Springer-Verlag, Berlin, Germany, fourth edition, 1994. ISBN 3-540-58033-6.

M. Hellwig. *Enwicklung und Anwendung parametrisierter Standard-Lastprofile*. PhD thesis, Technische Universität München, Department of Electrical Engineering and Information Technology, Institute for Power Engineering, 2003.

R. Jarrow, D. Ruppert, and Y. Yu. Estimating the Interest Rate Term Structure of Corporate Debt With a Semiparametric Penalized Spline Model. *Journal of the American Statistical Association*, 99(465):57–66, 2004.

G. Jones, Y. Leung, and H. Robertson. A Mixed Model for Investigating a Population of Asymptotic Growth Curves Using Restricted B-Splines. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(1):66–78, 2009.

T. Koch, H. Leövey, R. Mirkov, W. Römisch, and I. Wegner-Specht. Szenariogenerierung zur Modellierung der stochastischen Ausspeiselasten in einem Gastransportnetz. *Optimierung in der Energiewirtschaft, VDI-Berichte*, 2157:115–125, 2011.

F. Leisch. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004.

M.L. Mackenzie, C.R. Donovan, and B.H. McArdle. Regression Spline Mixed Model: A Forestry Example. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(4):394–410, 2005.

A. Martin, B. Geißler, C. Hayn, J. Humpola, T. Koch, T. Lehman, A. Morsi, M.E. Pfetsch, L. Schewe, M. Schmidt, R. Schultz, R. Schwarz, J. Schweiger, M.C. Steinbach, and B.M. Willert. Optimierung Technischer Kapazitäten in Gasnetzen. *Optimierung in der Energiewirtschaft, VDI-Berichte*, 2157:105–115, 2011.

R. Mirkov, H. Friedl, H. Leövey, W. Römisch, and I. Wegner-Specht. Sigmoid Models Utilized in Optimization of Gas Transportation Networks. In A.W. Bowman, editor, *Proceedings of the 25th International Workshop on Statistical Modelling*, pages 381–385. The University of Glasgow, 2010.

M.J.D. Powell. *Approximation theory and methods*. Cambridge University Press, New York, NY, USA, third edition, 1991. ISBN 0-521-29514-9.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

K.S. Riedel and K. Imre. Smoothing spline growth curves with covariates. *Communications in Statistics - Theory and Methods*, 22(7):1795–1818, 1993.

C. Ritz and J.C. Streibig. *Nonlinear Regression with R*. Use R! Springer, New York, NY, USA, 2008. ISBN 978-0-387-09615-5.

G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. Wiley's Series in Probability and Statistics. Wiley, Hoboken, NJ, USA, 2003. ISBN 978-0-471-47135-6.

Z.G. Vitezica, C. Marie-Etancelin, M. D. Bernadet, X. Fernandez, and C. Robert-Granie. Comparison of nonlinear and spline regression models for describing mule duck growth curves. *Poultry Science*, 89:1778–1784, 2010.

U. Wagner and B. Geiger. Gutachten zur Festlegung von standardlastprofilen Haushalte und Gewerbe für BGW und VKU. Technical report, Technische Universität München, Department of Electrical Engineering and Information Technology, Institute for Power Engineering, 2005.

E.J. Wegman and I.W. Wright. Splines in Statistics. *Journal of the American Statistical Association*, 78(382):351–365, 1983.