Chapter 13

Empirical observations and statistical analysis of gas demand data

Holger Heitsch, René Henrion, Hernan Leövey,
Radoslava Mirkov, Andris Möller, Werner Römisch,
Isabel Wegner-Specht

Abstract *In this chapter we describe an approach for the statistical analysis of gas demand data. The objective is to model temperature dependent univariate and multivariate distributions allowing for later evaluation of network constellations with respect to the probability of demand satisfaction. In the first part, methodologies of descriptive data analysis (statistical tests, visual tools) are presented and dominating distribution types identified. Then, an automated procedure for assigning a particular distribution to the measurement data of some exit point is proposed. The univariate analysis subsequently serves as the basis for establishing an approximate multivariate model characterizing the statistics of the network as a whole. Special attention is paid to the statistical model in the low temperature range.*

The goal of our data analysis consists in evaluating historical data on gas demand at exits of some gas transportation network. The results will be used to extract statistical information, which may be exploited later for modeling the gas flow in the network under similar temperature conditions. More precisely, the aim is to generate a number of scenarios of possible exit loads, which will be complemented in several subsequent steps to complete a nomination (see Chapter 14). Such scenarios are needed for validating the gas network and for calculating and maximizing its technical capacities.

The analysis will be based on historical measurement data for gas consumption, which is typically available during some time period, and on daily mean temperature data provided by a local weather service. Due to a high temperature-dependent proportion of heating gas, the gas demand is subject to seasonal fluctuations. During the warmer season the gas consumption decreases: hot water supply for households and process gas consumption are the only basic constituents.

The method for analyzing the data should be applicable to all exits, no matter what their distribution characteristics are, and should allow for multivariate modeling to take into account statistical dependencies of different exits of the network. Therefore, the use of local temperatures as in day-ahead prediction of gas demands is less appropriate. Rather, we introduce a reference temperature which is given as a weighted sum of several local

Table 13.1: Categorization of the data characteristics obtained for the data of the L-gas network. The values are given in percentage w.r.t. relative frequency of occurrence. Stationarity indicates that data do not contain significant trends over time.

Temperature dependency	Description	Stationarity	
		yes	no
dependent	sigmoidal	41.21	0.70
	sigmoidal and jumps to zero flow	12.52	0.28
	piecewise-linear (summer zero flow)	17.30	0.42
	band- and cluster-like structures	6.05	0.56
	non-regular cluster-like structures	1.55	0.56
	discrete data distribution	3.52	0.00
	other non-categorized relations	4.50	2.81
independent	completely positive	4.50	0.70
	jumps between zero and positive flow	0.14	0.00
	band-like structures and jumps	0.98	0.84
	uniform-like distribution	0.84	0.98
	local varying distribution	0.00	0.42
	discrete data distribution	0.70	0.14
	other non-categorized relations	2.53	0.56
	almost zero flow	3.66	0.00

temperatures and in this way is more representative for the entire network. Due to the stationarity of the gas flow model considered throughout this book, we consider daily averages of the gas demands at all exits, based on measurement data which is mostly given for smaller (e.g., hourly) time intervals.

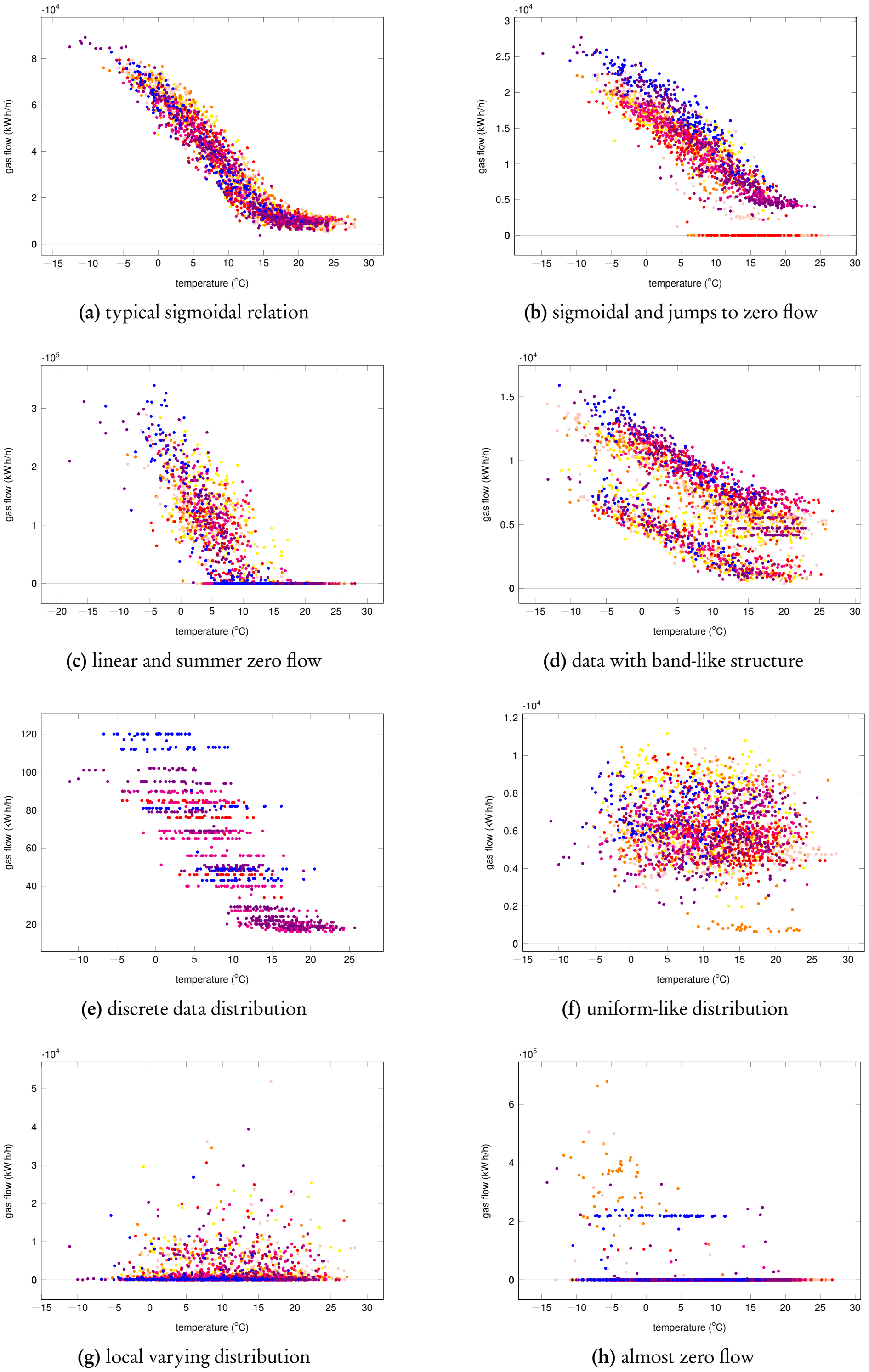
The statistical data analysis will consist of several steps, namely,

(i) visualization and categorization of the available data,
(ii) analysis of basic structural properties,
(iii) statistical tests,
(iv) definition of temperature classes,
(v) fitting univariate probability distributions for each exit and temperature class,
(vi) fitting multivariate distributions, and
(vii) forecasting gas demands for low temperatures.

Issues (i)–(iii) are discussed in Section 13.1, (iv) in Section 13.2, (v) in Section 13.3, (vi) in Section 13.4 and (vii) in Section 13.5. All tables and figures illustrating the different steps are based on historical data provided by the company Open Grid Europe GmbH (OGE) for its H-gas and L-gas transportation network. Both networks contain almost 800 exit points.

13.1 ▪ Descriptive data analysis and hypothesis testing

In this section, we discuss several methods to analyze gas demand data. These techniques form the basis for the succeeding steps to construct distributions.



(a) typical sigmoidal relation

(b) sigmoidal and jumps to zero flow

(c) linear and summer zero flow

(d) data with band-like structure

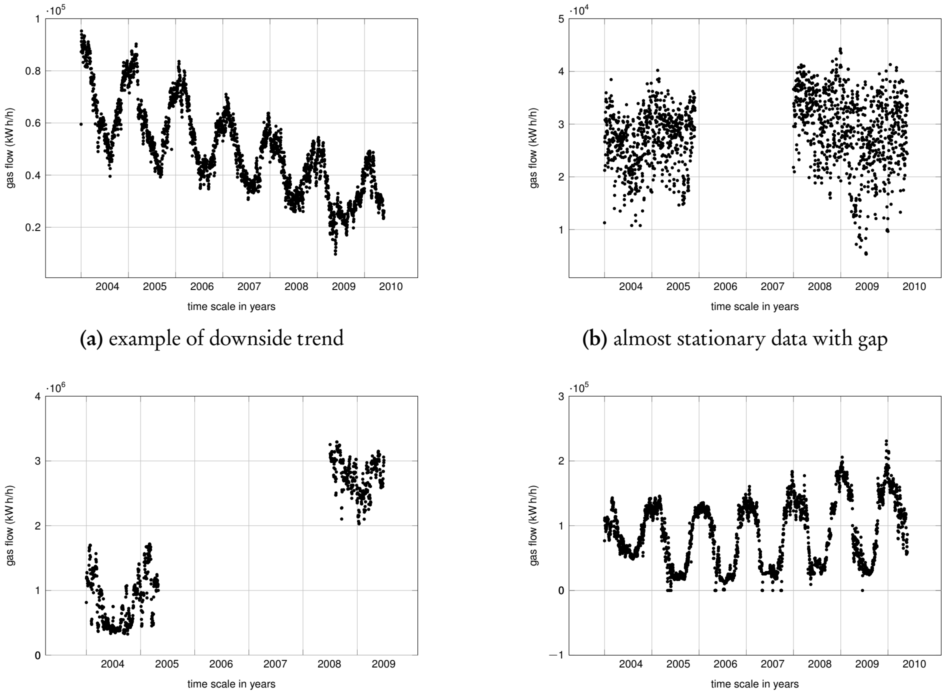
(e) discrete data distribution

(f) uniform-like distribution

(g) local varying distribution

(h) almost zero flow

Figure 13.1: Overview of the variety of data containing typical characteristics



(a) example of downside trend

(b) almost stationary data with gap

(c) non-stationary data with gap

(d) almost stationary data

Figure 13.2: Illustration of varying data behavior w.r.t. gaps and trends over time

13.1.1 ▪ Visualization and categorization

The first step of data analysis typically consists in the visualization and categorization of the available data for all exit points of a given network. Then basic structural properties (trends, clusters, temperature dependence etc.) should be analyzed, before in a third step different types of univariate (exit-based) statistical distributions (normal, uniform etc.) can be classified.

For the L-gas network and the given data base, Figure 13.1 illustrates the distribution of daily mean gas flows as a function of temperature, for selected exit points. Different years in the data base are visualized by different colors. As mentioned above, gas demands are mainly temperature dependent. The type of dependence varies between (piecewise) linear, sigmoidal and – in rare cases – other functional relations. Often the exit flow data are not positive, but contain zero flows for certain time periods without clear temporal or temperature dependence. Furthermore, one may be faced with band- or cluster-like structures.

Table 13.1 provides an overview of a general categorization in the L-gas network. Figure 13.2 reveals that one may have to cope also with the presence of outliers and of non-stationary characteristics – such as trends and breaks, which necessitate an appropriate data preparation prior to the actual model calibration (Brockwell and Davis 2002).

For gas demand forecast one typically uses non-linear regression models in which the functional relationship between the daily average gas flow and the daily mean temperature or a weighted mean temperature is described using sigmoid functions, see Cerbe (2008) and Leövey et al. (2011). For an interpretation of the model parameters in the context of gas flow modeling we refer, e.g., to Geiger and Hellwig (2002). Based on the visual analysis of the data originating from the given networks, we suspect that classic assumptions such as homogeneity of variance over the entire temperature range or approximate normality of residuals (Seber and Wild 2003) are frequently not met.

13.1.2 ▪ Statistical tests

In order to corroborate the latter observation one may apply suitable statistical tests. As far as homogeneity of variance for residuals of adjacent temperature intervals is concerned, the methodology of analysis of variance (ANOVA) can be employed. Here, a variety of statistical tests is at our disposal, e.g., Bartlett's test for normally distributed data or robust (distribution-free) alternatives such as the Siegel-Tukey test. For further details we refer to the comprehensive description in Hartung (2005) and Fahrmeir et al. (2007). The application of ANOVA to the available data showed that the assumption of homogeneous variances over the entire temperature interval is violated for a majority of exit points in the H-gas and L-gas network.

As far as the normality assumption is concerned, there exist numerous tests which can differ in their design and properties, compare D'Agostino and Stephens (1986), Hartung (2005), and Thode (2002). Probably the most prominent among these is the classical Kolmogorov-Smirnov goodness-of-fit-test. It checks whether a random variable obeys some fixed distribution.

To give a short description of the *Kolmogorov-Smirnov test*, let F_ξ denote the distribution function of a real random variable ξ on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, i.e.,

$$F_\xi(t) := \mathbb{P}(\xi \leq t) \quad \text{for all } t \in \mathbb{R}.$$

The *Kolmogorov distance* between the distributions $\mathbb{P} \circ \xi^{-1}$ and $\mathbb{P} \circ \eta^{-1}$ of two real random variables ξ and η is given as the uniform distance of the corresponding distribution functions, i.e.,

$$\mathbb{D}_K(\mathbb{P} \circ \xi^{-1}, \mathbb{P} \circ \eta^{-1}) := \sup_{t \in \mathbb{R}} \left| F_\xi(t) - F_\eta(t) \right|. \tag{13.1}$$

The observed data $\{x_1, \ldots, x_N\}$ induce a new random variable ξ^N, whose distribution (called *empirical distribution*) is defined by the empirical distribution function

$$F_N(t) := \frac{1}{N} \#\{x_i \mid x_i \leq t\}. \tag{13.2}$$

Assuming that the measurement data are ordered ($x_1 \leq \ldots \leq x_N$) the Kolmogorov-Smirnov test statistic is given by the Kolmogorov distance between the distributions of ξ and ξ^N, i.e.,

$$D_N = \sup_{t \in \mathbb{R}} \left| F_\xi(t) - F_N(t) \right| = \max_{1 \leq i \leq N} \left\{ F_\xi(x_i) - \tfrac{i-1}{N}, \tfrac{i}{N} - F_\xi(x_i) \right\},$$

under the null hypothesis that the measurement data $\{x_1, \ldots, x_N\}$, are drawn from the distribution of ξ. Significant advantages of the Kolmogorov-Smirnov test statistic are that it can be exactly determined and that it does not depend on the underlying distribution if the distribution function is continuous.

Moreover, the sequence $\sqrt{N} D_N$ converges in distribution to the random variable $\sup_{t \in \mathbb{R}} |B(t)|$, where B is the so-called Brownian bridge. In particular, it holds

$$\mathbb{P}(\sqrt{N} D_N) > x) \leq 2 \exp(-2x^2) \quad (N \in \mathbb{N})$$

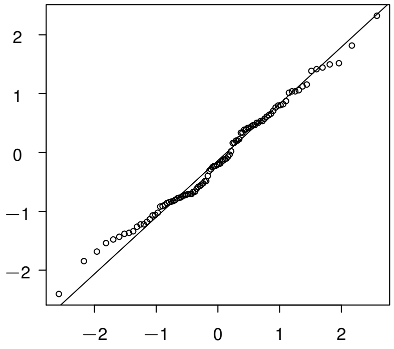


Figure 13.3: Statistical visualization of the data set: Q-Q plot for normal distribution

and

$$\lim_{N\to\infty} \mathbb{P}(\sqrt{N}D_N \leq x) = H(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2) \quad (x \in \mathbb{R}),$$

where the estimate is the so-called Dvoretsky-Kiefer-Wolfowitz inequality with the specific leading constant 2 due to Massart (1990) and the limit is due to Kolmogorov.

The goodness-of-fit test is based on the latter argument, namely, the null hypothesis is rejected at level α if

$$\sqrt{N}D_N > H^{-1}(1-\alpha).$$

The test is more sensitive in the middle of the distribution rather than at its tails.

Alternatively, the Shapiro-Wilk test has the highest power among the normal distribution tests. The test statistic used here is based on two different estimates for the variance: an estimate from the quantile-quantile plot (see below) and of the sample variance. Both estimates should be nearly identical, if the data are drawn from a normal distribution.

In the literature one may find a lot of other commonly used tests, see Hartung (2005) and Thode (2002). Applying these to the data base of the given networks, we observed that only in rare cases a unique result can be derived from the whole variety of the available tests.

Many common tests output the so-called p-value of the data $\{x_1,\ldots,x_N\}$. The p-value pval$(x_1,\ldots,x_N)$ is defined as the greatest lower bound on the set of all α such that the level α test based on $\{x_1,\ldots,x_N\}$ rejects the null hypothesis (see, e.g., Section 3.3 in Lehmann and Romano 2005). One tends to reject the null hypothesis if the p-value turns out to be less than a certain significance level, often 0.05. The smaller the p-value, the more strongly the test rejects the null hypothesis.

The automation described in Section 13.3 is exclusively based on the Kolmogorov-Smirnov test.

13.1.3 ▪ Visual checks

In addition to statistical tests, there are many visual checks to verify the distribution assumptions of individual data series. The most commonly used ones are histograms, quantile-quantile plots and box plots, see Hartung (2005) and Fahrmeir et al. (2007). In *quantile-quantile plots* (Q-Q plots) as shown in Figure 13.3, the quantiles (see Eq. (13.3) below) of the data series are plotted against the quantiles of the expected underlying distribution. In case of assuming the correct distribution, the points lie approximately on the bisector. With clear deviations from this reference line, the distribution has been specified wrongly. Strong deviations of the last points only indicate outliers. In box-plots, location and dispersion differences between two data series as well as outliers are made visible. These graphical tools are available on a case-by-case basis, but are not useful for a complete automation of the process. The visual checks served as a basis for selecting the classes of probability distributions utilized in Section 13.3.

13.1.4 ▪ Outlier detection

Another component of data analysis is the detection of outliers and specifications on how to deal with them. Outliers are individual values strongly deviating from the other data. If there are no systematic errors in the data collection, one can check by suitable tests whether a suspected outlier should be removed from or retained in the sample. A simple criterion for the identification of outliers, in particular for symmetric single-peaked distributions is provided by the so-called sigma-range. In case of a normal distribution data beyond the 2.5σ-range (covering realization of the random variable with a probability of 99%) are declared to be outliers.

For arbitrary distributions, Chebyshev's inequality yields conclusions about the portion of the data outside of the $\ell\sigma$-range. More exactly, it holds for arbitrary random variables with mean $\mathbb{E}[\xi] = \mu$ and finite variance $\mathrm{Var}(\xi) = \sigma^2$ that

$$\mathbb{P}(|\xi - \mu| < \ell\sigma) \geq 1 - \frac{1}{\ell^2} \quad \text{for all } \ell > 0.$$

Robust limits for the detection of outliers for many distribution types, especially for skewed distributions, can be derived on the basis of the quartiles and the interquartile range. The latter is defined for a random variable ξ with distribution function $F_\xi(t)$ as $\mathrm{IQR} := x_{0.75} - x_{0.25}$. Here x_p denotes the p-quantile

$$x_p := F_\xi^{-1}(p) = \inf\{t \in \mathbb{R} \mid F_\xi(t) \geq p\}. \tag{13.3}$$

There exists a variety of outlier tests, e.g., the David-Hartley-Pearson test, the Grubbs' test for normally distributed samples and average sample sizes, and the Dean-Dixon test for very small sample sizes. For further details, we refer to the relevant literature, see Hartung (2005). Removal of outliers has to be carried out very cautiously because especially for the multivariate analysis the sample size should be kept as large as possible. The automation in Section 13.3 and Section 13.4 refrains from outlier deletion.

13.2 ▪ Reference temperature and temperature intervals

As mentioned earlier, daily mean gas flows at some exit point depend on the local temperature. However, a numerically tractable model for this functional dependence is not available in general. An alternative approach consists in removing this dependence by introducing (i) a reference temperature for the H-gas and L-gas network, respectively, and (ii) a subdivision of the temperature range into sufficiently small and properly sized intervals in order to arrive at statistically relevant univariate distributions at each exit point. More precisely, we propose to proceed as follows:

1. Start with historical data for gas demands at the exits of the gas network on the one hand and for temperatures at certain preselected measuring stations on the other hand.

Calculate for each day d of the historical time period a *reference temperature* $T^{\mathrm{ref}}(d)$ by using a weighted average and the daily mean demand $D(n,d)$ for each exit point n.

2. The relevant temperature range (for the H-gas and L-gas network from −15 °C to 30 °C) is subdivided into intervals $(T^i, T^{i+1}]$, $i = 1,\ldots,I$, which are

   ▹ small enough to neglect the temperature dependence of demands within the interval and

   ▹ large enough to contain a sufficient amount of data required for statistical modeling.

   (For the H-gas and L-gas network, intervals of two degrees Celsius were selected for the interior of the temperature range and (−15 °C,−2 °C], (20 °C,30 °C] for its boundary to have reasonable amount of data available.)

3. For each temperature interval and each exit point of the network the data of gas demands for days with reference temperature belonging to the given interval are filtered:

$$S(i,n) := \left\{ D(n,d) \,\middle|\, T^{\mathrm{ref}}(d) \in (T^i, T^{i+1}] \right\} \quad \forall i \; \forall n.$$

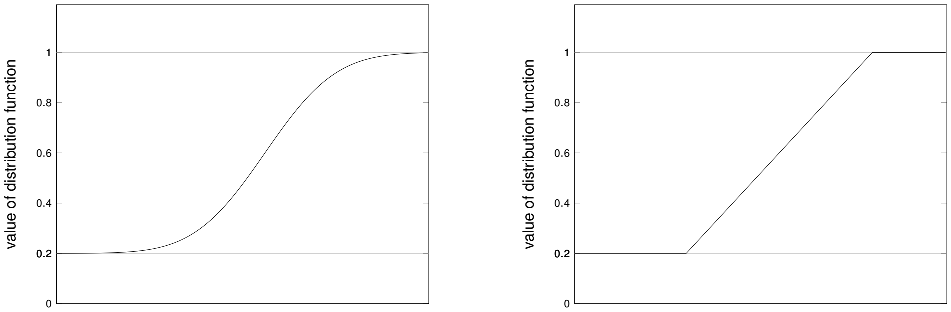
In this way the temperature dependence of gas demand is modeled for the whole gas network rather than for local parts of it. Instead of modeling this temperature dependence by an explicit formula, a subdivision into small intervals is used which allows us to establish univariate statistical models with homogeneous variance within each interval at each exit. Moreover, whenever possible, correlations in gas demand between different exit points can be taken into account for the multivariate statistical model of each temperature interval (see Section 13.4). In order to avoid confusion with respect to reference or local temperature, these intervals will be called *temperature classes* from now on. In addition to these temperature classes, the filtering procedure described above is also carried out with respect to *day classes* in order to model significantly different behavior of gas demand for specific days, namely: *working day*, *weekend* and *holiday*.

13.3 ▪ Univariate distribution fitting

The next step consists in finding a univariate statistical model for the distribution of gas demands in each class $S(i,n)$. This is a two-step procedure where in the first step an appropriate class of distributions (e.g., normal distribution) has to be found and in the second step the associated parameters (e.g., mean value, standard deviation) for this class have to be determined. For the H-gas and L-gas network, the following selection of distributions turned out to be relevant as a result of our prior descriptive data analysis (statistical tests, visual inspection): *normal distribution*, *log-normal distribution*, *uniform distribution*, *Dirac distribution*.

The need to incorporate the Dirac distribution arises mainly from frequent observations of zero demands as mentioned above, but occasionally also of constant positive demands. Visualization of the data suggested the presence of mixtures of these distribution classes, e.g., Dirac and normal distribution. We then speak of *shifted* normal, uniform etc. distributions. Therefore we extend our statistical model to positive combinations of the distribution classes above, such that the determination of weights in this combination is also part of the modeling process.

To provide an example, consider the case of the shifted normal distribution: Assume that the relative frequency of zero loads in the data sets equals $w \in [0,1]$, whereas the remaining (positive) elements of the data set follow a normal distribution with mean μ and standard deviation σ. Then, the overall distribution function F for the data set would



(a) Shifted normal distribution

(b) Shifted uniform distribution

Figure 13.4: Illustration of shifted distribution functions (see Eq. (13.4) in case of normal distribution), where in this example the shift is given by $w = 0.2$.

be modeled as:

$$F = wF_{\delta} + (1 - w)F_{\mathcal{N}(\mu,\sigma)}, \tag{13.4}$$

where

$$F_{\delta}(t) = \begin{cases} 0, & \text{if } t < 0, \\ 1, & \text{it } t \geq 0, \end{cases}$$

denotes the distribution function of the Dirac distribution at zero and

$$F_{\mathcal{N}(\mu,\sigma)}(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \mathrm{d}x$$

denotes the distribution function of the normal distribution $\mathcal{N}(\mu,\sigma)$. Figure 13.4 (a) illustrates the distribution function of such shifted normal distribution, where the shift with respect to a pure normal distribution function is reflected by the intercept on the y-axis. Similar shifted forms are considered for the other distributions, e.g., shifted uniform distribution on some interval see Figure 13.4 (b).

We aim at an automated procedure for statistical modeling. This means that both the assignment of a specific distribution class (e.g., shifted normal) to a data set $S(i,n)$ and the parameter estimation for this class (e.g., w, μ, σ) have to be carried out by a numerical procedure. As an assignment criterion one may use the Kolmogorov distance Eq. (13.1) which is justified, for instance, by the Kolmogorov-Smirnov test discussed above, but also by stability results in stochastic optimization (see Römisch 2003).

More precisely, we assign to a given data set $S(i,n)$ that distribution from our portfolio which realizes the smallest Kolmogorov distance to the empirical distribution function in our data set. This idea is illustrated in Figure 13.5: The points $\{x_1,\ldots,x_N\}$ distributed along the x-axis represent the measurement data $S(i,n)$ for some exit point n and temperature class i. These points induce an empirical distribution function as defined in Eq. (13.2), see the staircase function in Figure 13.5. Assume that we have two candidate distributions which we want to assign to the given data set, e.g., normal distribution or uniform distribution. Then our choice will be made according to the candidate realizing a smaller Kolmogorov distance to the empirical distribution function (see vertical bars in Figure 13.5, indicating the positions at which the largest deviation of the corresponding distribution functions occur).

The numerical computation of the Kolmogorov distance between a distribution with distribution function F and the empirical distribution function S_N associated with the

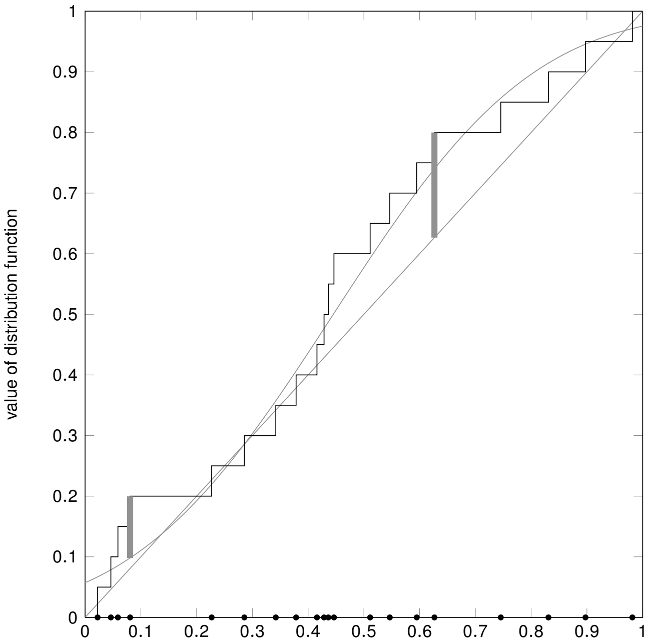


Figure 13.5: Kolmogorov distance between the empirical distribution and the best fitting normal and uniform distribution, respectively. The empirical distribution function (staircase function) is induced by 20 sample points. Due to smaller Kolmogorov distance, in this example preference would be given to the normal distribution.

measurement data $\{x_1, \ldots, x_N\}$ can be carried out according to the simple expression

$$\max_{i=1,\ldots,N} \max \left\{ \left| F(x_i) - \frac{1}{N} \#\{x_j \mid x_j \le x_i\} \right|, \left| F(x_i) - \frac{1}{N} \#\{x_j \mid x_j < x_i\} \right| \right\}. \tag{13.5}$$

For the computation, it has to be taken into account that the candidate distributions themselves depend on parameters (e.g., mean and standard deviation for the normal distribution or interval limits for uniform distribution). One possibility to identify these parameters would rely on estimating them from the given measurement data (e.g., arithmetic mean and empirical standard deviation for the normal distribution and minimum and maximum measurements for the interval limits of the uniform distribution). It appears, however, more natural to subordinate the parameter identification to the same criterion of minimum Kolmogorov distance as discussed above. More precisely, we are led to solve the following optimization problem:

$$\begin{aligned} \min_{p} \max_{i=1,\ldots,N} \max \Bigg\{ & \left| F(p, x_i) - \frac{1}{N} \#\{x_j \mid x_j \le x_i\} \right|, \\ & \left| F(p, x_i) - \frac{1}{N} \#\{x_j \mid x_j < x_i\} \right| \Bigg\}. \end{aligned} \tag{13.6}$$

Here, $F(p, \cdot)$ refers to the distribution function of the specific class in the portfolio depending on the parameter vector p (e.g., $p = (\mu, \sigma)$ for the normal distribution). According to Eq. (13.5), the minimization problem above identifies the parameter vector p in a way that the best fit to the empirical distribution function is realized within the given class. Similar optimization problems are solved for other distribution classes (uniform, log-normal,

Table 13.2: Percentage of univariate distributions for the H-gas network for all temperature classes: normal (ND), shifted normal (shND), lognormal (logND), shifted lognormal (shLogND), Dirac distributed (Dirac), uniform (UD), and shifted uniform (shUD)

Temp. Class	ND	shND	logND	shLogND	Dirac	UD	shUD
(−15 °C, −2 °C)	40.40	8.75	34.01	5.39	8.08	1.68	1.35
(−2 °C, 0 °C)	40.40	8.08	31.31	7.07	8.75	3.03	1.01
(0 °C, 2 °C)	41.41	7.74	30.30	9.76	7.41	0.67	2.69
(2 °C, 4 °C)	42.42	13.13	22.56	11.78	6.73	0.67	2.02
(4 °C, 6 °C)	41.75	12.12	21.89	14.14	7.07	0.34	2.69
(6 °C, 8 °C)	41.08	15.49	18.86	12.46	7.07	2.02	3.03
(8 °C, 10 °C)	38.72	23.57	15.15	10.10	7.07	1.35	3.70
(10 °C, 12 °C)	32.32	20.20	20.88	17.17	6.73	0.34	1.68
(12 °C, 14 °C)	26.60	17.51	24.58	19.53	7.41	0.00	4.04
(14 °C, 16 °C)	21.55	21.55	27.61	15.49	10.10	0.67	3.03
(16 °C, 18 °C)	23.23	20.20	26.26	17.85	9.09	0.00	3.37
(18 °C, 20 °C)	24.58	17.17	23.91	21.55	10.44	0.00	1.68
(20 °C, 30 °C)	24.92	17.17	25.25	15.82	14.48	1.01	1.01

shifted versions etc.). The choice of the final statistical model is then made according to the class realizing the smallest of these optimal values along with the associated parameter vector p. Evidently, Eq. (13.6) is equivalent to the following nonlinear optimization problem:

$$\begin{aligned}
\min_{t,p} \quad & t \\
& F(p,x_i) - \frac{1}{N}\#\{x_j \mid x_j \le x_i\} \le t && \text{for all } i = 1,\dots,N, \\
& \frac{1}{N}\#\{x_j \mid x_j \le x_i\} - F(p,x_i) \le t && \text{for all } i = 1,\dots,N, \\
& F(p,x_i) - \frac{1}{N}\#\{x_j \mid x_j < x_i\} \le t && \text{for all } i = 1,\dots,N, \\
& \frac{1}{N}\#\{x_j \mid x_j < x_i\} - F(p,x_i) \le t && \text{for all } i = 1,\dots,N.
\end{aligned}$$

As starting values for the unknown parameter vector p in this problem one may use the classic empirical estimates based on the measurement data. The solution of these optimization problems for all exits of the H-gas and the L-gas network leads to the results compiled in Table 13.2 and Table 13.3. In case the true distribution type is approximately normal, the classical estimator (based on estimating mean and variance) and the optimized estimators (based on solving Eq. (13.6)) are almost identical.

13.4 ▪ Multivariate distribution fitting

So far, our statistical models describe the statistical behavior of individual exit points of the network. To capture the correlations and other relationships among the exits, the underlying multivariate distributions need to be estimated.

Estimating a multivariate distribution is closely related to an analysis of correlations between the single univariate distributions. In the context of gas networks, correlations reflect tendencies of similar or opposite behavior in gas consumption between certain

Table 13.3: Percentage of univariate distributions for the L-gas network for all temperature classes: normal (ND), shifted normal (shND), lognormal (logND), shifted lognormal (shLogND), Dirac distributed (Dirac), uniform (UD), and shifted uniform (shUD)

Temp. Class	ND	shND	logND	shLogND	Dirac	UD	shUD
(−15 °C, −2 °C)	58.01	2.55	35.32	3.97	2.70	3.26	0.43
(−2 °C, 0 °C)	51.49	3.97	40.85	3.12	2.84	3.12	0.85
(0 °C, 2 °C)	48.09	4.96	40.85	6.24	2.98	2.41	0.71
(2 °C, 4 °C)	60.00	9.36	24.82	7.38	2.41	0.85	1.28
(4 °C, 6 °C)	57.16	11.49	25.39	7.80	2.55	1.42	0.43
(6 °C, 8 °C)	55.04	18.72	20.43	8.37	2.27	0.71	0.71
(8 °C, 10 °C)	49.79	18.72	20.71	10.21	2.84	2.27	1.56
(10 °C, 12 °C)	41.70	21.84	23.69	12.77	2.98	0.99	2.27
(12 °C, 14 °C)	27.66	20.57	32.77	19.86	2.55	0.28	2.55
(14 °C, 16 °C)	27.38	18.01	32.06	22.70	3.12	0.28	2.70
(16 °C, 18 °C)	29.65	17.73	29.50	20.43	5.25	0.57	2.70
(18 °C, 20 °C)	33.48	15.74	25.67	22.27	6.10	0.57	2.13
(20 °C, 30 °C)	32.48	14.61	26.52	17.87	10.78	1.42	2.41

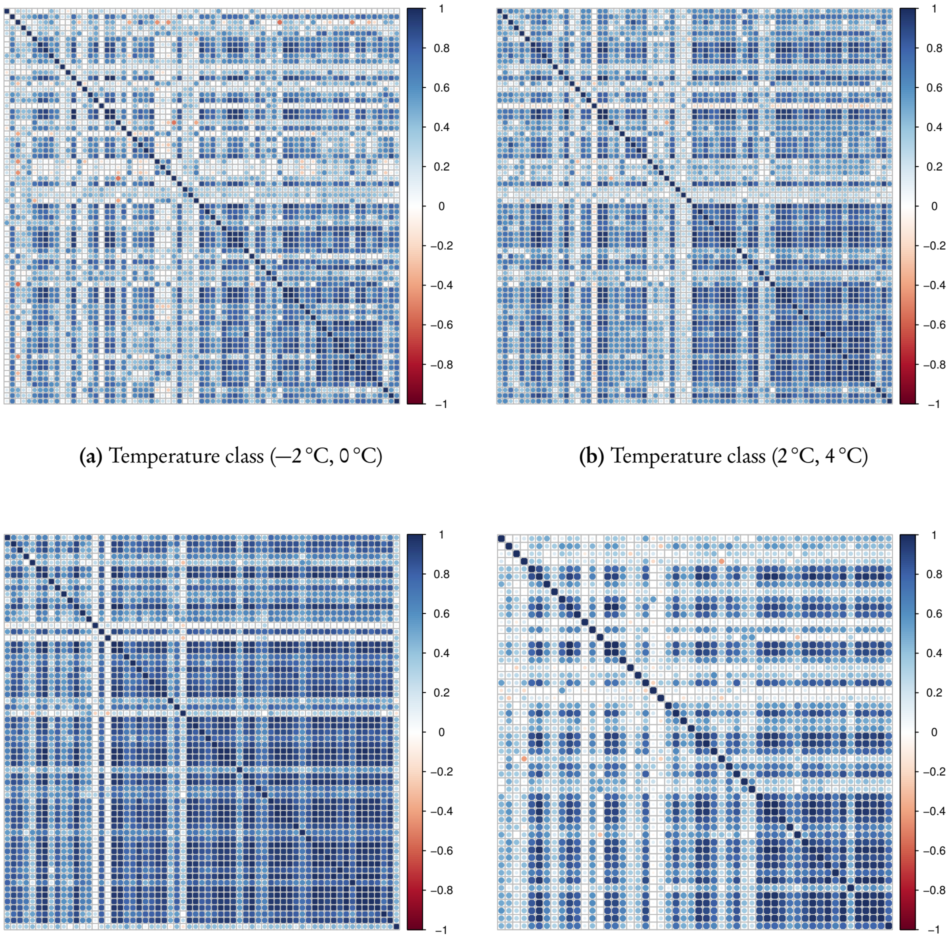
groups of exit points. To give a simplified idea, for instance, households could exhibit a common behavior in gas demands while this may be uncorrelated with the behavior of industrial clients. Figure 13.6 plots the pairwise correlations between gas demands of exits from a certain area of the L-gas network. In the respective plots, one recognizes certain substructures, namely groups of exits being relatively strongly correlated within the group but only weakly correlated with exits from different groups. It is also revealed that the correlation pattern is a function of temperature. In the L-gas network, negatively correlated exits occur more frequently than in the H-gas network (not plotted here). Presumably, this is due to functional or contractual relationships reflected in the data.

As in the univariate case, a multivariate distribution is uniquely characterized by its density (if it exists) or, more generally, by its distribution function. Determining such joint (network-related) distribution on the basis of the individual (exit-wise) distributions may be very hard or even impossible in a situation where the latter are of very different character: recall from the distribution types discussed in the previous section that we are faced not only with continuous but also with discrete or shifted distributions for which it is not evident how they would fit to a joint multivariate distribution. On the other hand, for exits obeying a univariate normal distribution, it appears natural to group them in order to establish a joint multivariate distribution. The latter is characterized by the density

$$f(x) = (2\pi)^{-k/2} (\det \Sigma)^{-1/2} \exp\left[-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right], \tag{13.7}$$

where k refers to the dimension (number of exit points), μ is the mean vector, and Σ the covariance matrix. The parameters μ and Σ can be estimated from the exit-wise mean values and variances as well as from the correlations between exits. The same procedure can be applied to exits with a lognormal distribution by simple transformations.

Summarizing, regarding gas demand data one is faced with the situation that a subset of exit points is appropriate for establishing a partial multivariate model, whereas other exits (exhibiting Dirac, uniform, or shifted distributions) are difficult to join with the previous ones to set up a total multivariate model. Therefore, a simplified approach providing such total multivariate model would consist in assuming these remaining exit points to



(a) Temperature class (—2 °C, 0 °C)

(b) Temperature class (2 °C, 4 °C)

(c) Temperature class (8 °C, 10 °C)

(d) Temperature class (18 °C, 20 °C)

Figure 13.6: Correlation plots for different temperature classes in an area of the L-gas network

have independent distributions with all other exits. This allows to calculate the joint distribution of the total model as the product of the multivariate normal distribution with density Eq. (13.7) obtained from exit points with univariate normal or lognormal distributions on the one hand and all the univariate distributions from the remaining exits on the other hand.

A reliable estimation of the multivariate normal distribution with density Eq. (13.7) requires a sufficient amount of data. Given k exit points and a sample size N, it is known that $N \geq k$ should hold, since in case of $N < k$ the sample covariance matrix may be singular and a regularized estimator is needed. It is also known that $N = O(k)$ samples suffice (see Vershynin 2012). As a rule of thumb, the number of samples in each exit point should exceed $3k$ to provide a stable estimate of the covariance matrix Σ. This requirement turned out not to be satisfied for the subgroup of normal or lognormal exits in the network due to the limited amount of data. As a remedy, the correlation structure as it became visible in Figure 13.6 can be exploited in order to find subgroups of these exits such that the data requirement is fulfilled for these smaller subgroups and such that the distribution between different subgroups is almost independent (indicated by small correlation between exit points of different subgroups). In this way, similar to the total model discussed before, the partial multivariate normal model is obtained as a product of smaller multivariate normal distributions each of which is estimated in a sufficiently stable manner. The analysis of block structures in the correlation matrix can be automated by means of p-median greedy heuristics based on the distance $1-\rho_{i,j}^2$ between exit points i and j, where $\rho_{i,j}$ refers to the correlation coefficient between these exits. The greedy heuristics are also used for scenario reduction in Section 14.2.2. Evidently, the larger the correlation (positive or negative), the smaller the distance between these exits making it more likely that they are gathered in a common group.

13.5 ▪ Forecasting gas flow demand for low temperatures

In the following we study historical data of exit loads of the considered gas networks, in order to make a reliable and realistic prediction of the future exit loads for low temperatures. In this particular case, the prediction of gas consumption at the exit points is extremely important, since it is usually very high in cold seasons. We utilize parametric as well as semi-parametric non-linear logistic regression models to estimate the gas flow in dependence of the temperature. The relationship between load flow and temperature is closely related to empirical models for growth data, which are frequently employed in natural and environmental sciences, and sometimes in social sciences and economics. Some examples of such models and their applications can be found in Jones, Leung, and Robertson (2009), Vitezica et al. (2010), and Jarrow, Ruppert, and Yu (2004).

The articles (Hellwig 2003; Geiger and Hellwig 2002; Wagner and Geiger 2005) suggest the use of sigmoidal growth models for the description of typical gas load profiles in the energy sector. An overview of methods useful for understanding the complexity of gas transportation relying on the mentioned parametric 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 a time variable x. The general assumption is, apart from the underlying random fluctuation that 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 using the data at hand. The same type of models occurs when the explanatory variable x is not the time, but the increasing intensity of some other factor. We observe a change (in general a reduction) of gas consumption with increased 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 (see Wegman and Wright 1983; Eilers and Marx 2010; Eilers and Marx 1996). This approach emphasizes a modeling based on smooth regression, where the penalty controls the amount of smoothing. We follow in particular a similar approach as proposed in (Bollaerts, Eilers, and Mechen 2006), where the authors incorporate shape constraints into the P-splines model.

Typical exit points in the considered gas networks are public utilities, industrial consumers, and storages, as well as exit points on national borders and market crossings. An important aspect to consider in the forecast is the so-called *design temperature*. The design temperature is defined as the lowest temperature at which the network 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° and −16° Celsius. Such low mean daily temperatures are very uncommon in Germany, and we rarely encounter load flow data available at the design temperature. For this reason, prediction is usually the only way to estimate gas loads at the design temperature of the network, and we investigate several possible models suitable for the forecast.

For the aim of forecasting, we search for a fitting curve that can be interpreted as a curve of expected values in dependence of temperature. The expected values estimated in this way can be used in Chapter 14 as input parameters necessary for sampling gas load under particular distribution for low temperature intervals where non-sufficient data are available. The type of distribution assigned to these low temperature intervals containing almost no data is usually taken to coincide with the one of the neighboring temperature intervals. The neighboring interval is chosen in such a way that its estimated distribution is sufficiently reliable and stable. Thus, these temperature intervals corresponding to the exceptionally low temperatures share the same distributional properties, i.e., deviation and multivariate information as variance and correlation w.r.t. the mentioned nearest warmer temperature interval, but differ in their expected value resulting from the fitted curve. We refer to Chapter 14 for further discussion on this issue.

The data sets considered for fitting a curve of expected values in gas networks are usually large, including in some cases many periods (years) of data. Therefore for the aim of fitting a curve it is sometimes recommendable to cluster the data in convenient subintervals of temperatures and replace the data within a subinterval through its empirical average. This is done in order to reduce the size of the data set and therefore the dimensionality of the underlying optimization problems, in order to avoid very ill conditioned numerical problems. This strategy also allows the elimination of undesired outliers, by considering only subintervals where the sample size is big enough to be *significant*. To this end, we follow a criteria based on the Central Limit Theorem and define a sample size N on a subinterval to be significant if $N \geq 16\hat{\sigma}^2/(U\hat{\mu})^2$, $N \geq 2$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation from the i.i.d. sample, and U is the number of deviation units measured in terms of $|\hat{\mu}|$. The latter condition means that we ask for the sample size N to be big enough such that the width of a 95 % confidence interval for the standard error of $\hat{\mu}$ can be covered by fixed U units of $|\hat{\mu}|$.

13.5.1 ▪ Parametric models

In the case of parametric models, the basic assumption is that the growth curve belongs to a well known parametric family of curves. The physical interpretability of parameters in the model usually motivates the choice of the growth curve. Based on agreements between the network operators (see [KoV]), we take the following sigmoidal growth model to describe the dependence of gas consumption on temperature:

$$y_i = S(t_i|\theta) + \varepsilon_i\,.$$

Here y_i denotes the standardized daily mean gas flow, and the corresponding expected value (or mean) curve parameterized in $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ is given by

$$S(t_i|\theta) = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\dfrac{\theta_2}{t_i - 40}\right)^{\theta_3}}\,. \tag{13.8}$$

The curve depends on the predictor t_i, which stands for the weighted four-day mean temperature with weights given in the following form

$$t_i = \sum_{j=0}^{3} w_j t_{ij}, \quad w_0 = \frac{8}{15},\ w_1 = \frac{4}{15},\ w_2 = \frac{2}{15},\ w_3 = \frac{1}{15}, \tag{13.9}$$

where t_{i0}, t_{i1}, t_{i2}, and t_{i3} are the temperatures corresponding to the days $i, i-1, i-2, i-3$. Finally, ε_i is an error term reflecting zero mean and constant variance.

The articles (Geiger and Hellwig 2002; Cerbe 2008) introduce this kind of models for description of typical gas loads in dependence of temperature. According to the description of the log-logistic model provided by Ritz and Streibig (2008), the parameters θ_1 and θ_4 in Eq. (13.8) represent upper and lower horizontal asymptotes on the curve, and the other two parameters describe the shape of the decrease of the (logistic like) curve. 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. Parameter 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, and θ_3 measures flexibly the dependence in the heating period.

The authors in (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 Eq. (13.9) is suggested. The weights given by Eq. (13.9) are obtained from the standardized geometric series with basis 2 applied to the temperature of the last four days, i.e.,

$$t_i = \frac{t_{i0} + \frac{t_{i1}}{2} + \frac{t_{i2}}{4} + \frac{t_{i3}}{8}}{1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \sum_{j=0}^{3} w_j t_{ij}\,.$$

Based on these facts, German gas transportation companies agreed to use the *sigmoidal function* $S(t_i|\theta)$ defined in Eq. (13.8) with the explanatory variable t_i given by Eq. (13.9) to describe the dependence of gas loads on temperature and to forecast the gas consumption at the design temperature.

Several generalizations of the basic sigmoid model have been considered in order to improve the forecasting of gas loads (see Friedl, Mirkov, and Steinkamp 2012; Leövey et al. 2011). The sigmoid model and its generalizations so far represent a rough characterization of the mean gas load. The resulting forecast usually underestimates the mean responses for low temperatures. An alternative approach based on semi-parametric models has been also considered in order to achieve a more accurate forecast for low temperatures, and will be described in the following section.

13.5.2 ▪ Semi-parametric models

The nuances missed by the sigmoidal models, as well as the numerical difficulties that arise in the resulting non-convex nonlinear optimization problems, motivates the search for alternative and, in some sense, simpler models that overcome these difficulties. One possibility is to use semi-parametric models, such as locally weighted regression (Cleveland 1979) or spline models. Unfortunately, locally weighted regression models are not suitable for prediction. Many authors propose some variant of spline regression for this kind of problems, see, e.g., Jones, Leung, and Robertson (2009), Vitezica et al. (2010), Jarrow, Ruppert, and Yu (2004), Mackenzie, Donovan, and McArdle (2005), Cadorso-Suárez et al. (2010), and Riedel and Imre (1993).

In our case, we choose the penalized splines (P-splines) approach, based on Wegman and Wright (1983), Eilers and Marx (1996), Eilers and Marx (2010), and Bollaerts, Eilers, and Mechen (2006). This choice is motivated by its simplicity and flexibility. The optimization problems resulting from the fitting problem are convex and solutions can be obtained through solving a system of linear equations. The advantage of P-splines over B-splines is the easy control of smoothness as well as the simple way to handle spline knots, i.e., their number and their positions. As emphasized by Jarrow, Ruppert, and Yu (2004), another advantage of the P-splines method is that knots can be chosen automatically. The number of knots should be sufficiently large to accommodate the non-linearity of the underlying data. Additional shape constraints lend even more flexibility to the model, since the shape of the curve on boundaries can be adjusted if necessary.

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

$$y_i = S_\Delta(t_i) + \varepsilon_i,$$

where y_i is usually taken to be a daily mean gas flow at a particular exit point of the network, and t_i stands for the weighted four-day mean temperature, with the weights w described in Eq. (13.9). The function $S_\Delta(t)$ is given by the linear combination of basis spline functions $B_j, j = 1, \ldots, m$, on the mesh Δ, given by

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

and ε_i, $i = 1, \ldots, n$, are random noise terms reflecting zero mean and constant variance. The functions B_j are basis functions of the B-spline of degree q. The mesh Δ results by taking first an equidistant grid with $m - q$ segments extended over the set of data that includes only subintervals with significant sample size, and then we modify the grid by replacing the least and maximal knots with knots at the design temperature t_{min} and at the maximal temperature t_{max} correspondingly. The resulting mesh Δ contains $m - q - 1$ inner knots.

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

$$\sum_{i=1}^{n} (y_i - S_\Delta(t_i))^2,$$

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

$$\sum_{i=1}^{n} (y_i - S_\Delta(t_i))^2 + \lambda \sum_{j=k+1}^{m} (\delta^k(a_j))^2,$$

where $\delta^k(a_j)$ denotes the k-th order finite differences of the coefficients a_j of the corresponding B-splines. For the work described here, we use in particular the second order finite difference $\delta^2(a_j) := a_j - 2a_{j-1} + a_{j-2}$, $j = 3, \ldots, m$.

The first order shape constraints can be added to the P-splines approach, following Bollaerts, Eilers, and Mechen (2006) or by imposing linear constraints on the spline coefficients a_j and solving the resulting constrained least-squares optimization problem.

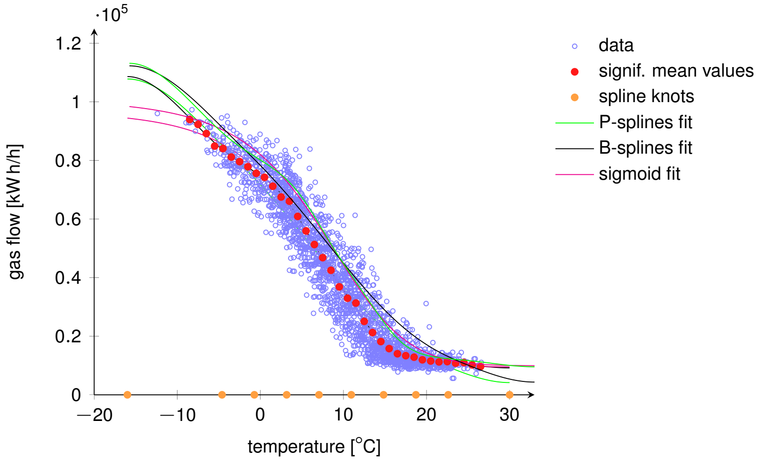


Figure 13.7: Shown are the P-splines, B-splines ($\lambda = 0$) and sigmoid fitting mean curves with 9 cubic splines based on significant mean values on 1 °C width subintervals, with $U = 0.4$, $t_{min} = -16$, $t_{max} = 30$.

Several available commercial optimization solvers can be used for this purpose. In our approach, we include first-order boundary derivative constraints

$$\frac{\partial S_\Delta}{\partial t}(t_{min}) = \frac{\partial S_\Delta}{\partial t}(t_{max}) = 0,$$

aiming to simulate a constant consumption beyond the design temperature t_{min} and the considered maximal temperature t_{max}. Positivity constraints where also added to ensure that the fitted mean curve remains positive. Different criteria for a robust choice of the smoothing parameter λ can be founded in Lee and Cox (2010). We adopted the strategy called absolute cross validation (ACV) by Lee and Cox (2010, Section 3.2), which yields very good results for the forecasting of the mean curve in the considered gas networks.

Based on data from a typical statistical exit point of the gas network, Figure 13.7 compares P-splines, B-splines ($\lambda = 0$), and sigmoid fitting mean curves.