17

Statistical model of segment-specific relationship between natural gas consumption and temperature in daily and hourly resolution

Marek Brabec, Marek Malý, Emil Pelikán and Ondřej Konár
Department of Nonlinear Modeling, Institute of Computer Science, Academy of Sciences of the Czech Republic
Czech Republic

1. Introduction

In this chapter, we will describe a statistical model which was developed from first principles and from empirical behavior of the real data to characterize the relationship between the consumption of natural gas and temperature in several segments of a typical gas utility company's customer pool. Specifically, we will deal with household and small+medium (HOU+SMC) size commercial customers. For several reasons, consumption modeling is both challenging and important here. The essential fact is that these segments are quite numerous in terms of customer numbers. It leads to three practically significant consequences.

- First, their aggregated consumption constitutes an important part of the total gas consumption for a particular day.
- Secondly, their consumption depends strongly on the ambient temperature. Hence, the temperature lends itself as a nice and cheap-to-obtain, exogeneous predictor. The temperature response is nonlinear and quite complex, however. Traditional, simplistic approaches to its extraction are not adequate for many practical purposes.
- Further, the number of customers is high, so that their individual follow-up in fine time resolution (say daily) is not feasible from financial and other points of view. Routinely, their individual data are available only at a very coarse (time-aggregated) level, typically in the form of approximately annual consumption totals obtained from more or less regular meter readings. When daily consumption is of interest, the available observations need to be disaggregated somehow, however.

Disaggregation is necessary for various practical purposes – for instance for the routine distribution network balancing, for billing computations related to the natural gas price changes (leading to the need for pre- and post-change consumption part estimates), etc. As required by the market regulator, the resulting estimates need to be as precise as possible, and hence they need to use available information effectively and correctly. Therefore, they should be based on a good, formalized model of the gas consumption. Since the main driver of the natural consumption is temperature, any useful model should reflect the consumption response to temperature as closely as possible. It ought to follow basic qualitative features of the relationship (consumption is a decreasing function of temperature having both lower and upper asymptotes), but it needs to incorporate also much finer details of the relationship observed in empirical data.

Our model tries to achieve just this and a bit more, as we will describe in the following paragraphs. It is based on our analyses of rather large amounts of real consumption data of unique quality (namely of fine time resolution) that was obtained during several projects our team was involved in during the last several years. These include the Gamma project, Standardized load profiles (SLP) projects in both the Czech Republic and Slovakia, as well as the Elvira project (Elvira, 2010). Consumption-to-temperature relationships were analyzed there in order to be able to model/describe them in a practically usable way.

Our resulting model is built in a stratified way, where the strata had been defined previously via formal clustering of the consumption dynamics profiles (Brabec at al., 2009). The stratification concerns the values of model parameters only, however. The form of the model is kept the same in all strata, both in order to retain simplicity advantageous for practical implementation and for saving the possibility of a relatively easy (dynamic) model calibration (Brabec et al., 2009a). Model parameters are estimated from data in a formalized way (based on statistical theory). The data consist of a sample of consumption trajectories obtained through individualized measurements (obtained in rare and costly measurement campaigns for nationwide studies mentioned above).

Construction of the model keeps the same philosophy as our previous models that have been in practical use in Czech and Slovak gas utility companies (Brabec et al., 2009), (Vondráček et al., 2008). It is modular, stressing physical interpretation of its components. This is useful both for practical purposes (e.g. the ability to estimate certain latent quantities that are not accessible to direct measurement but might be of practical interest) and for model criticism and improvement (good serviceability of the model).

The model we present here is substantially different from the standardized load profile (SLP) model we published previously (Brabec et al., 2009) and from other gas consumption models (Vondráček et al., 2008) in that it has no standard-consumption (or consumption under standard conditions) part. It is advantageous that the model is more responsible to the temperature changes, especially in years whose temperature dynamics is far from being "standard" and in transition (spring and fall) periods even during close-to-normal years. Absence of the smooth standard-consumption part also simplifies the interpretation of various model parts. It calls for expansion of the temperature response function. Here, we start from the approach (Brabec at al., 2008), but we expand it substantially in three important ways:

- Shape of the temperature response is estimated in a flexible, nonparametric way (so that we let the empirical data to speak for themselves, without presupposing any a priori parametric shape).
- Dynamic character of the temperature response and mainly its lag structure is captured in much more detail.

- The model now allows for temperature*(type of the day) interaction. In plain words, this means that is allows for different temperature responses for different day of week.

Numerous papers have discussed various aspects of modeling, estimation and prediction of natural gas consumption for various groups of customers such as residential, commercial, and industrial. Similar tasks are solved in the context of electricity load. Load profiles are typically constructed using a detailed measurements of a sample of customers from each group. Other, methods include dynamic modeling (historical load data are related to an external factor such as temperature) or proxy days (a day in history is selected which closely matches the day being estimated). The optimal profiling method should be chosen based on cost, accuracy and predictability (Bailey, 2000). Close association between gas demand and outdoor temperature has been recognized long time ago, so the first approaches to modeling were typically based on regression models with temperature as the most important regressor. Among such models, nonlinear regression approaches to gas consumption modeling prevail (Potocnik, 2007). The concept of heating degree days is sometimes used to suppress the temperature dependency during the days when no heating is needed (Gil & Deferrari, 2004).

In addition to the temperature, weather variables like sunshine length or wind speed are studied as potential predictors. Among other important explanatory variables mentioned in the literature one can find calendar effects, seasonal effects, dwelling characteristic, site altitude, client type (residential or commercial customer), or character of natural gas end-use. Economical, social and behavioral aspects influence the energy consumption, as well. Data on many relevant potential predictors are not available. Regression and econometric models may include ARMA terms to capture the effects of latent and time-varying variables. Another large group of models is based on the classical time series approach, especially on Box-Jenkins methodology (Lyness, 1984), or on complex time series modifications.

In the following, we will first describe the model construction in a formalized and general way, having in mind its practical implementation, however. Then, we will illustrate its performance on real data.

2. Model description and estimation of its parameters

2.1 Segmentation

As mentioned in the Introduction already, we will deal here only with customers from the household and small+medium size commercial segments (HOU+SMC). The segmentation is considered as a prerequisite to the statistical modeling which will be stratified on the segments. In the gas industry (at least in the Czech Republic and Slovakia), the tariffs are not related to the character of the consumption dynamics, unlike in the (from this point of view, more fortunate) electricity distribution (Liedermann, 2006). Therefore, the segmentation has to be based on empirical data. In order to be practical, it has to be based on time-invariant characteristics of customers which are easily obtainable from routine gas utility company databases. These include character of customer (HOU or SMC), character of the consumption (space heating, cooking, hot water or their combinations; technological usage).

Here, we used hierarchical agglomerative clustering (Johnson & Wichern, 1988) of weekly standardized consumption means averaged across customers having the same values of selected time-invariant characteristics. Then, upon expert review of the resulting clusters, we used them as segments, similarly as in (Vondráček et al., 2008). This way, we have $K = 8$ segments (4 HOU + 4 SMC in the Czech Republic and 2 HOU + 6 SMC in Slovakia).

2.2 Statistical model of consumption in daily resolution

Here we will formulate a fully specified statistical model describing natural gas consumption Y_{ikt} of a particular (say the i -th, $i = 1,\ldots,n_k$) customer of the k -th segment ($k = 1,\ldots,K$) on during the day $t = 1,2,\ldots$ (using julian date starting at a convenient point in the past). In fact, in order to deal with occasional zero consumptions (that would produce mathematically troublesome results in the development later), we define Y_{ikt} as the consumption plus a small constant (we used 0.005 m^3 when consumption was measured in m^3/100). Another, more complicated possibility is to model zero consumption process more explicitly is described in (Brabec et al., 2008).

We stress that the model is built from down to top (from individual customers) and it is intended to work for large regions, or even on a national level. It has been implemented in the Czech Republic and Slovakia separately. They are of the same form but they have different parameters, reflecting differences in consumption, gas distribution, measurement etc. Then we have:

$$Y_{ikt} = p_{ik}.f_{kt} + \varepsilon_{ikt} = p_{ik}.\exp\left(\sum_{j=1}^{5}\alpha_{jk}.I_{t\in D_j} + \chi_k.I_{t\in Christmas} + \phi_k.I_{t\in Easter} + \zeta_{kt}\right) + \varepsilon_{ikt} \quad (1)$$

where $I_{condition}$ is an indicator function. It assumes value of 1 when the *condition* in its argument is true and 0 otherwise. The model (1) has several unknown parameters (that will have to be estimated from training data somehow).

We will now explain their meaning. α_{jk} is the effect of the j -th type of the day ($j = 1,\ldots,5$). Note that different segments have different day type effects (because of the subscripting by k). The notation is similar to the so called textbook parametrization often used in the ANOVA and general linear models' context (Graybill, 1976; Searle, 1971). We haste to add that, for numerical stability, the model is actually fitted in the so called sum-to-zero (or contr.sum) parametrization

$$\mu' = \sum_{j=1}^{5}\alpha_{jk},\ \alpha'_{jk} = \alpha_{jk} - \sum_{j=1}^{5}\alpha_{jk},\ j = 1,\ldots,5 \quad (2)$$

(Rawlings, 1988). In other words, we reparametrize the model (1) to the sum-to-zero for numerical computations and then we reparametrize the results back to the textbook parametrization for convenience. Table 1 shows how different types of the day $D_1,\ldots,D_5$ are defined by specifying for which particular triplet ($t-1,t,t+1$) a particular day type holds. Non- working days are the weekends and (generic) bank holidays of any kind. On the other hand, χ_k and ϕ_k are effects of special Christmas and Easter holidays. Note that these effects act on the top of the generic holiday effect, so that the total holiday effect e.g. for 25th of December is (on the log scale) the sum of generic holiday (given by the day type 4, from Table 1) and Christmas effects. *Christmas* period is (in the Central European implementations of the model) defined to consist of days of December, 23, 24, 25, 26, while *Easter* period is defined to consist form the Wednesday, Thursday, Friday, Saturday of the week before the Easter Monday. ζ_{kt} is the temperature correction which is the most important part of the model with quite rich internal structure that we will explain in detail in the next section. p_{ik} is a multiple of the so called expected annual consumption (scaled as a daily consumption average) for the i-th customer. It is estimated from past consumption record (typically 3 calendar years) of the particular customer. For instance, if we have m roughly annual consumption readings $Y_{ik,\Im_{i1}},\ldots,Y_{ik,\Im_{im}}$ in the intervals $\Im_{i1} = [t_{i1}, t_{i2}], \ldots, \Im_{im} = [t_{i,2m-1}, t_{i,2m}]$, we compute

$$\hat{p}_{ik} = \frac{Y_{ik,\Im_{i1}} + \ldots + Y_{ik,\Im_{im}}}{t_{im} - t_{i1} + 1} \tag{3}$$

and then condition on that estimate (i.e., we take the $\hat{p}_{ik}$ for the unknown p_{ik}) in all the development that follows. That way, we buy considerable computational simplicity, compared to the correct estimation based on nonlinear mixed effects model style estimation (Davidian & Giltinan, 1995; Pinheiro & Bates, 2000) at the expense of neglecting some (relatively minor) part of the variability in the consumption estimates. It is important, however that the integration period for the $\hat{p}_{ik}$ estimation is long enough.

Note that (1) immediately implies a particular separation

$$\mu_{ikt} \equiv p_{ik}.f_{kt} \tag{4}$$

of substantial practical importance. In fact, (4) achieves multiplicative separation of the individual-specific but time-invariant and common across individuals but time-varying terms. Obviously, the separation is additive on the log scale.

ε_{ikt} is an additive random error term (independent across i,k,t) which describes variability of individual customers around a central tendency of the consumption dynamics. In accord with the heteroscedasticity of the consumptions observed in practice, we assume that $\varepsilon_{ikt} \sim N(0, \sigma_k^2.\mu_{ikt})$, i.e. that the error is distributed as a normal (or Gaussian) random variable with zero expected value and variance $\sigma_k^2.\mu_{ikt}$ (which means that variance to mean ratio is allowed to differ across segments). This means that also the observable consumption Y_{ikt} has a normal distribution, $Y_{ikt} \sim N(\mu_{ikt}, \sigma_k^2.\mu_{ikt})$, with expected value μ_{ikt} (i.e. the true consumption mean for a situation given by calendar effects and temperature is given by μ_{ikt}), variance $\sigma_k^2.\mu_{ikt}$, and coefficient of variation $\frac{\sigma_k}{\sqrt{\mu_{ikt}}}$. This is a bit milder variance-to-mean relationship than that used in (Brabec et al., 2009). The distribution is heteroscedastic (both over individuals and over time). Specifically, variability increases for times when the mean consumption is higher and also for individuals with higher average consumption (within the same segment). These changes are such that the coefficient of variation decreases within a segment, but its proportionality factor is allowed to change among segments to reflect different consumption volatility of e.g. households and small industrial establishments.

Taken together, it is clear that the model (1) has multiplicative correction terms for different calendar phenomena which modulate individual long term daily average consumption and a correction for temperature.

Type of the day code, j	Previous day ($t-1$)	Current day (t)	Next day ($t+1$)
1	working	working	working
2	working	working	non- working
2	non- working	working	non- working
3	non- working	working	working
4	working	non- working	non- working
4	non- working	non- working	non- working
5	non- working	non- working	working
5	working	non- working	working

Table 1. Type of the day codes

2.3 Temperature response function

Temperature response function ζ_{kt} is in the core of model (1). Here, we will describe how it is structured to capture details of the consumption to temperature relationship:

$$\zeta_{kt} = \left(\sum_{j=1}^{5} \psi_{jk}.I_{t\in D_j}\right).\left(1+\exp\left(\beta_k.\frac{\sum_{j=0}^{9} T_{t-j}}{10}\right)\right).\left(\rho_k(T_t)+\omega_k.\sum_{j=1}^{7}\delta_k^{j-1}.\rho_k(T_{t-j})\right), \quad (5)$$

where T_t is a daily temperature average for day t. We use a nation-wide average based on official met office measurements, but other (more local) temperature versions can be used. Even though a more detailed temperature info can be obtained in principle (e.g. reading at several times for a particular day, daily minima, maxima, etc.), we go with the average as with a cheap and easy to obtain summary.

$\rho_k(.)$ is a segment-specific temperature transformation function. It is assumed to be smooth and monotone decreasing (as it should to conform with principles mentioned in the Introduction). Since it is not known a priori, it has to be estimated from the data. Here we use a nonparametric formulation. In particular, we rely on loess smoother as a part of the GAM (generalized additive model) specified by (1) and (5), (Hastie & Tibshirani, 1990, Hastie et al., 2001).

It is easy to see that the right-most term in the parenthesis represents a nonlinear, but time invariant filter in temperature. In the transformed temperature, $\widetilde{T}_{kt} \equiv \rho_k(T_t)$, it is even a linear time invariant filter. In fact, it is quite similar to the so called Koyck model used in econometrics (Johnston, 1984). It can be perceived as a slight generalization of that model allowing for non-exponential (in fact even for non-monotone) lag weight on nonlinear temperature transforms $\widetilde{T}_{kt}$. $\omega_k > 0$ and $\delta_k^j > 0, j = 1, \ldots, 7$ are the parameters which characterize shape of the lag weight distribution. The behavior is somewhat more complex than geometrical decay dictated by the Koyck scheme. While the weights decay geometrically from ω_k at lag 1 (with the rate given by δ_k), they allow for arbitrary (positive) lag-zero-to- lag-one weight ratio (given by ω_k). In particular, they allow for local maximum of the lag distribution at lag one, which is frequently observed in empirical data. The parametrization uses weight of 1 for zero lag within the right-most parenthesis in order to assure identifiability (since the general scaling is provided by the two previous parentheses).

The term in the middle parenthesis essentially modulates the temperature effect seasonally. The moving average in temperature modifies the effect of left and right parentheses terms slowly, according to the "currently prevailing temperature situation", that is differently in year's seasons. In a sense, this term captures (part of) the interaction between the season and temperature effect - we use the word "interaction" in the typical linear statistical models' terminology sense of the word here (Rawlings, 1988). The impact is controlled by the parameter β_k. Note that the weighing in the 10-day temperature average could be non-uniform, at least in principle. Estimation of the weights is extremely difficult here so that we stick to the uniform weighting.

The left-most parenthesis contains an interaction term. It mediates the interaction of nonlinearly transformed temperature and type of the day. In other words, the temperature effect is different on different types of the day. This is a point that was missing in the SLP model formulation (Brabec et al., 2009) and it was considered one of its weaknesses – because the empirical data suggest that the response to the same temperature can be quite different if it occurs on a working day than in it occurs on Saturday, etc. The (saturated) interaction is described by the parameters $\psi_{jk}, j = 1, \ldots 5$. For numerical stability, they are estimated using a similar reparametrization as that mentioned in connection with α_{jk} after model (1) formulation in the section 2.2.

Consumption estimate $\hat{Y}_{ikt}$ (we will denote estimates by hat over the symbol of the quantity to be estimated) for day t, individual i of segment k is obtained as

$$\hat{Y}_{ikt} = \hat{\mu}_{ikt} = \hat{p}_{ik}.\hat{f}_{kt}. \quad (6)$$

Therefore, it is given just by evaluating the model (1), (5) with unknown parameters being replaced by their estimates.
This finishes the description of our gas consumption model (GCM) in daily resolution, which we will call GCMd, for shortness.

2.4 Hourly resolution

The GCMd model (1), (5) operates on daily basis. Obviously, there is no problem to use it for longer periods (e.g. months) by integrating/summing the outputs. But when one needs to operate on finer time scale (hourly), another model level is necessary. Here we follow a relatively simple route that easily achieves an important property of "gas conservation". In particular, we add an hourly sub-model on the top of the daily sub-model in such a way that the daily sum predicted by the GCMd will be redistributed into hours. That will mean that the hourly consumptions of a particular day will really sum to the daily total. To this end, we will formulate the following working model:

$$\log\left(\frac{q_{kth}}{1-q_{kth}}\right) = \lambda_{kth} = I_{t\in work}.\sum_{j=1}^{24} I_{j=h}.\varphi_{jk}^{w} + I_{t\in nonwork}.\sum_{j=1}^{24} I_{j=h}.\varphi_{jk}^{n} + \varepsilon_{kth} \quad (7)$$

where we use $\log(.)$ for the natural logarithm (base e). Indicator functions are used as before, now they help to select parameters (φ) of a particular hour for a working (w) and nonworking (n) day. This is an (empirical) logit model (Agresti, 1990) for proportion of gas consumed at hour h of the day t (averaged across data available from all customers of the given segment k):

$$q_{kth} = \frac{\sum_{i\in k} Y_{ikth}}{\sum_{i\in k}\sum_{h'} Y_{ikth'}} \quad (8)$$

with Y_{ikth} being consumption of a particular customer i within the segment k during hour h of day t. The logit transformation assures here that the modeled proportions will stay within the legal (0,1) range. They do not sum to one automatically, however. Although a multinomial logit model (Agresti, 1990) can be posed to do this, we prefer here (much) simpler formulation (7) and following renormalization. Model (7) is a working (or approximative) model in the sense that it assumes iid (identically distributed) additive error ε_{kth} with zero mean and finite second moment (and independent across k,t,h). This is not complete, but it gives a useful and easy to use approximation.

Given the φ_{hk}^{w} and φ_{hk}^{n}, it is easy to compute estimated proportion consumed during hour h and normalize it properly. It is given by

$$\tilde{q}_{kth} = \frac{\dfrac{1}{1+\exp(-\lambda_{kth})}}{\sum_{h' \in t} \dfrac{1}{1+\exp(-\lambda_{kth'})}} \tag{9}$$

Amount of gas consumed at hour h of day t is then obtained upon using (1) and (9). When we replace the unknown parameters (appearing implicitly in quantities like μ_{ikt} and $\tilde{q}_{kth}$) by their estimates (denoted by hats), as in (6), we get the GCM model in hourly resolution, or GCMh:

$$\hat{Y}_{ikth} = \hat{\mu}_{ikt} . \hat{\tilde{q}}_{kth} \tag{10}$$

In the modeling just described, the daily and hourly steps are separated (leading to substantial computational simplifications during the estimation of parameters). Temperature modulation is used only at the daily level at present (due to practical difficulty to obtain detailed temperature readings quickly enough for routine gas utility calculations).

3. Discussion of practical issues related to the GCM model

3.1. Model estimation

Notice that real use of the model described in previous sections is simple both in daily and hourly resolution, once its parameters (and the nonparametric functions $\rho_k(.)$) are given. For instance, its SW implementation is easy enough and relies upon evaluation of a few fairly simple nonlinear functions (mostly of exponential character). Indeed, the implementation of a model similar to that described here in both the Czech Republic and Slovakia is based on passing the estimated parameter values and tables defining the $\rho_k(.)$ functions (those need to be stored in a fine temperature resolution, e.g. by 0.1 ºC) to the gas distribution company or market operator where the evaluation can be done easily and quickly even for a large number of customers.

The separation property (4) is extremely useful in this context. This is because that the time-varying and nonlinear consumption dynamics part f_{kt} needs to be evaluated only once (per segment). Individual long-term-consumption-related p_{ik}'s enter the formula only linearly and hence they can be stored, summed and otherwise operated on, separately from the f_{kt} part.

It is only the estimation of the parameters and of the temperature transformations that is difficult. But that work can be done by a team of specialists (statisticians) once upon a longer period. We re-estimate the parameters once a year in our running projects.

For parameter estimation, we use a sample of customers whose consumption is followed with continuous gas meters. There are about 1000 such customers in the Czech Republic and about 500 in Slovakia. They come from various segments and were selected quasi-randomly from the total customer pool. Their consumptions are measured as a part of large SLP projects running for more than five years. Time-invariant information (important for classification into segments) as well as historical annual consumption readings are obtained from routine gas utility company databases. It is important to acknowledge that even though the data are obtained within a specialized project, they are not error-free. Substantial effort has to be exercised before the data can be used for statistical modeling (model specification and/or parameter estimation). In fact, one to two persons from our team work continuously on the data checking, cleaning and corrections. After an error is located, gas company is contacted and consulted about proper correction. Those data that cannot be corrected unambiguously are replaced by "missing" codes. In the subsequent analyses, we simply assume the MCAR (missing at random) mechanism (Little & Rubin, 1987).

As we mentioned already, the model is specified and hence also fitted in a stratified way – that is separately for each segment. Parameter estimation can be done either on original data (individual measurements) or on averages computed across customers of a given segment. The first approach is more appropriate but it can be troublesome if the data are numerous and/or contain occasional gross errors. In such a case the second might be more robust and quicker.

For the functions ρ_k, we assume that they are smooth and can be approximated with loess (Cleveland, 1979). Due to the presence of both fixed parameters and the nonparametric ρ_k's, the model GCMd is a semiparametric model (Carroll & Wand, 2003). Apart from the temperature correction part, the structure of the model is additive and linear in parameters, after log transformation, therefore it can be fitted as a GAM model (Hastie & Tibshirani, 1990), after a small adjustment. Naturally, we use normal, heteroscedastic GAM with variance being proportional to the mean, logarithmic link and offset into which we put $\log(p_{ikt})$ here. The estimation proceeds in several stages, in the generalized estimating equation style (Small & Wang, 2003). We start the estimation with estimation of the function ρ_k. To that end, we start with a simpler version of the model GCMd which formally corresponds to a restriction with parameters $\psi_{jk} \equiv 1, \beta_k = -\infty, \omega_k = 0$ being held. The $\hat{\rho}_k$ obtained from there is fixed and used in the next step where all parameters are re-estimated (including $\psi_{jk}, \beta_k, \omega_k$). The β, ω, δ parameters that appear nonlinearly in the temperature correction (5) are estimated via profiling, i.e. just by adding an external loop to the GAM fitting function and optimizing the profile quasilikelihood (McCullagh & Nelder, 1989) $Q_P(\beta, \omega, \delta) = \max_{others} Q(\beta, \omega, \delta, others)$ across β, ω, δ, where "others" denotes all other parameters of the model. This is analogous to what had been suggested in (Brabec et al., 2009).

Hourly sub-model needed for GCMh is estimated by a straightforward regression. Alternatively, one might use weighting and/or GAM (generalized linear model) approach.

For practical computations, we use the R system (R Development Core Team, 2010), with both standard packages (gam, in particular) and our own functions and procedures.

3.2 Practical applications of the model and typical tasks which it is used for

The model GCM (be it GCMd or GCMh) is typically used for two main tasks in practice, namely redistribution and prediction. First, it is employed in a retrospective regime when known (roughly annual) total consumption readings need to be decomposed into parts corresponding to smaller time units in such a way that they add to the total. In other words, we need to estimate proportions corresponding to the time intervals of interest, having the total fixed. When the total consumption $Y_{ik,[t_{1i},t_{2i}]}$ over the time interval $[t_{1i},t_{2i}]$ is known for an i-th individual of the k-th segment and it needs to be redistributed into days $t \in [t_{1i},t_{2i}]$, we use the following estimate:

$$\hat{Y}_{ikt}^{R} = \frac{Y_{ik,[t_{1i},t_{2i}]}.\hat{Y}_{ikt}}{\sum_{t'=t_{1i}}^{t_{2i}} \hat{Y}_{ikt'}} = \frac{Y_{ik,[t_{1i},t_{2i}]}.\hat{f}_{kt}}{\sum_{t'=t_{1i}}^{t_{2i}} \hat{f}_{kt'}} \tag{11}$$

where $\hat{Y}_{ikt}$ has been defined in (6). Disaggregation into hours would be analogous, only the GCMh model would be used instead of the GCMd. Such a disaggregation is very much of interest in accounting when the price of the natural gas changed during the interval $[t_{1i},t_{2i}]$ and hence amounts of gas consumed for lower and higher rates need to be estimated. It is also used when doing a routine network mass balancing, comparing closed network inputs and amounts of gas measured by individual customers' meters (for instance to assess losses). The disaggregated estimates might need to be aggregated again (to a different aggregation than original readings), in this context. The estimate of the desired consumption aggregation both over time and customers is obtained simply by appropriate integration (summation) of the disaggregated estimates (11):

$$\hat{Y}_{I,[T_1,T_2]}^{R} = \sum_{i,k \in I} \sum_{t=T_1}^{t=T_2} \hat{Y}_{ikt}^{R} \tag{12}$$

where I is a given index set. It might e.g. require to sum consumptions of all customers of two selected segments, etc.

Secondly, one might want to have prospective estimates of consumption over the interval which lies, at least partially, in future. Redistribution of the known total is not possible here, and the estimates have to be done without the (helpful) restriction on the total. They will have to be based on $\hat{Y}_{ikt}$ alone. It is clear that such estimates will have to be less precise and hence less reliable, in general. This is even more true in the situation when the average annual consumption changes systematically, e.g. due to the external economic conditions (like crisis) which the GCM model does not take into account. At any rate, the disagreggated estimates can then be used to estimate a new aggregation in a way totally parallel to (12), i.e. as follows:

$$\hat{Y}_{I,[T_1,T_2]} = \sum_{i,k \in I} \sum_{t=T_1}^{t=T_2} \hat{Y}_{ikt} \tag{13}$$

It is important to bear on mind that the estimates (both $\hat{Y}_{ikt}^R$ and $\hat{Y}_{ikt}$, as well as their new aggregations) are estimates of means of the consumption distribution. Therefore, they are not to be used directly e.g. for maximal load of a network or similar computations (mean is not a good estimate of maximum). Estimates of the maxima and of general quantiles (Koenker, 2005) of the consumption distribution are possible, but they are much more complicated to get than the means.

3.3 Model calibration

In some cases, it might be useful to calibrate a model against additional data. This step might or might not be necessary (and the additional data might not be even available). One can think that if the original model is good (i.e. well calibrated against the data on which it was fitted), it seems that there should be no space for a further calibration. It might not be necessarily the case at least for two reasons.

First, the sample of customers on which the model was developed, its parameters fitted, and its fit tested might not be entirely representative for the total pool of customers within a given segment or segments. The lack of representativity obviously depends on the quality of the sampling of the customer pool for getting the sample of customers followed in high resolution to obtain data for the subsequent statistical modeling (model "training" or just the estimation of its parameters). We certainly want to stress that a lot of care should be taken in this step and the sampling protocol should definitely conform to principles of the statistical survey sampling (Cochran, 1977). The sample should be definitely drawn at random. It is not enough to haphazardly take a few customers that are easy to follow, e.g. those that are located close to the center managing the study measurements. Such a sample can easily be substantially biased, indeed! Taking the effort (and money) that is later spent in collecting, cleaning and modeling the data, it should really pay off to spend a time to get this first phase right. This even more so when we consider the fact that, when an inappropriate sampling error is made, it practically cannot be corrected later, leading to improper, or at least, inefficient results. The sample should be drawn formally (either using computerized random number generator or by balloting) from the list of all relevant customers (as from the sampling frame), possibly with unequal probabilities of being drawn and/or following stratified or other, more complicated, designs. It is clear, that to get a representative sample is much more difficult than usual, since in fact, we sample not for scalar quantities but for curves which are certainly much more complicated objects with much larger space for not being drawn representatively in all of their (relevant) aspects. It might easily happen that while the sample is appropriate for the most important aspects of the consumption trajectory, it might not be entirely representative e.g. for summer consumption minima. For instance, the sample might over-represent those that do consume gas throughout the year, i.e. those that do not turn off their gas appliances even when the temperature is high. The volume predicted error might be small in this case, but when being interested in relative model error, one could be pressed to improve the model by recalibration (because the small numerators stress the quality of the summer behavior substantially).

Secondly, when the model is to be used e.g. for network balancing, it can easily happen that the values which the model is compared against are obtained by a procedure that is not entirely compatible with the measurement procedure used for individual customer readings and/or for the fine time resolution reading in the sample. For instance, we might want to compare the model results to amount of gas consumed in a closed network (or in the whole gas distribution company). While the model value can be obtained by appropriate integration over time and customers easily, for instance as in (13), obtaining the value which this should be compared to is much more problematic than it seems at first. The problem lies in the fact that, typically there is no direct observation (or measurement) of the total network consumption. Even if we neglect network losses (including technical losses, leaks, illegal consumption) or account for them in a normative way (for instance, in the Czech Republic, there are gas industry standards that describe how to set a (constant) loss percentage) and hence introduce the first approximation, there are many problems in practical settings. The network entry is measured with a device that has only a finite precision (measurement errors are by no means negligible). The precision can even depend on the amount of gas measured in a complicated way. The errors might be even systematic occasionally, e.g. for small gas flows which the meter might not follow correctly (so that summer can easily be much more problematic than winter). Further, there might be large customers within the network, whose consumption need to be subtracted from the network input in order to get HOU+SMC total that is modeled by a model like GCM. These large customers might be followed with their own meters with fine time precision (as it is the case e.g. in the Czech Republic and Slovakia), but all these devices have their errors, both random and systematic. From the previous discussion, it should be clear now that the "observed" SMC+HOU totals

$$Z_{..t} = \left(input\right)_t - \left(sum\,of\ nonHOUSMC\ customers\right)_t - \left(normative\ losses\right)_t \tag{14}$$

have not the same properties as the direct measurements used for model training. It is just an artificial, indirect construct (nothing else is really feasible in practice, however) which might even have systematic errors. Then the calibration of the model can be very much in place (because even a good model that gives correct and precise results for individual consumptions might not do well for network totals).

In the context of the GCM model, we might think about a simple linear calibration of $Z_{..t}$ against $\sum_{i,k} \hat{Y}_{ikt}$ (where it is understood that the summation is against the indexes corresponding to the HOU+SMC customers from the network), i.e. about the calibration model described by the equation (15) and about fitting it by the OLS, ordinary least squares (Rawlings, 1988) i.e. by the simple linear regression:

$$Z_{..t} = \kappa_1 + \kappa_2 . \sum_{i,k} \hat{Y}_{ikt} + error_t \,. \tag{15}$$

Conceptually, it is a starting point, but it is not good as the final solution to the calibration. Indeed, the model (15) is simple enough, but it has several serious flaws. First, it does not acknowledge the variability in the $\sum_{i,k} \hat{Y}_{ikt}$. Since it is obtained by integration of estimates obtained from random data, it is a random quantity (containing estimation error of $\hat{Y}_{ikt}$'s). In particular, it is not a fixed explanatory variable, as assumed in standard regression problems that lead to the OLS as to the correct solution. The situation here is known as the measurement error problem (Carroll et al., 1995) in Statistics and it is notorious for the possibility of generating spurious regression coefficients (here calibration coefficients) estimates. Secondly, the (globally) linear calibration form assumed by (15) can be a bit too rigid to be useful in real situations. Locally, the calibration might be still linear, but its coefficients can change smoothly over time (e.g. due to various random disturbances to the network).

Therefore, we formulate a more appropriate and complete statistical model from which the calibration will come out as one of its products. It is a model of state-space type (Durbin & Koopman, 2001) that takes all the available information into account simultaneously, unlike the approach based on (15):

$$\begin{aligned} &Y_{ikt} = \mu_{ikt} + \varepsilon_{ikt} \qquad\qquad k = 1, \cdots, K \\ &Z_{..t} = \exp(\gamma_t).\sum_{k=1}^{K} \tau_k . \sum_{i=1}^{n_k} Y_{ikt} + \eta_t \\ &\gamma_t = \gamma_{t-1} + \xi_t \\ &\varepsilon_{ikt} \sim N\left(0, \sigma_k^2 . \mu_{ikt}\right), \eta_t \sim N\left(0, \sigma_\eta^2\right), \xi_t \sim N\left(0, \sigma_\xi^2\right) \end{aligned} \tag{16}$$

Here, we take the GCMd parameters as fixed. Their unknown values are replaced by the estimates from the GCMd model (1), (5) fitted previously (hence also μ_{ikt} appearing explicitly in the first K equations, as well as in the error specification and implicitly in the $(K+1)$-th equation are fixed quantities). Therefore, we have only the variances σ_k^2 , σ_η^2 , σ_ξ^2 as unknown parameters, plus we need to estimate the unknown γ_t's. In the model (16), the first $K+1$ equations are the measurements equations. In a sense they encompass simultaneously what models (1), (5) and (15) try to do separately. There is one state equation which describes possible (slow) movements of the linear calibration coefficient $\exp(\gamma_t)$ in the random walk (RW) style (Kloeden & Platen, 1992). The RW dynamics is imposed on the log scale in order to preserve the plausible range for the calibration coefficients (for even a moderately good model, they certainly should be positive!). The random error terms are specified on the last line. We assume that ε , η and ξ are mutually independent and that each of them is independent across its indexes (t and i,k). For identifiability, we have to have a restriction on τ_k's (that is on the segment-specific changes of the calibration). In general, we prefer the multiplicative restriction $\prod_{k=1}^{K} \tau_k = 1$, but in practical applications of (16), we took even more restrictive model with $\tau_k \equiv 1$.

Although the model (16) can be fitted in the frequentist style via the extended Kalman filter (Harvey, 1989), in practical computations we prefer to use a Bayesian approach to the estimation of all the unknown quantities because of the nonlinearities in the observation operator. Taking suitable (relatively flat) priors, the estimates can be obtained from MCMC simulations as posterior means. We had a good experience with Winbugs (2007) software.

Advantage of the model (16) is that, apart from calibration, it provides a diagnostic tool that might be used to check the fitted model. For instance, comparing the results of the GCMd model (1), (5) alone to the results of the calibration, i.e. of (1), (5), (15), we were able to detect that the GCMd model fit was OK for the training data but that it overestimated network sums over the summer, leading to further investigation of the measurement process at very low gas flows.

4. Illustration on real data

In this paragraph, we will illustrate performance of the GCMd on real data coming from various projects we have been working with. Since these data are proprietary, we normalize the consumptions deliberately in such a way, that they are on 0-1 scale (zero corresponds to the minimal observed consumption and one corresponds to the maximal observed consumption). This way, we work with the data that are unit-less (while the original consumptions were measured in $m^3/100$).

Figure 1 illustrates that the gas consumption modeling is not entirely trivial. It shows individual normalized consumption trajectories for a sample of customers from HOU4 (or household heaters') segment that have been continuously measured in the SLP project. Since considerable overlay occurs at times, the same data are depicted on both original (left) and logarithmic (right) consumption scale. Clearly, there is a strong seasonality in the data (higher consumption in colder parts of the year), but at the same time, there is a lot of inter-individual heterogeneity as well. This variability prevails even within a single (and rather well defined) customer segment, as shown here. Some individuals show trajectories that are markedly different from the others. Most of the variability is concentrated to the scale, which justifies the separation (4). Due to the normalization, we cannot appreciate the fact that the consumptions vary over several orders of magnitude between seasons, which brings further challenges to a modeler. Note that model (1) deals with these (and other) complications through the particular assumptions about error behavior and about multiplicative effects of various model parts.

Figure 2 plots logarithm of the normalized consumption against the mean temperature of the same day for the data sampled from the same customer segment as before, HOU3. Here, the normalization (by subtracting minimum and scaling through division by maximum) is applied to the ratios $\frac{Y_{ikt}}{p_{ik}}$ as to the quantities more comparable across individuals. Clearly, the asymptotes are visible here, but there is still substantial heterogeneity both among different individual customers and within a customer, across time (temperature response is different at different types of the day, etc., as described by the model (1)). This second, within individual variability is exactly where the model (5) comes into play. All of this (and more) needs to be taken into account while estimating the model.

After motivating the model, it is interesting to look at the model's components and compare them across customer segments. They can be plotted and compared easily once the model is estimated (as described in the section 3.1). Figure 3 compares shapes of the nonlinear temperature transformation function $\rho_k(.)$ across different segments, k . It is clearly visible that the shape of the temperature response is substantially different across different segments – not only between private (HOU) and commercial (SMC) groups, but also among different segments within the same group. The segments are numbered in such a way that increasing code means more tendency to using the natural gas predominantly for heating. We can observe that, in the same direction, the temperature response becomes less flat. When examining the curves in a more detail, we can notice that they are asymmetric (in the sense that their derivative is not symmetric around its extreme). For these and related reasons, it is important to estimate them nonparametrically, with no pre-assumed shapes of the response curve. The model (5) with nonparametric ρ_k formulation brings a refinement e.g. over previous parametric formulation of (Brabec at al., 2009), where one minus the logistic cumulative distribution function (CDF) was used for temperature response as well as over other parametric models (including asymmetric ones, like 1-smallest extreme value CDF) that we have tried. Figure 4 shows $\exp(\alpha_{1k}),\ldots,\exp(\alpha_{5k})$'s of model (1), which correspond to the (marginal) multiplicative change induced by operating on day of type 1 though 5. Indeed, we can see that HOU1 consisting of those customers that use the natural gas mostly for cooking have more dramatically shaped day type profile (corresponding to more cooking over the weekends and using the food at the beginning of the next week, see the Table 1). Figure 5 shows a frequency histogram for normalized p_{ik}'s from SMC2 segment (subtracting minimum p_{ik} and dividing by maximum p_{ik} in that segment).

One could continue in the analysis and explore various other effects or their combinations. For instance, there might be considerable interest in evaluating μ_{ikt} for various temperature trajectories (e.g. to see what happens when the temperature falls down to the coldest day on Saturday versus shat happens when that is on Wednesday). This and other computations can be done easily once the model parameters are available (estimated from the sample data). Similarly, one can be interested in hourly part of the model. Figure 6 illustrates this viewpoint. It shows proportions of the daily total consumed at a particular hour for the HOU1 segment. They are easily calculated from (9), when parameters of model (7) have been estimated. For this particular segment of those customers that use the gas mostly for cooking, we can see much more concentrated gas usage on weekends and on holidays (related to more intensive cooking related to lunch preparation).

How does the model fit the data? Figure 7 illustrates the fit of the model to the HOU4 (heaters') data. This is fit on the same data that have been used to estimate the parameters. Since the model is relatively small (less than 20 parameters for modeling hundreds of observations), signs of overfit (or of adhering to the training data too closely, much more closely than to new, independent data) should not be too severe. Nevertheless, one might be interested in how does the model perform on new data and on larger scale as well. The problem is that the new, independent data (unused in the fit) are simply not available in the fine time resolution (since the measurement is costly and all the available information should be used for model training). Nevertheless, aggregated data are available. For instance, total (HOU+SMC) consumptions for closed distribution networks, for individual gas companies and for the whole country are available from routine balancing. To be able to compare the model fit with such data, we need to integrate (or re-aggregate) the model estimates properly, e.g. along the lines of formula (13). When we do this for the balancing data from the Czech Republic, we get the Figure 8. The fit is rather nice, especially when considering that there are other than model errors involved in the comparison (as discussed in the section 3.3) – note that the model output has not been calibrated here in any way.

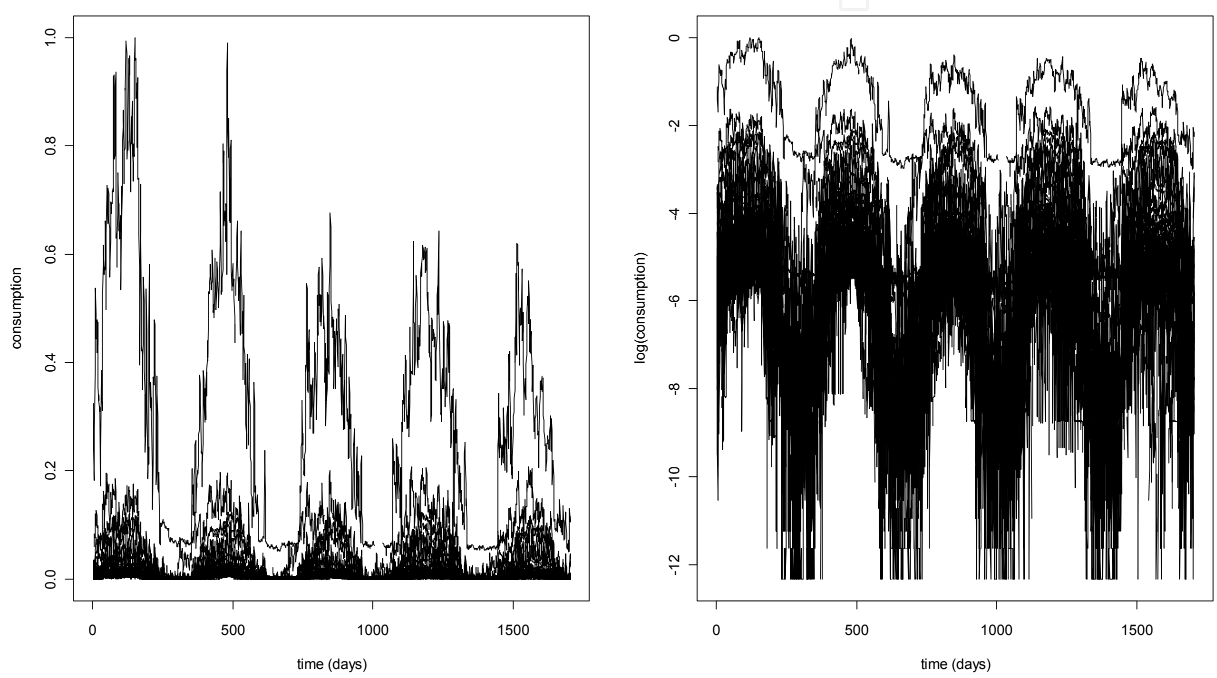


Fig. 1. Overlay of individual consumption trajectories (left – normalized untransformed, right – logarithmically transformed normalized consumptions). Day 1 corresponds to starting point of the SLP projects (October 1, 2004).

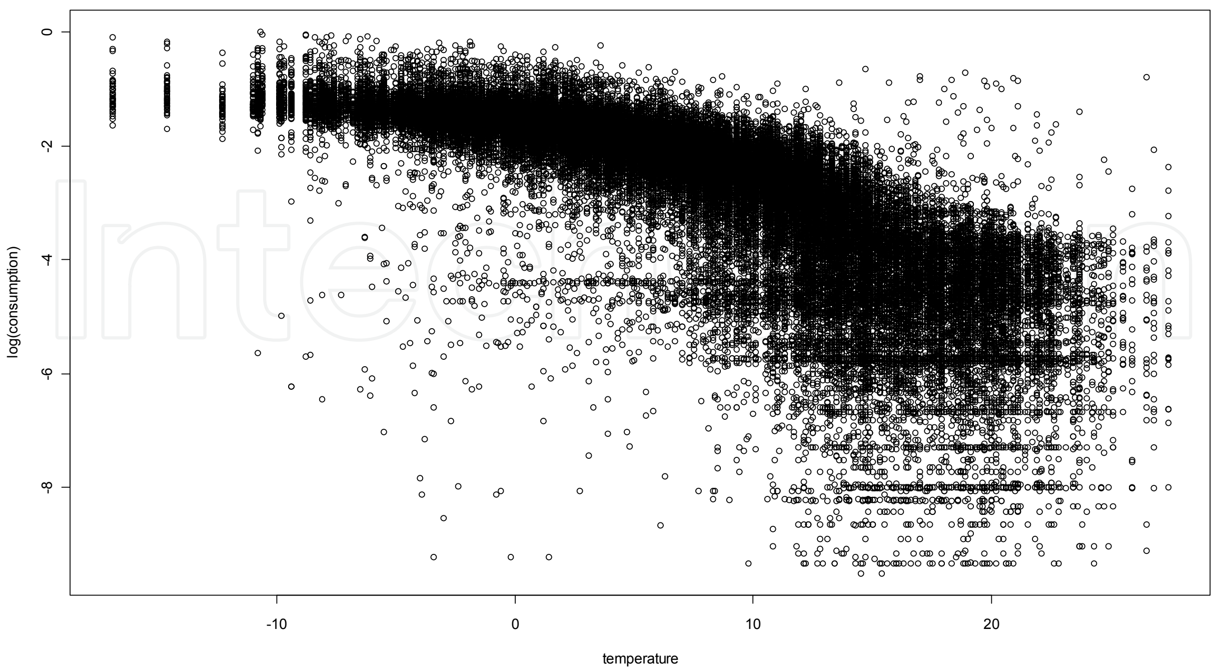


Fig. 2. Logarithmically transformed normalized consumption against current day average temperature.

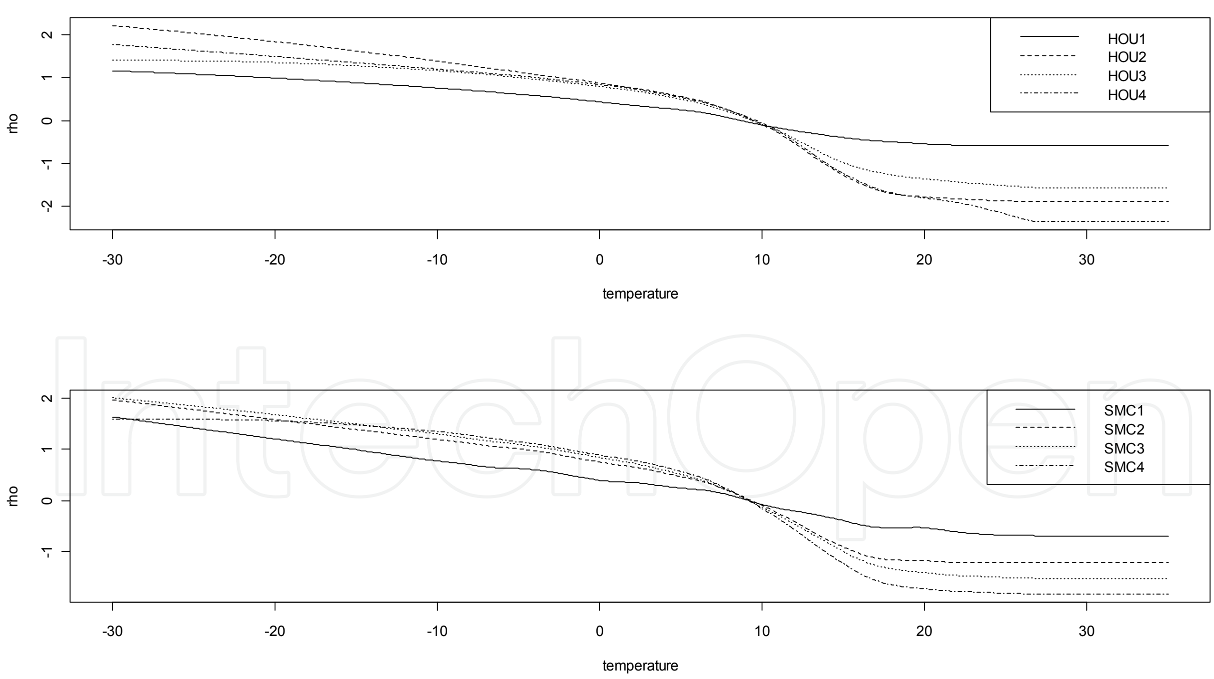


Fig. 3. Temperature response function $\rho_k(.)$ of (5), compared across different HOU and SMC segments.

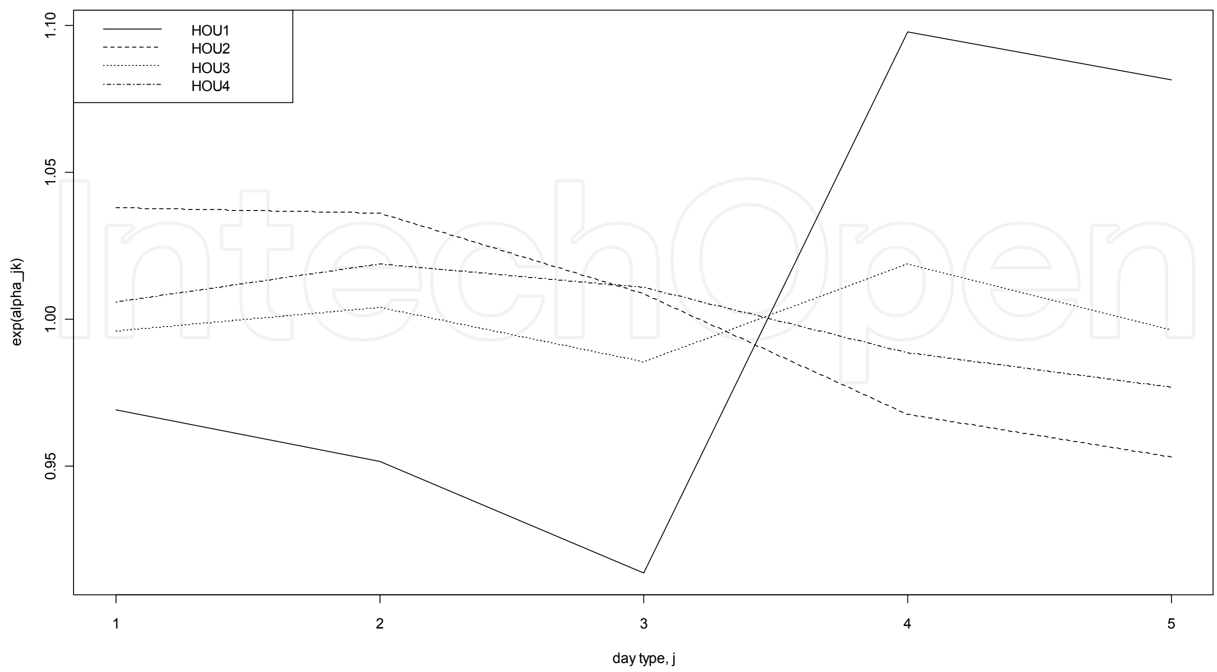


Fig. 4. Marginal factors of day type, $\exp(\alpha_{jk})$ from model (1).

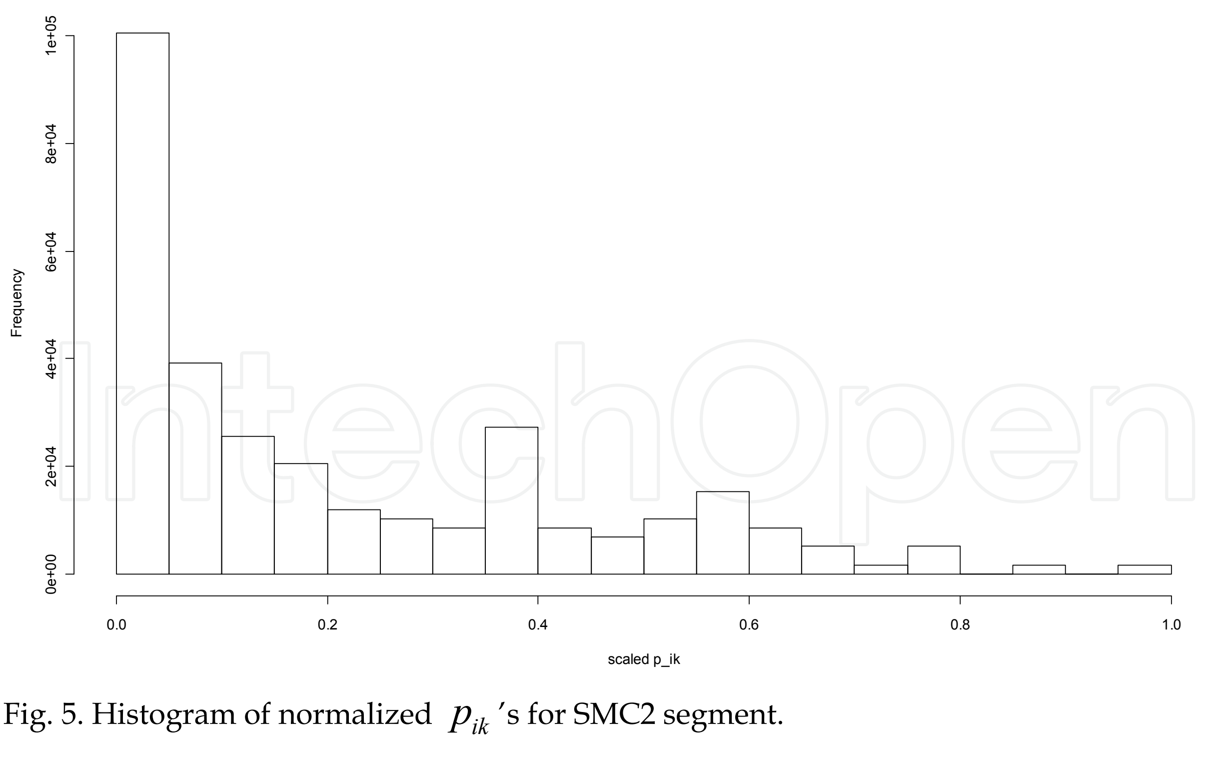


Fig. 5. Histogram of normalized p_{ik}'s for SMC2 segment.

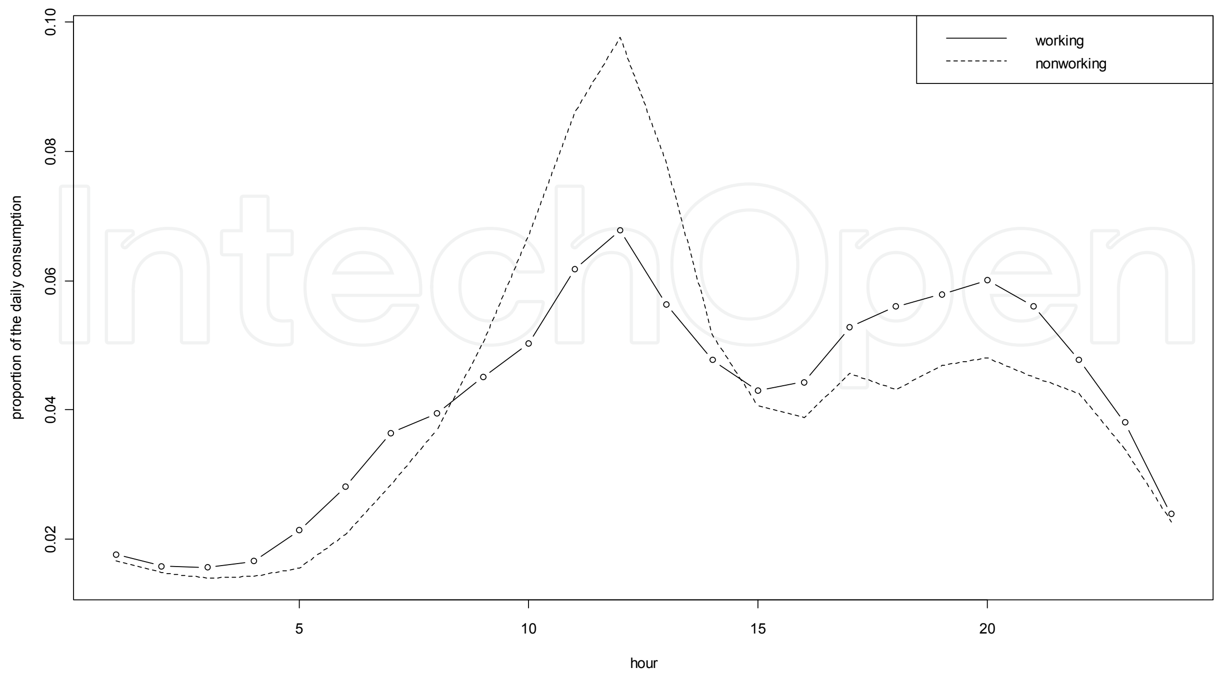


Fig. 6. Proportions of daily consumption totals consumed in a particular hour of the day, i.e. $\widetilde{q}_{kth}$'s from (9), compared between working and nonworking day for HOU1 segment (i.e. for „cookers“).

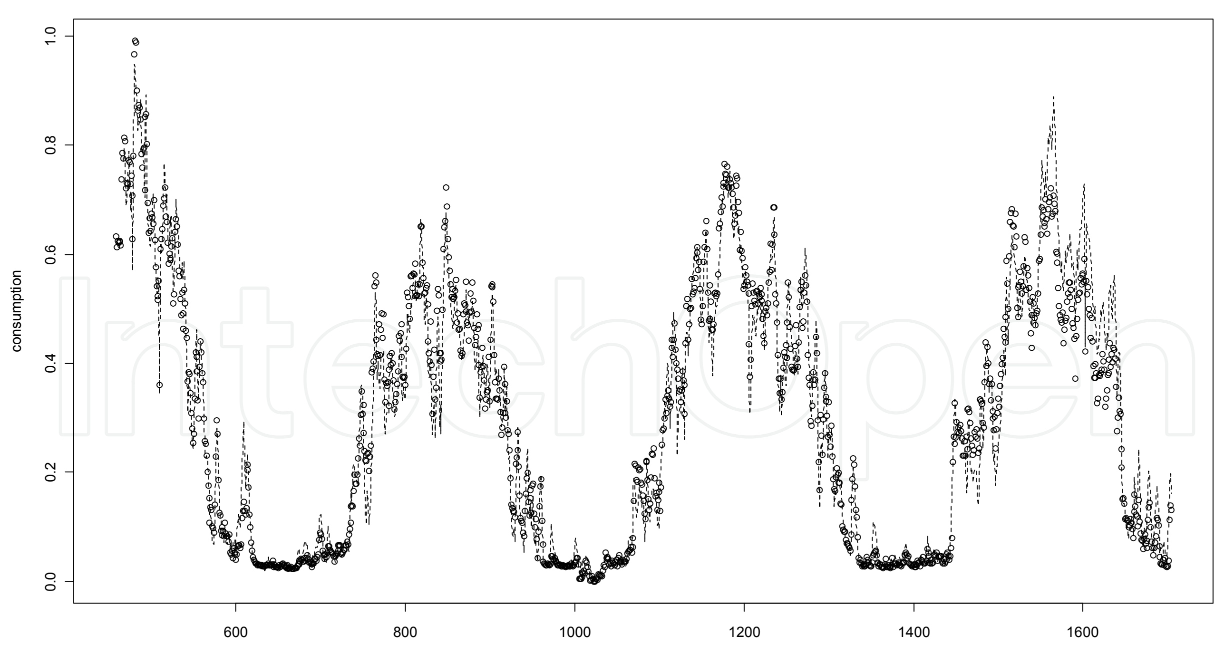


Fig. 7. Fit of the model (1) to the HOU4 data (normalized consumptions as dots and normalized model output as a dotted line).

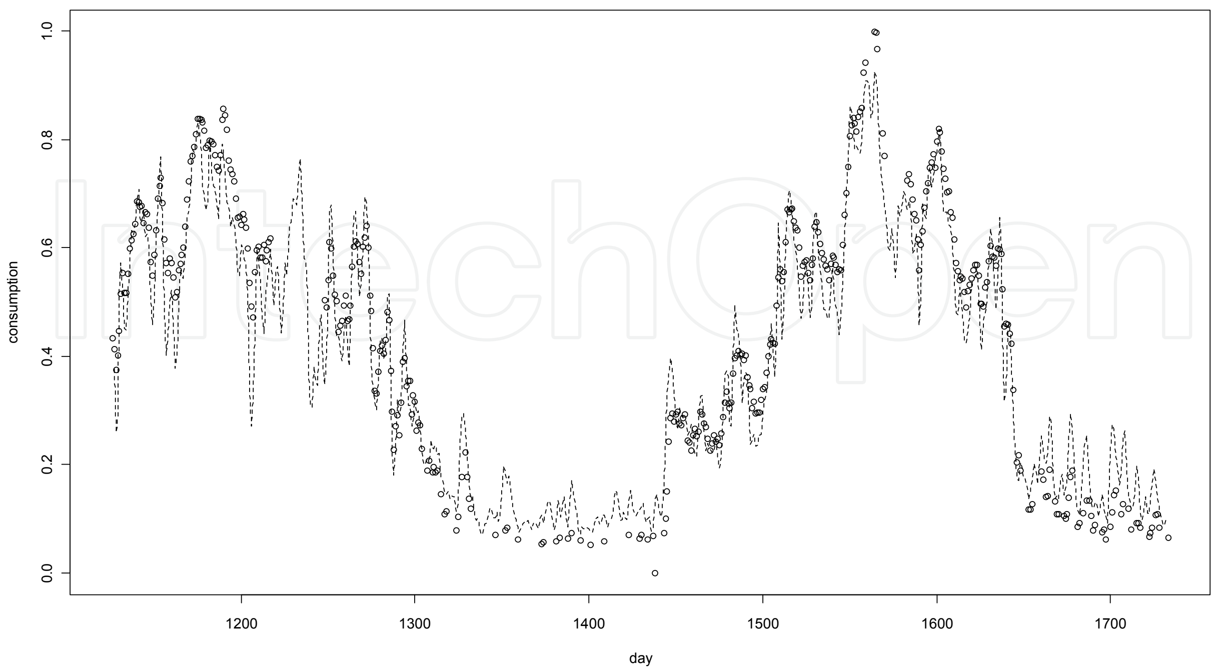


Fig. 8. Fit of the model (1) after disaggregation and re-aggregation of normalized model output, according to (13) on the CR total HOU+SMC consumption data over a period of more than a year.

5. Future work and discussion of some open problems

The model GCM, as described in previous sections uses a single temperature average for all customers. That might be perfectly appropriate if it is employed within a gas company operating over a relatively small and homogeneous region. Even for larger and less homogeneous regions, it can provide good approximations - as we know from its nationwide implementations in both the Czech Republic and Slovakia (both of them being relatively small countries, admittedly). For large and climatologically heterogeneous countries, it might be useful to "regionalize" the GCM in the sense that the global temperature T_t entering the formula (5) would be replaced by the local temperature relevant for the i,k -th customer, i.e. by T_{ikt} . Obviously, it would not be practical to require temperature measurements for each individual customer. Therefore, T_{ikt} would on the i,k index only through the relation of being included in some more local region for which the temperature daily average would be available separately (e.g. county). Technically, this is very simple indeed. Nevertheless, such an improvement requires appropriate (regionally) stratified sample.

The calibration model (16) can be expanded to cover not only proportional but also additive biases. Note that, compared to the full linear calibration of (15), model (16) assumes that the additive bias is zero. The assumption is in line with what we experienced in practice, but for other situations, the model (16) can be expanded by one more state equation to have time-varying intercept as well.

Another useful way of expanding (16) is to drop the $\tau_k \equiv 1$ restriction (while keeping the multiplicative identifiability-related restriction). That might be useful in case when different segments would show very different proportional biases. In our experience, the segment-specific multipliers are very difficult to estimate, however.

The GCM model is very efficient computationally and easy to comprehend conceptually because it implies the relation (4), i.e. the multiplicative separation of the individual-specific but time-invariant and common but time-varying parts. Lack of interaction between the two parts (i.e. between the individual and dynamical parts) is important in practice because it eases implementation substantially. Sums of the p_{ik}'s and sums of the f_{kt}'s can be formed separately when doing the integrations like (12) and (13). The log-additive GCM model certainly captures substantial part of the consumption behavior. If more detailed modeling is attempted, p_{ik}'s might be allowed to follow a time trend (e.g. in connection with changes in economy or with building insulation trends, etc.). Willingness to expand the model along these lines might be hampered by the fact that the impact of this should not be overwhelming however, when the GCM model parameters are re-estimated periodically in relatively short periods (e.g. annually), as suggested. Furthermore, trend common to everybody (within a segment) might not be strong enough to matter at all. More useful would be to assume a trend, but to allow the trend to change the trend from individual to individual. In other words, to allow the individual*dynamics interaction (where the * and the word interaction are used in the statistical sense, as explained before) instead of the additivity of the two terms currently assumed. Obviously, full (or saturated) interaction in the analysis of variance (ANOVA) model style (Graybill, 1976) style is out of question here (since it would not be even estimable). Nevertheless, it is possible to attempt for a more parsimonious model where only part of the interaction (with less degrees of freedom than the saturated interaction) would be specified. Particularly promising route is to allow for time-varying p_{ikt}, i.e. for individual p_{ikt}'s to follow time series models implying slow, but individual-specific dynamics. This is an interesting topic, we have been working on recently (Brabec et al., 2007; Brabec et al., 2008a).

6. Conclusion

In this chapter, we have introduced a gas consumption model GCM for household and small medium customers in daily and hourly resolution and showed how it can be used for various practical tasks, including estimations of consumption aggregates integrated over time and/or customers as well as network related balancing. A model similar to the implementation described here has been running in nationwide system in the Czech Republic and Slovakia for several years already.

The model has a moderately rich structure but it has been built with very strong accent on easy and efficient practical implementation in a gas company or energy market operator environment. It is built in a modular way, enhancing serviceability and making local adjustments to somewhat different conditions rather easy. For more complicated

adjustments, we might be a help a new user with the statistical modeling part if it would result in an interesting project.
The GCM model is built from the first principles, in close contact with empirical behavior of the observed consumptions. It is specified in formal terms as a full blown statistical model (not only mean behavior but also variability assumptions and distributional behavior are given by the model). Our practical experience in natural gas modeling has been strongly supporting the idea that rigorous statistical formulation always pays off here and that it is to be preferred to a haphazard ad hoc or even black box type approaches. There is a lot of structure and many systematic features that a good gas consumption model should follow closely in order to be useful.

7. Acknowledgement

The work was partly supported by the grant 1ET400300513 of the Grant Agency of the Academy of Sciences of the Czech Republic as well as by the Institutional Research Plan AV0Z10300504 'Computer Science for the Information Society: Models, Algorithms, Applications'. We would like to acknowledge important support from the M100300904 project of the Academy of Sciences of the Czech Republic. We also would like to thank to the people from the RWE GasNet, formerly West Bohemian Gas Distribution Company (J. Bečvář, J. Čermáková and others) and to V. Jilemnický from RWE Plynoprojekt for their help and willingness to discuss gas distribution background problems and issues.

8. References

Agresti, A. (1990). *Categorical data analysis*. John Wiley. New York.

Bailey, J. (2000). Load profiling for retail choice: Examining a complex and crucial component of settlement. *Electricity Journal.* 13, 69-74

Brabec, M.; Konár, O.; Malý, M.; Pelikán, E.; Vondráček, J. (2009). A statistical model for natural gas standardized load profiles. *JRSS C - Applied Statistics*. 58, 1, 123-139

Brabec, M.; Malý, M.; Pelikán, E.; Konár, O. (2009a). Statistical calibration of the natural gas consumption model. *WSEAS transactions on systems*. 8, 7, 902-912

Brabec, M.; Konár, O.; Pelikán, E.; Malý, M. (2008). A nonlinear mixed effects model for prediction of natural gas consumption by individual customers. *International Journal of Forecasting*. 24, 659-678

Brabec, M.; Konár, O.; Pelikán. E.; Malý, M. (2008a). Hierarchical model for estimation of yearly sums from irregular longitudinal data. *Book of abstracts, ISF symposium on forecasting*, Nice, France, page 139

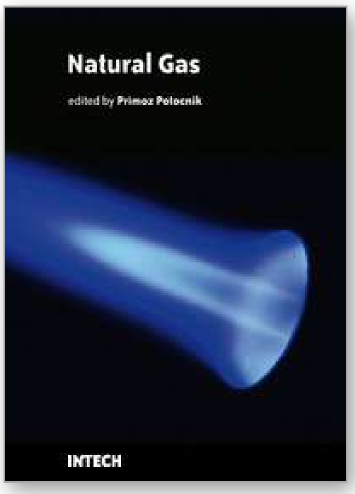
Brabec, M.; Konár, O.; Malý,M.; Pelikán, E.; Vondráček, J. (2007). State space model for aggregated longitudinal data. *Abstract Book, 27th International Symposium on Forecasting*, New York 24.-27.6.2007, page 46, ISF.

Carroll, R. J. D.; Ruppert, L. A.; Stefanski. (1995). *Measurement error in nonlinear models*. Chapman & Hall/CRC. London.

Carroll, R. J. & Wand, M. P. (2003). *Semiparametric regression*. Cambridge University Press. Cambridge.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*. 74, 829-836

Cochran, W. G. (1977). *Sampling techniques*. John Wiley. New York.
Davidian, M. & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. Chapman and Hall. London.
Durbin, J. & Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford University Press. Oxford.
Elvira (2010). Elvira project webpage, http://www.cs.cas.cz/nlm/elviraindex-en.htm
Gil S.; Deferrari J. (2004). Generalized model of prediction of natural gas consumption. *Transactions of the ASME*. 126, 90-97
Graybill, F. A. (1976). *Theory and application of the linear model*. Wadsworth & Brooks–Cole. Pacific Grove.
Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press. Cambridge.
Hastie, T. & Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall. New York.
Hastie, T.; Tibshirani, R.; Friedman, J. (2001). *The elements of statistical learning*. Springer. New York.
Johnson, R. A. & Wichern, D. W. (1988). *Applied multivariate statistical analysis*. Englewood Cliffs: Prentice Hall. New Jersey.
Johnston, J. (1984). *Econometric Methods*. McGraw Hill. New York.
Kloeden, P. E. & Platen, E. (1992). *Numerical solution of stochastic differential equations*. Springer. New York.
Liedermann, P. (2006). Standardized load profiles for electricity supply—surrogate method for billing customers without continuous measurement (in Czech). *Energetika*, 56, 402–405
Lyness F.K. (1984). Gas demand forecasting. *Statistician*. 33, 9-21
Koenker, R. (2005). *Quantile regression*. Cambridge University Press. Cambridge.
Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley. New York.
McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall. London.
Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. Springer. New York.
Potočnik,P.; Thaler, M.; Govekar, E.; Grabec, I.; Poredoš, A. (2007). Forecasting risks of natural gas consumption in Slovenia. *Energy Policy*. 35, 4271-4282
Rawlings, J. O. (1988). *Applied regression analysis: A research tool*. Wadsworth & Brooks Cole. Pacific Grove.
R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/. Accessed on 23 March 2010.
Searle, S. R. (1971). *Linear models*. John Wiley. New York.
Small, C. G. & Wang, J. (2003). *Numerical methods for nonlinear estimating equations*. Clarendon Press. Oxford.
Vondráček, J.; Pelikán, E.; Konár, O.; Čermáková, J.; Eben, K.; Malý, M.; Brabec, M. (2008). A statistical model for the estimation of natural gas consumption. *Applied Energy*. 85, 5, 362-370
Winbugs (2007). *Winbugs with Doodle Bugs*. Version 1.4.3. (6th August 2007) http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml. Medical Research Council. United Kingdom.



Natural Gas
Edited by PrimoÃ…Â¾ PotoÃ„Ânik

ISBN 978-953-307-112-1
Hard cover, 606 pages
Publisher Sciyo
Published online 18, August, 2010
Published in print edition August, 2010

The contributions in this book present an overview of cutting edge research on natural gas which is a vital component of world's supply of energy. Natural gas is a combustible mixture of hydrocarbon gases, primarily methane but also heavier gaseous hydrocarbons such as ethane, propane and butane. Unlike other fossil fuels, natural gas is clean burning and emits lower levels of potentially harmful by-products into the air. Therefore, it is considered as one of the cleanest, safest, and most useful of all energy sources applied in variety of residential, commercial and industrial fields. The book is organized in 25 chapters that cover various aspects of natural gas research: technology, applications, forecasting, numerical simulations, transport and risk assessment.