Statistical calibration of the natural gas consumption model

MAREK BRABEC, MAREK MALÝ, EMIL PELIKÁN, ONDŘEJ KONÁR
Department of nonlinear modeling
Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Praha 8
CZECH REPUBLIC
mbrabec@cs.cas.cz http://www.cs.cas.cz/nlm

Abstract: - We will discuss a problem pertinent to many situations in which a statistical model is developed on a sample of individuals (describing their trajectories), but then it is applied on a much larger population of interest. Typical examples occur in natural gas and other energy consumption contexts. Due to various deficiencies of the original sample-only based model, due to possible inconsistencies between sample and the population as a whole, and/or due to inherently different nature of the available sample and population data, calibration arises as a natural way to improve original model. We start with a simple approach and proceed to introduce a formal and flexible, time-varying statistical model of state-space nature, from which such a calibration will come out as one of the products. The calibration model is quite general and hence it can be used far beyond the particular context of natural gas consumption modeling in which it was originally motivated.

Key-Words: linear calibration, natural gas consumption modeling, Bayesian approach, statistical model, time-varying calibration, state-space model

1 Introduction

It is customary to fit/train a statistical model (i.e. to estimate its unknown parameters) on a random sample of units taken from the population of interest and then to use the estimated model for the entire population. Such an approach is not always problem-free, as we will show on example arising in the context of natural gas consumption modeling.

There might be inherent discrepancies among the sample data and the domain in which the model is to be used. These need to be solved by a sort of calibration procedure. Situation might not be as easy as that solvable by a straightforward linear calibration of the model output. First, there might be inherent nonlinearity, leading to more complicated calibration functions and/or calibrations that are time-varying. Second, the form of the sample data and the data on which the model is to be used in practice might be different. For instance, sample data and target population inference might relate to different *functionals* of the process of interest (e.g. weighted sums, or convolutions with a kernel of a completely known form).

In the natural gas consumption context, utility companies often need to estimate total (or mean) consumption of large groups of customers (e.g. entire segments of customer population defined by the type of gas use, like space heating, water heating, cooking, etc.) over a given period (e.g. calendar month, year, week). Since the gas meters of household and small+middle size commercial customers (HOU+SMC) are routinely red as consumption sums in roughly annual intervals (whose position and length varies from individual to individual), it is not feasible to obtain estimates for short time periods from the routine data themselves. On the other hand, it is possible to obtain a sample of relevant customers and follow their gas consumption for a while, in fine time resolution (daily). Such a sample is expensive and hence it cannot be exceedingly large.

Nevertheless, it can be used to fit a model describing how the consumption depends on temperature, time of the year and other calendar effects and estimate its parameters. The resulting model can then be used for decomposition of time-aggregated routine (roughly annual) data into shorter interval consumption estimates and/or for related regularization (estimation of consumption over regular intervals, same for all customers, like calendar months, calendar years, etc. from irregular individual reading intervals).

There are various models available for practical use, see for instance the standardized load profiles (SLP) model [3] or the so called Gamma model [15], [13] and many others, including those for electricity demand [1], [20]. Alternative approaches, including those based on semiparametric, e.g. wavelet [17], [18] or radial basis functions [19] modeling are possible.

2 Problem Formulation

Quality of the statistical model can be obviously verified on sample data, using well known principles of statistical model checking, [14] (this might include both in-sample and out-of-sample comparisons, as one can look at both lack of fit and prediction abilities of the model). When checking the model performance “at large” – on sizeable groups of customers, one faces a much less standard problem. This indeed does occur frequently in practice, e.g. when a particular gas distribution company needs to evaluate consumption sums for substantial parts (or even totals) of various segments of their customer pool (e.g. of the entire HOU segment, etc.).

First, we have to check generalizability of the model from sample to the entirety of target population. In some sense, this is an inverse problem to that solved in [2]. Second, population data are usually of vastly different resolution (aggregation). For fundamental reasons mentioned previously, we do not have access to individual consumption trajectories in daily resolution for the complete customer pool (nor for any of its substantial parts). But, from a mass balancing (comparing inputs, outputs, standard losses, etc.), performed regularly by the gas company, one can have time series of total daily consumptions for all HOU+SMC customers for a closed balancing system. The data can come from nationwide network, individual company’s network or local closed network (LCN).

Such data can be used not only for checking the quality of consumption model, but also for its improvement. Clearly, the total daily consumptions represent an additional source of information, as in the calibration of computer models on empirical data [10].

Different aggregation level for sample and total data (both over time and over individuals) makes the process a bit more challenging. In fact, the situation resembles the modifiable aerial unit problem (MAUP), known and well studied in Spatial Statistics as well as in other contexts, see [5].

In this paper, we will describe one possible approach to consumption model calibration, illustrate its application on real data and discuss its byproducts important for model building as well as for understanding the data generating process.

3 Calibration model

We start with a statistical model of consumption, formulated and fitted on sample data. It will be calibrated subsequently.

3.1 Model for sample data (having high resolution across time and individual customers)

For instance, we might start with the following simplification of the SLP model [3] which is used routinely in both Czech Republic and Slovakia. We have for the natural gas consumption (in m^3) Y_{ikt} of i-th customer in k-th segment on day t:

$$Y_{ikt} = p_{ik}\, f_t(T_t, T_{t-1}; \theta_k)\varepsilon_{ikt} \tag{1}$$

where

- p_{ik} is the mean daily consumption for i-th customer of k-th segment (obtained from his/her historical annual data of at least 3 years, corrected for temperature, actual length of the historical period, as well as for calendar effects). It corresponds to the long-term individual consumption expressed in per day basis.
- stratifying on k-th segment, $f_t(T_t, T_{t-1}; \theta_k)$ is a nonlinear function of current (T_t) and one day lagged (T_{t-1}) temperature (and of certain calendar effects), involving (vector of) unknown parameters θ_k. Concretely, we have

$$\begin{aligned} f_t(T_t, T_{t-1}; \theta_k) &= \alpha_k + \sum_{j=0}^{6} \delta_{jk} I(day\ t\ is\ weekday\ j) + \\ &\kappa_{hk} I(t\ is\ a\ bank\ holiday) + \kappa_{ck} I(t\ is\ Christmas) + \\ &\kappa_{ek} I(t\ is\ Easter) + \beta_{0k} \exp\left(-\exp\left(\frac{T_t - \mu_k}{\sigma_k}\right)\right) + \\ &\beta_{1k} \exp\left(-\exp\left(\frac{T_{t-1} - \mu_k}{\sigma_k}\right)\right) \end{aligned} \tag{2}$$

with $I(.)$ being an indicator function and with obvious identifiability restrictions on weekday effect (days are coded so that 0 corresponds to Sunday and 6 to Saturday) in ANOVA [8] style ($\sum_{j=0}^{6} \delta_{jk} = 0$), as in [3]. Temperature correction part was developed in [4]. It is motivated by the nonlinear character of the temperature response. It is intriguing to try to interpret and motivate the correction function in relationship to automatic temperature control algorithms in use today and to the new developments in natural gas burners. The most important feature here is the presence of both lower and upper asymptotes and necessarily decreasing character of the consumption to temperature relationship. While lower asymptote is obvious (but not reflected in various ad hoc models in practical use) – due to heating switch-off during periods of high temperature, the upper asymptote might not be that intuitive. It occurs due to final capacity of installed heating devices (on very cold days, the consumption cannot go to infinity, and even before the physical limitations occur, there is a change in consumer behavior, leading to more economic behavior, maintenance of smaller indoor temperature to prevent huge heat losses and dramatic increase of gas energy bills).

- model errors are independently and lognormally distributed, $\varepsilon_{ikt} \sim LN\left(0, \sigma_k^2\right)$. Generally, lognormal distribution is skewed and it has been observed to fit the empirical data nicely in previous studies, [3]. There are also other arguments for its use in the present context (nice statistical properties leading to relatively simple maximum likelihood estimation of model (1) or its competitors, the fact that it preserves nonnegative character of consumption in estimates automatically without any explicit restrictions on parameters, etc.).

For original consumption models that are analogous to (1), or at least similar to it in their structure, one can proceed with the calibration in analogous way as described below, perhaps with obvious small modifications. This is the case of various models in practical use, e.g. of the model called GAMMA, [15]. For structurally different models, we would modify (1) appropriately (on case by case basis) and proceed with the calibration as previously.

3.2 Additional data from network balancing

From network balancing, we get the data of the form:

$$Z_{..t} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ikt}, \qquad (3)$$

As we see, there is double summation (integration) here – being performed across K segments and individual members of a particular segment (there are n_k of them for k-th segment). The sets of individuals included in the sample used for model development and in the balancing network are either disjoint (in case of LCN's) or the sample constitutes a negligible fraction of the number of customers involved in balancing (in case of company-wide or nation-wide networks and samples of roughly hundreds of customers). This means that daily sum data and sample data are independent (or essentially independent).

Balancing data $Z_{..t}$ might cover the same time interval as the sample data or a different one. When they cover the same interval, we are dealing with "model generalization" across either individuals only. On the other hand, if they cover a different interval, the model generalization in question concerns both generalization across individuals and across time.

Moreover, the process of balancing is complicated and involves summation as well as subtraction of various random quantities, each of different precision. For instance, readings of gas meters of very large commercial customers whose data are available daily are subtracted from readings of gas meters on pipeline input. The result is corrected for normatively determined losses and for several other known processes. The net result is that the statistical properties of $Z_{..t}$'s are different from the properties of individual Y_{ikt}'s. In particular, the error behavior is different. It is complicated in detail, but since there are many small sources of variability, it tends to behave more normally (as seen from the central limit theorem and from the empirical data behavior itself). This should be reflected in a statistical calibration model.

3.3 Simple calibration

It is natural to begin to tackle the calibration problem from a simple starting point. Natural and perhaps the simplest procedure to start with is to use a straightforward linear regression of $Z_{..t}$ on appropriately summed estimates from model (1), that is on $\sum_{k=1}^{K}\sum_{i=1}^{n_k} \hat{Y}_{ikt} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} p_{ik}\, f_t\left(T_t, T_{t-1}; \hat{\theta}_k\right)$, using ordinary least squares (OLS). Actually, this is what is done in practice quite often to check appropriateness of the model (and "rectifying" it). Formally, that amounts to:

$$Z_{..t} = \vartheta_1 + \vartheta_2 \sum_{k=1}^{K}\sum_{i=1}^{n_k} \hat{Y}_{ikt} + \psi_t, \qquad (4)$$

where $\psi_t \sim N\left(0, \sigma_\psi^2\right)$ and $\vartheta = \left(\vartheta_1, \vartheta_2\right)'$ are parameters to estimate. Alternatively, it is possible to be even more parsimonious and to work with a restricted version of (4), with $\vartheta_1 = 0$. This would spare one parameter and it would correspond to the multiplicative-only calibration version (which is often, but not always reasonable - as long as it is reasonable to expect that the intercept is negligible, based on, say physical considerations, as it is here).

Approach described by (4) or by the restricted version of it is simple enough. It might serve well for a first orientation in the problem, but it is inappropriate from many points of view. Most of the arguments against it are simple and use only basic statistical theory. The largest problem is that such a simplistic approach treats the model estimates as fixed quantities (while they are estimates obtained from random quantities and hence themselves random, as (1) shows clearly). Such an approach would give a distorted picture about the precision of the calibrated model. Further, the shape of the calibration (inherent to (4)) is severely limited. In fact, (4) insists on linear calibration, with coefficients constant in time. This is not flexible enough for many processes, especially those which are not stable in time (natural gas consumption of an entire customer pool is an archetypal example – due to the changing prices, changing customer habits, improvement in both heating devices and home insulation, etc., etc.). From operational point of view, it does not allow for any feedback into θ_k estimation. It just modifies the resulting predictor in a rather ad hoc manner.

Formally, the procedure amounts to using system of estimating equations, [11], [7] of triangular $\begin{matrix} g_1(Y_1,\theta)=0 \\ g_2(Y_2,\theta,\vartheta)=0 \end{matrix}$, with $\theta=\left(\theta_1',\ldots,\theta_K'\right)'$, rather than rectangular $\begin{matrix} g_1(Y_1,\theta,\vartheta)=0 \\ g_2(Y_2,\theta,\vartheta)=0 \end{matrix}$ form (in case of regression (4), they are used only in an approximate way, obviously). Triangular estimating equations amount to a simplification, but unfortunately to a substantial loss of information as well.

3.4 Time-varying calibration model

Unlike in (4), the real discrepancy between the model and balancing output might be more complicated. While the linear calibration (4) with the restriction $\vartheta_1=0$ is reasonable locally, it might not be plausible in global terms. In particular, the calibration coefficient ϑ_2 often changes over time (showing periodicity and trend). This is because there might be additional processes operating on network balancing data that are not manifesting themselves on sample data (see the example below). It also allows for the nonlinearity of the calibration.

It motivates us to introduce the following model:

$$Z_{..t}=\vartheta_{2t}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{Y}_{ikt}+\psi_t . \qquad (5)$$

It might not be immediately clear how to perform such a calibration, especially in the view of the fact that (5) does not specify a complete statistical model.

Instead, we reformulate an analog of (5) as a proper (fully specified) stochastic model. In fact, we will make use of the nice to work with state-space formulation, [9]

$$\begin{aligned} Y_{ikt} &= p_{ik} f_{kt}\left(T_t,T_{t-1}\right)\varepsilon_{ikt} \qquad k=1,\ldots,K \\ Z_{..t} &= \exp(\gamma_t)\sum_{k=1}^{K}\lambda_k\sum_{i=1}^{n_k}Y_{ikt}+\omega_t \qquad (6) \\ \omega_t &\sim N\left(0,\sigma_\omega^2\right). \end{aligned}$$

This is an extended time-varying (but proportional-only) calibration model (with $\vartheta_{2t}=\exp(\gamma_t)$) and segment-specific proportion adjustment λ_k (with $\prod_{k=1}^{K}\lambda_k=1$ for identifiability). The model (6) is a nonlinear model that intertwines all segments and hence prevents stratification on segment. This is not a complication for a practical model use but for the estimation. Computation is harder (as one cannot utilize the "divide and conquer" strategy inherent in the stratification). Parameter space is larger and to get the right hand side, we have to integrate across segments (and individuals, but because of multiplicative structure of the first set of K equations in (6) and conditioning on p's, this is easy). Maximum likelihood parameter estimation is (obviously) still possible but numerically complicated when K is large.

To alleviate the estimation problems, we follow a Bayesian approach, instead, using MCMC (essentially the Gibbs sampling, [16], [12]) to simulate from posterior distributions, using flat (not very informative) priors for variances (inverse gamma with both parameters equal to 0.001) and *proper* uniform priors for other parameters (bounds given by somehow inflated physically plausible intervals). In order to guard against possible autocorrelation in the MCMC posterior sample, we use thinning (1:20).

Note that the model is quite general. Specifically, it is not tied exclusively to the natural gas consumption. Hence, it might be used in other contexts as well (for calibration of other sample-derived consumption model, using total consumption - in other energy/utility contexts (e.g. electricity, water consumption), but also in biology, chemistry and other measurements.

4 Real data example

We will utilize the model (6) on data from the Czech load profile construction project (see [3] for details). There, we have $K = 6$ (3 HOU and 3 SMC) segments that are used by the natural gas utility company in practice. The data are proprietary, so that we are not allowed to publish them in the original form. In order to prevent disclosing (valuable and sensitive) consumption scale information, we will scale the data to 0-1 range (subtracting minimum observed consumption and dividing by data range). Obviously, the original consumptions are wildly different in different segments (as they correspond to customer types using the natural gas for widely different purposes, ranging from cooking to heating or technological use).

We will use a restricted version of the model (6), insisting on no change of segment proportion ($\lambda_k \equiv 1$). To generate the posterior, we choose 20000 burn-in and then use 100000 simulations (with 1:100 thinning).

The model (6) was fitted simultaneously to sample data and a particular LCN data. Fig. 1a to Fig. 1f show how the model (6) estimates and the sample data (separately for separate segments). Days are coded from the beginning of LCN data (depicted in Fig. 3a). The fit is very close to what would be obtained from model (1), i.e. when only the sample data were fitted (with no total available for calibration). This is illustrated in Fig. 2a to Fig. 2f for individual HOU and SMC segments. Note that the agreement between model (1) and model (6) is usually relatively close, but not always – particular exceptions are HOU23 and SMC34. In other words, it is clear that the calibration part model expansion does not distort the fit on the sample data – as it really should not! On the other hand, the change in the fit quality is much more dramatic for the network sums. Fig. 3a and Fig. 3b show that the model (1) overestimates severely and systematically during summers (periods of low gas consumption). This can be seen both from the comparison of time series in Fig. 3a and (perhaps in a more detail) also on residual plot for the same data in Fig. 3b. It is a good news that the multiplicative time-varying calibration built into the model (6) removes the deficiency. Notably, it also removes the secondary problem connected to local underestimation around day 200 in the original consumption model.

Note that, in this example, calibration is performed on time points of two types: those seen in the sample data as well as those not seen by the sample (stretching both to right and left endpoints of sampling interval, as is clear from horizontal axis range comparison between Fig. 1a to Fig. 1f and Fig. 3a).

It is revealing to inspect the trajectory of calibration factors. Fig. 4 shows the $\vartheta_{2t} = \exp(\gamma_t)$ estimates for the LCN, that is the estimates of the time-varying multiplicative calibration factors. Clearly, $\vartheta_{2t} \equiv 1$ would correspond to the unbiased original model (that would not need any calibration). The more the ϑ_{2t} deviates from 1, the more correction the original model needs (ϑ_{2t}'s larger than one correspond to underestimation of the original model and vice versa for ϑ_{2t}'s smaller than one). In fact, $100.(\vartheta_{2t} - 1)$ gives the percentage of over/underestimation of the original model. This is useful for careful inspection of the original model and for finding its weak spots. But not only that, it can be used directly for its improvement!

It is one of the main advantages of the model (6) and of the approach presented here that the amount of over/underestimation (and hence of the correction needed to alleviate it) varies over the time. In particular, it is not forced to be constant as it would be in the case of the simple linear calibration (4). On the other hand, if the data would be such that the over/underestimation factor would be constant, it would adapt to it (the price being paid would consist of only somewhat smaller efficiency).

Taken overall, the Fig. 4 shows (similarly as in Fig. 3a, 3b, but in a much more detailed way) that the original non-calibrated (sample-based) model is essentially correct during the winters, but that it is overestimating during the summers (and it needs to be brought down by the calibration process).

It is practically important that the trajectory in Fig. 4 reveals substantial discrepancy between the sample and totality the LCN. Fit on the sample data is reasonable, but the fit on the totals is deficient and needs calibration. Therefore, it is not the model to be blamed here, but disagreement between the two sources of data. In subsequent analyses, we found that this is not a specific feature of the particular LCN. Similar phenomenon occurs if we use other (even much larger) network totals for calibration. Discussions with gas utility experts suggested several possible mechanisms for the discrepancy, including systematic bias in certain types of gas meters during the periods of low consumption, some misrepresentation of the sample as well as other, more subtle reasons. Since the shape of discrepancies is similar to what we saw in the context of semi-individual modeling of larger customers [4], it seems that one of the possibly more important factors might be the (holiday and regular maintenance related) switch-off. Such a switch-off might be relatively rare and not very well represented in the sample (it will probably be concentrated to and more synchronized for larger customers) and not predictable from information routinely available in customer databases. Switch-off represents an additional process that should be modeled in future. Ideally via some explicit model, or at least semi-empirically, via similar state-space model approach as in (6), perhaps with a richer time-series structure.

Note that the estimates shown in preceding figures are based on standard posterior output and hence they correspond to smoother (i.e. $E(X_t \mid data_1,\ldots,data_T)$ for X_t as a quantity of interest and data available for time points $1,\ldots,T$). Filter ($E(X_t \mid data_1,\ldots,data_t)$), or h-step-ahead predictor ($E(X_{t+h} \mid data_1,\ldots,data_t)$) can be obtained via MCMC as well (albeit with more computational effort), which might be useful for certain predictive practical tasks.

When we refit the model (6) without the $\lambda_k \equiv 1$ restriction, we find out that half of the segments (HOU1, HOU23, SMC34) have λ's whose credible intervals exclude 1, showing some possibility of fit improvements in future. The λ estimates are rather unstable, however, so that we do not report the results here.

It is attractive that the posterior distributions are not very far from being normal (in accord with their well known asymptotic normality, [6]). This is not only assuring but also interesting from practical point of view. For large-scale deployment of the calibration model, it is a bit of hindrance that, in order to produce estimate of consumption for a particular day, one needs to store (and use) complete estimate of posterior distribution of model parameters. For practitioners, it is not entirely straightforward to use, say 1000 sets of parameters and generate (posterior mean) estimate of consumption from there as the mean of nonlinear function of parameters, with respect to their posterior distribution, i.e. $p_{ik} E_{\theta_k|data} f_t(T_t, T_{t-1}; \theta_k)$. Implementation of this is not difficult in the research context, using experimental software, but it might be more problematic in the real world of a gas utility company, where one meets a tremendous pressure for simplicity under all circumstances.

Are there any simplifications at sight at all, here? Obviously, $p_{ik} E_{\theta_k|data} f_t(T_t, T_{t-1}; \theta_k)$ is not the same as $p_{ik} f_t(T_t, T_{t-1}; E(\theta_k \mid data))$. Nevertheless, it is relatively close (at least in our situation) and hence usable as a crude approximation, if needed.

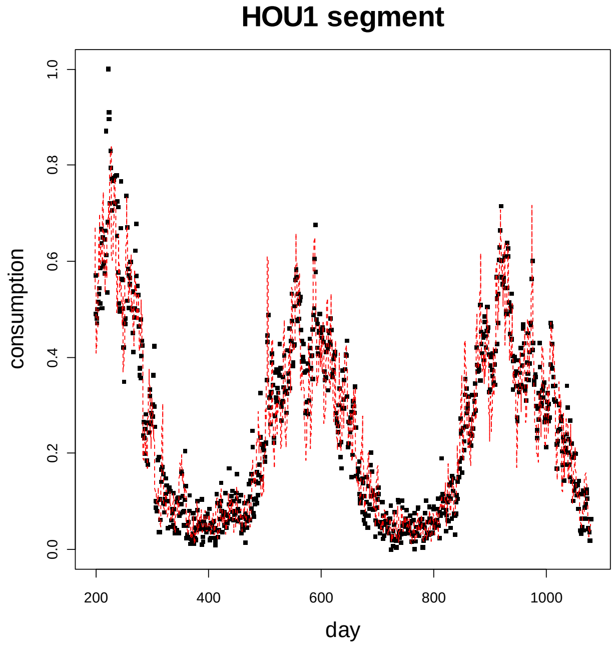

Fig. 1a: Consumptions in HOU1 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

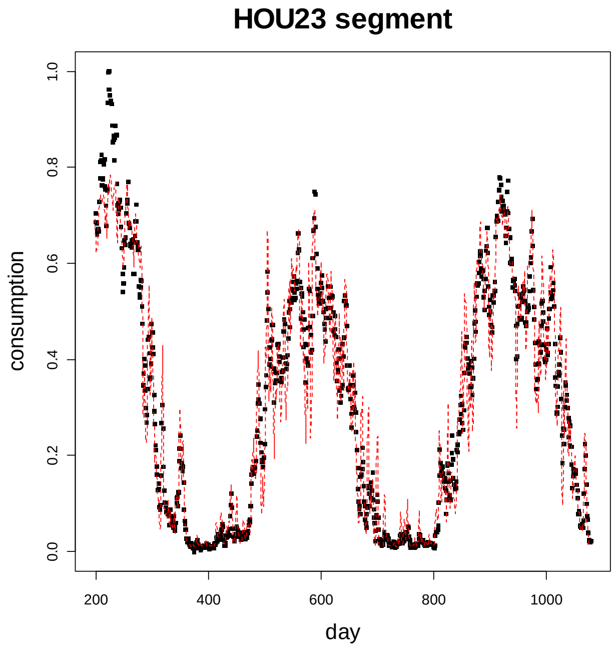

Fig. 1b: Consumptions in HOU23 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

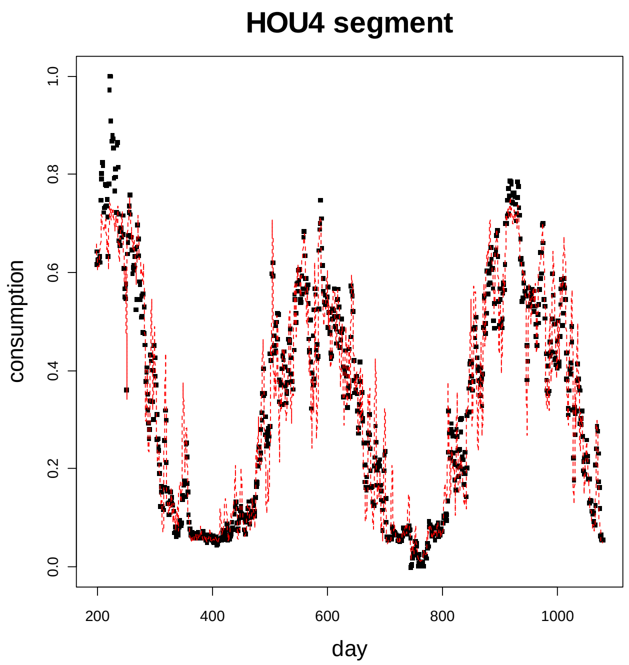

Fig. 1c: Consumptions in HOU4 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

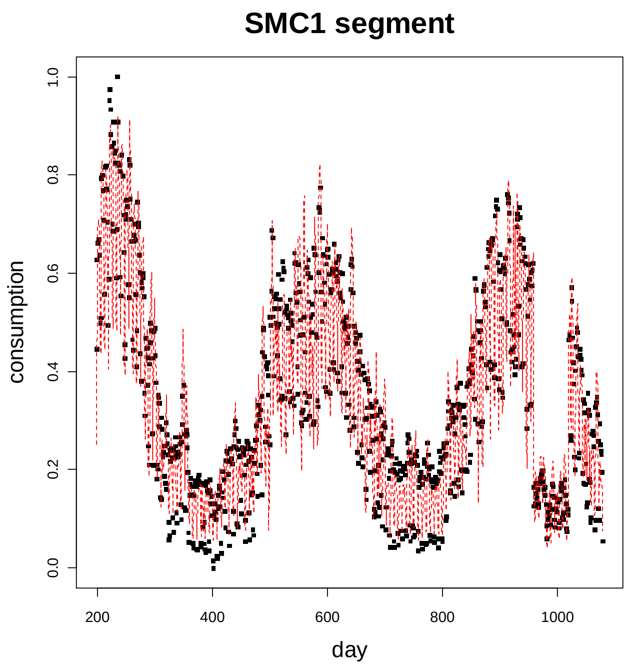

Fig. 1d: Consumptions in SMC1 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

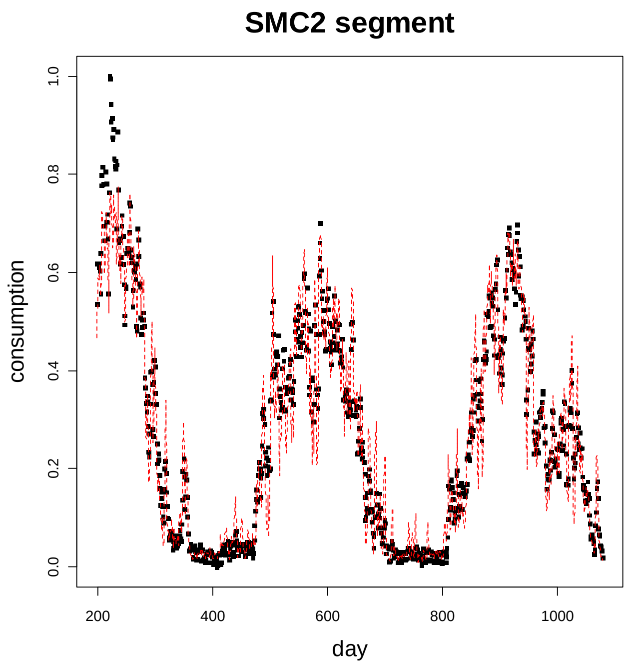

Fig. 1e: Consumptions in SMC2 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

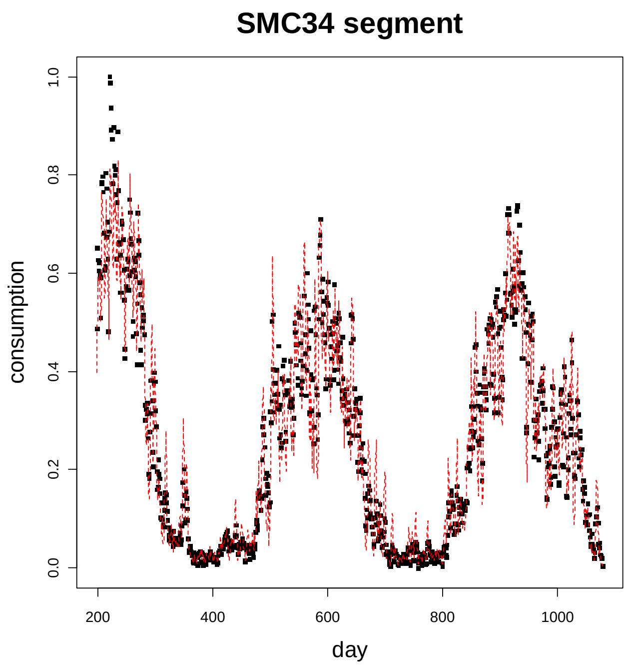

Fig. 1f: Consumptions in SMC34 segment of the sample (or "training") data. Empirical means are plotted as dots, the model (6) estimates as lines.

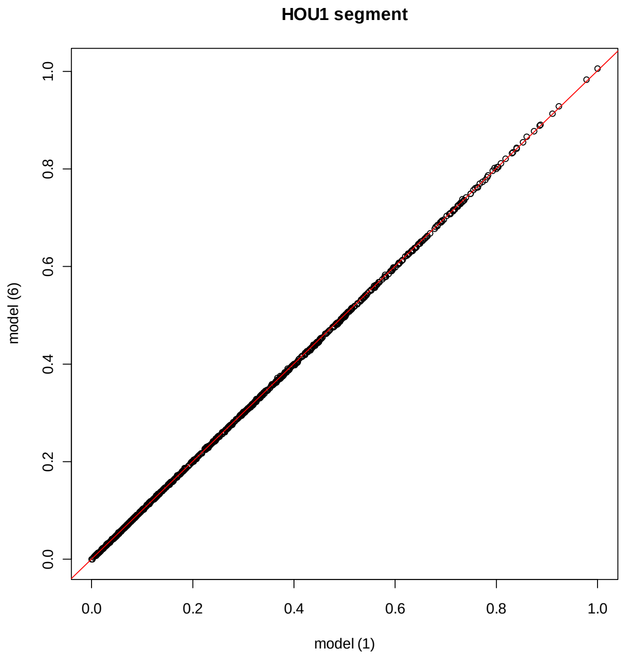

Fig. 2a: Estimates from the model (1) and the model (6) for HOU1 segment (line corresponding to identity mapping is plotted to ease the comparison).

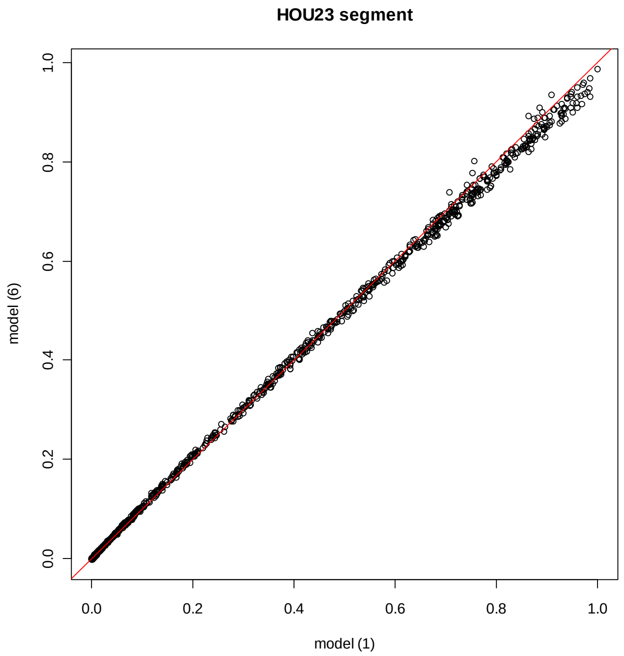

Fig. 2b: Estimates from the model (1) and the model (6) for HOU23 segment (line corresponding to identity mapping is plotted to ease the comparison).

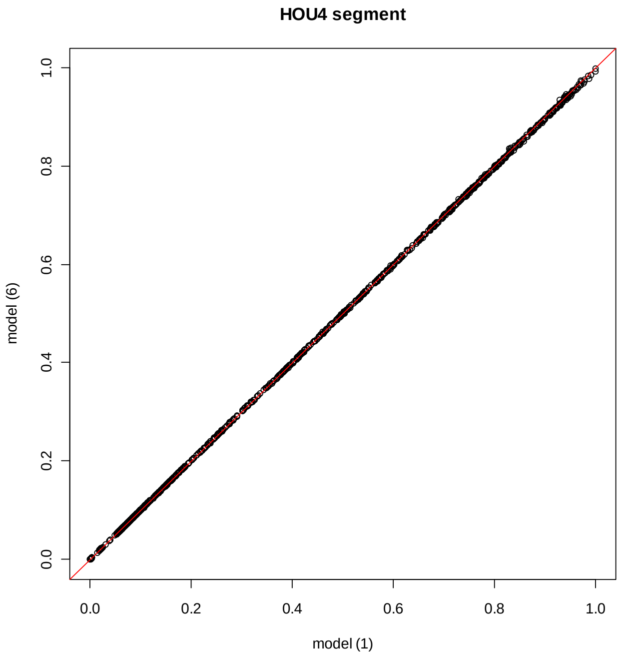

Fig. 2c: Estimates from the model (1) and the model (6) for HOU4 segment (line corresponding to identity mapping is plotted to ease the comparison).

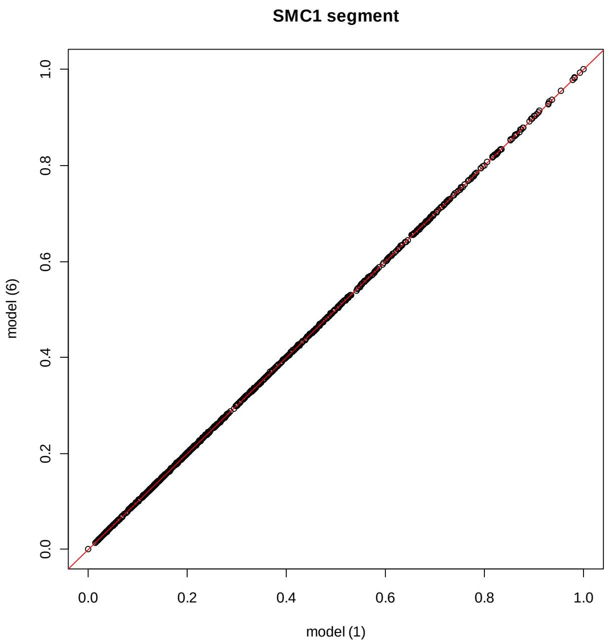

Fig. 2d: Estimates from the model (1) and the model (6) for SMC1 segment (line corresponding to identity mapping is plotted to ease the comparison).