Statistical surveillance in gas distribution networks

M. Brabec

Institute of Computer Science,
Czech Academy of Sciences

mbrabec@cs.cas.cz

Energy Days 2018
Praha, September 20-22, 2018

Natural gas distribution network surveillance

- It is important to monitor how an energy distribution network works
  - – both for routine checking,
  - – and to discover patterns useful for future management and business decisions
- In addition to overall balances, it is useful to explore also lower level of distribution hierarchy

  lower to medium level of aggregation
- Here, we will examine possibilities for statistically based surveillance of natural gas distribution at the *regulation station* level

Regulation station

- A hub serving a closed local distributional network for natural gas
- The local network contains hundreds to thousands of individual end-customers of small to medium consumption totals

  mostly households + small- and medium-size commercial customers
- Daily throughput recorded routinely

  Variation:

  - normal

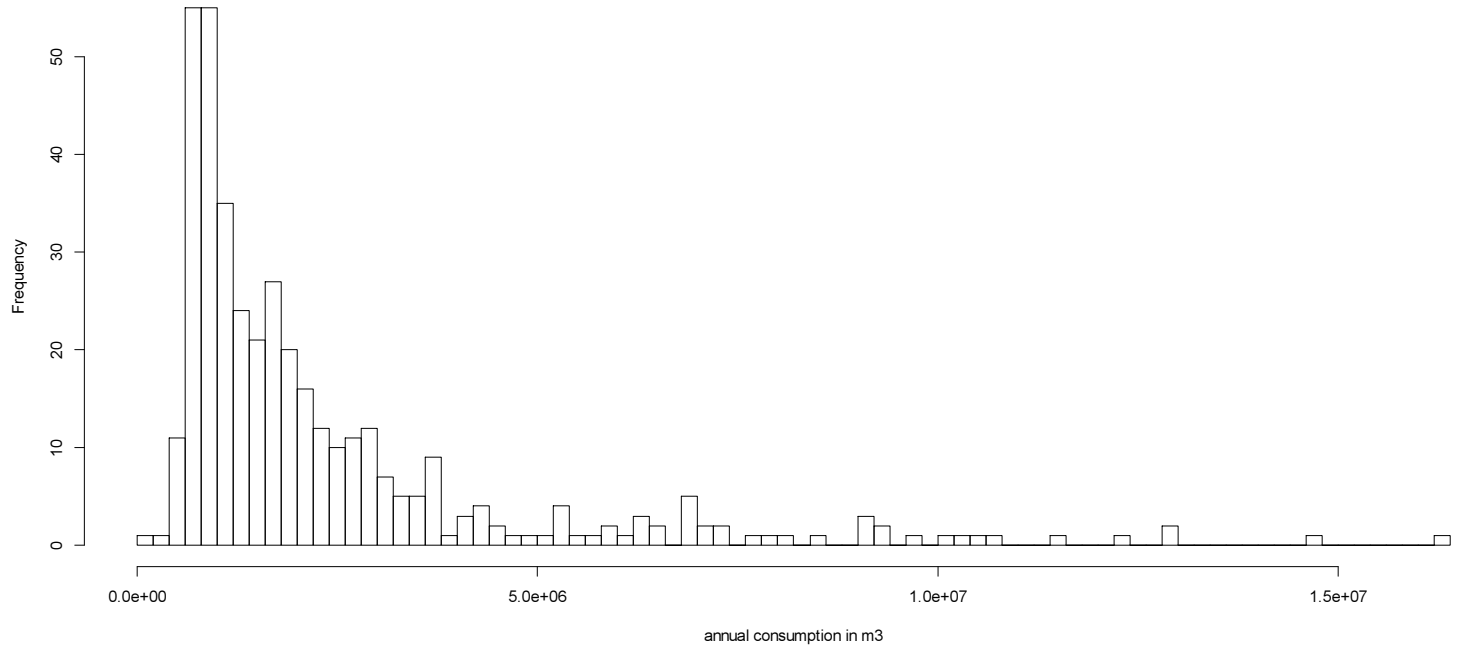
    ambient temperature, seasonality, etc.

  - irregular

    accidents, pressure decrease, thefts, unusual technological op.

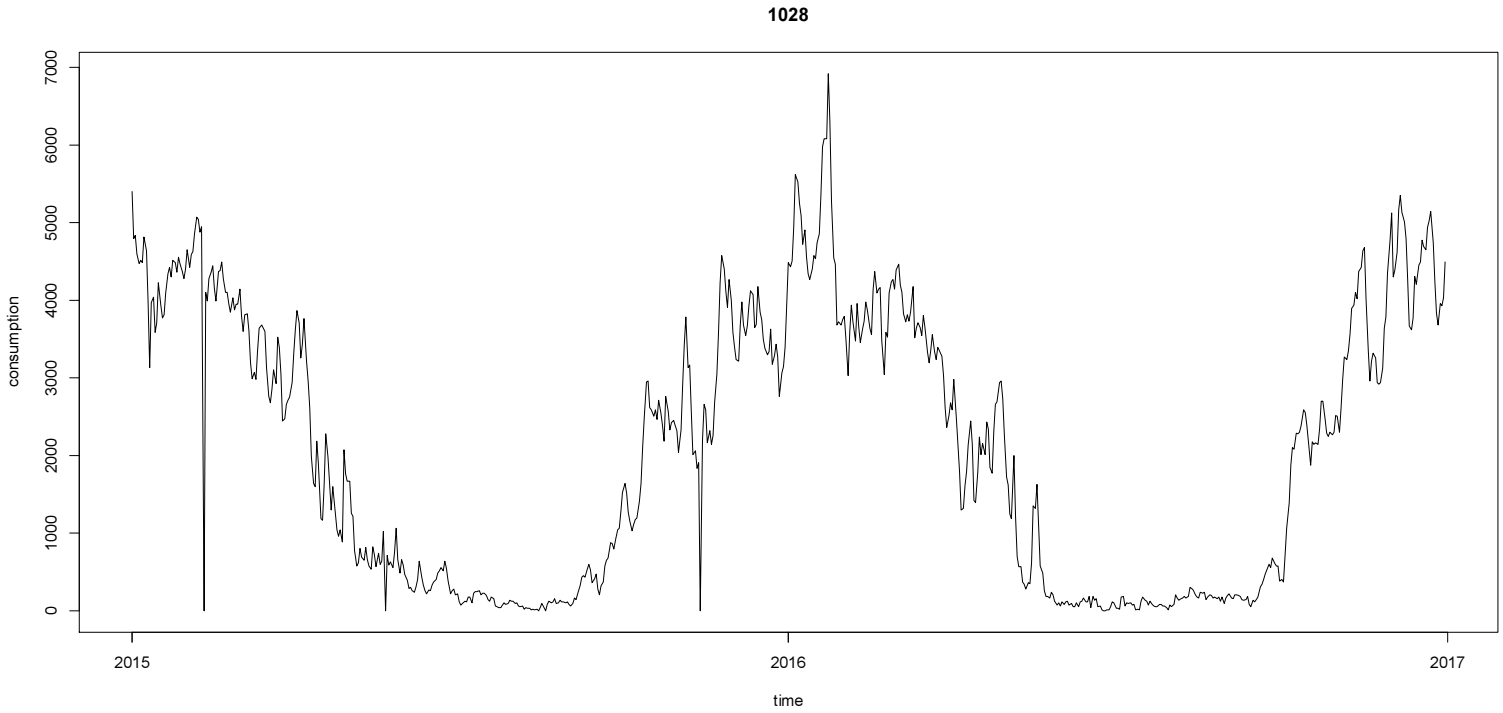
Regulation station annual consumption

distribution of a sample of stations in the Czech Republic

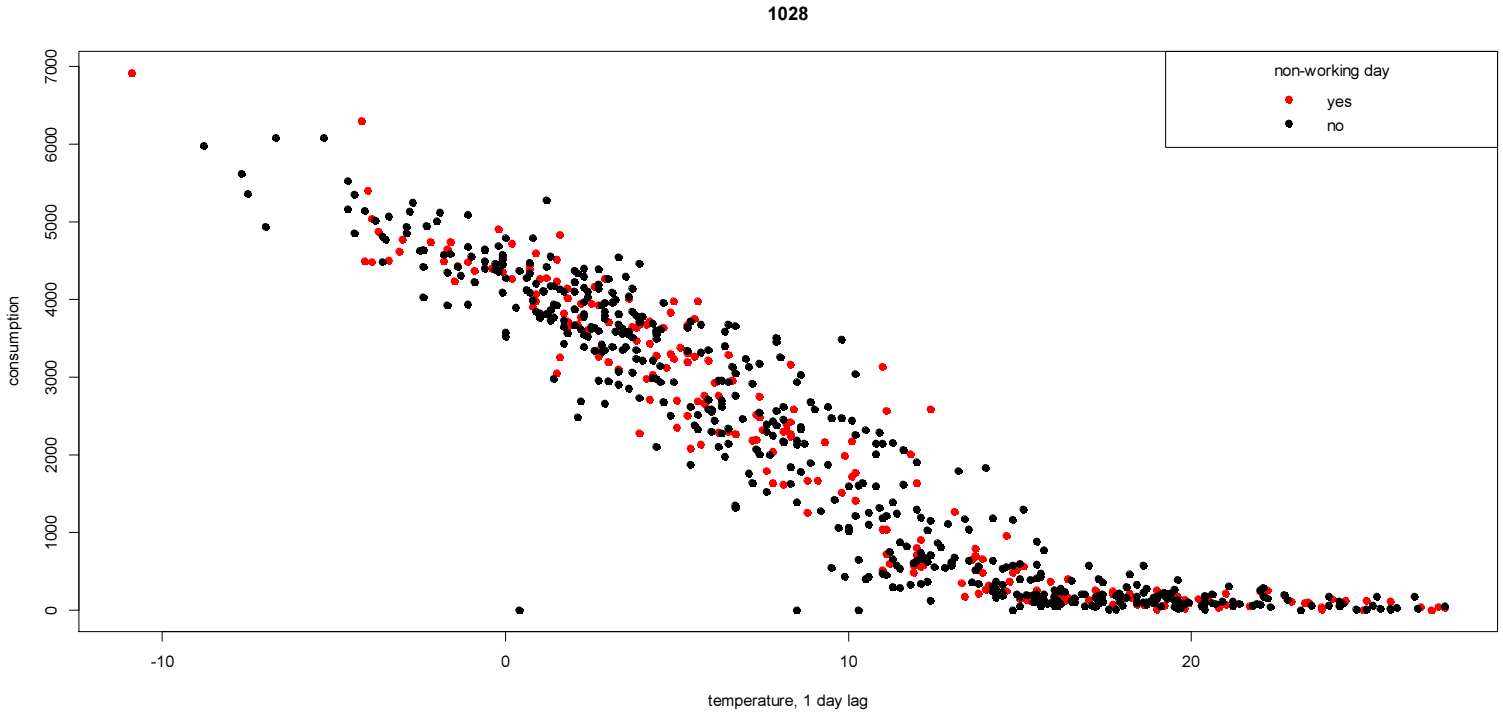


Consumption trajectory

(example of one regulation station)



Temperature is the main driver



GAM model, M1

$$Y_t = exp\left(\beta_0 + \beta_1 . log(Y_{t-1}) + s_T(T_{t-1}) + s_D(T_t - T_{t-1})\right) + \varepsilon_t$$

with

- Y_t consumption on day *t*
- T_t temperature on day *t*
- T_{t-1} temperature difference, current-previous day
- unknown smooth functions s_T ,s_D
- via spline basis expansion $s(x) = \sum_{k=1}^{K} a_k . b_k(x)$
  and penalization $a \sim N\left(0, \frac{1}{\lambda} . P^{-1}\right)$
- with heavy-tailed error distribution $\varepsilon_t \sim t(0, \sigma^2, \nu)$
  to be able to cope with occasional outliers

Model identification – parameter estimation

- Penalized likelihood

  GAM as a penalized GLM

  $$Loglik(\beta, a) + Pen(\lambda) = Loglik(\beta, a) + \sum_k \lambda_k . a'_k . V_k^{-1} . a_k$$

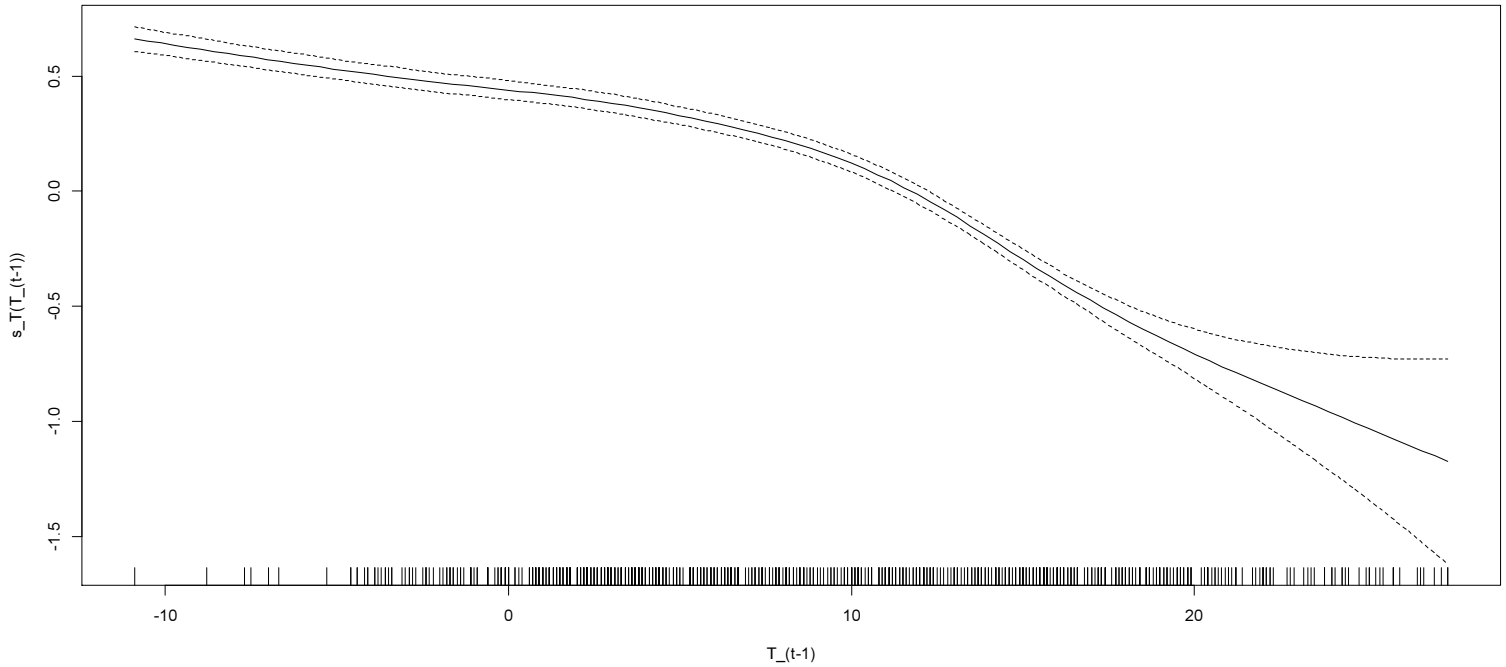
- Coeficients – IRLS (iterative least squares)

- λ_k crossvalidation,

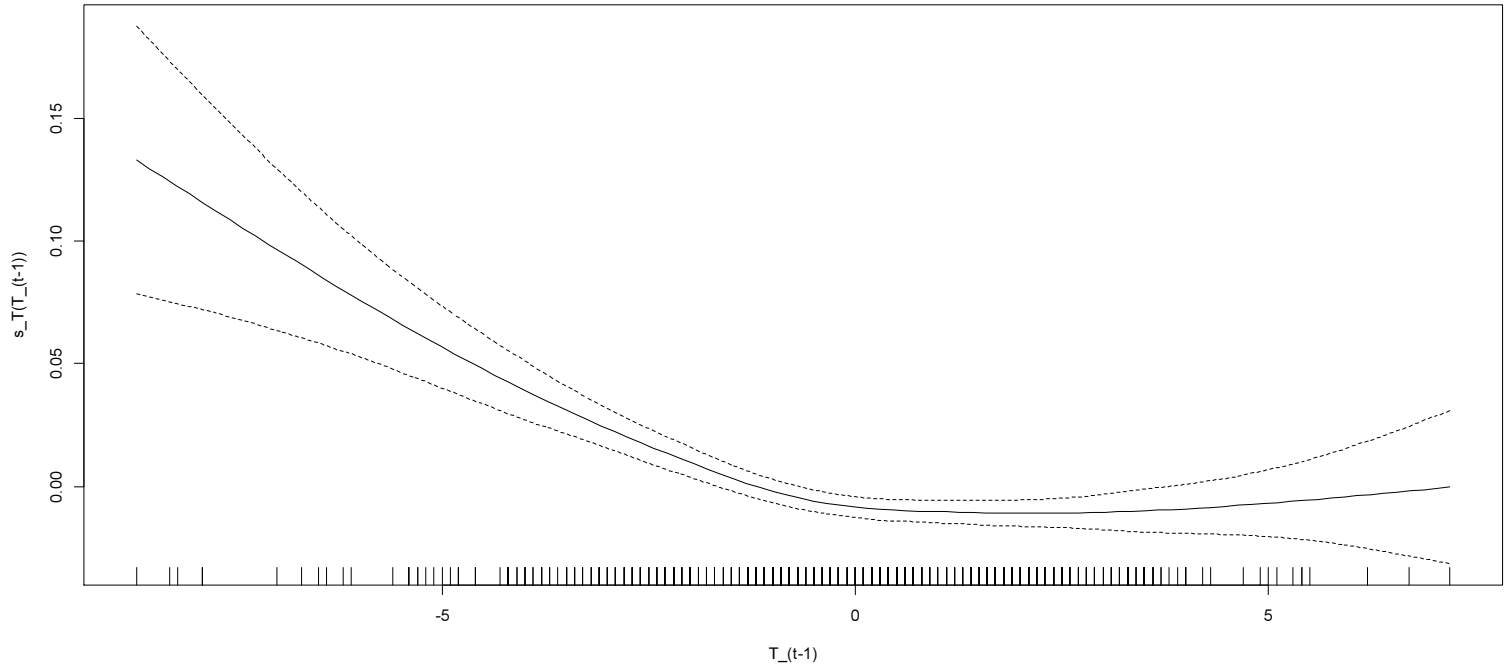
  generalized crossvalidation, REML

Example of smooth component extraction

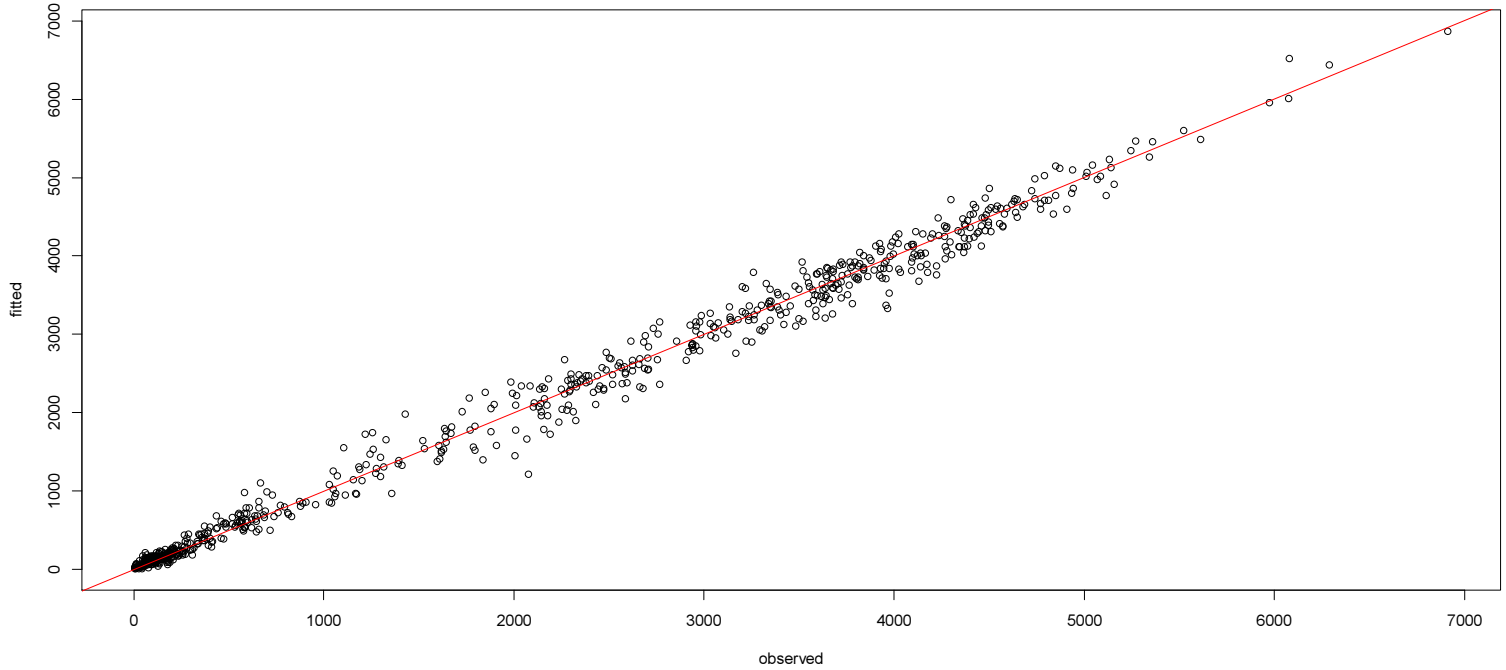
lag1 temperature effect



Temperature difference effect

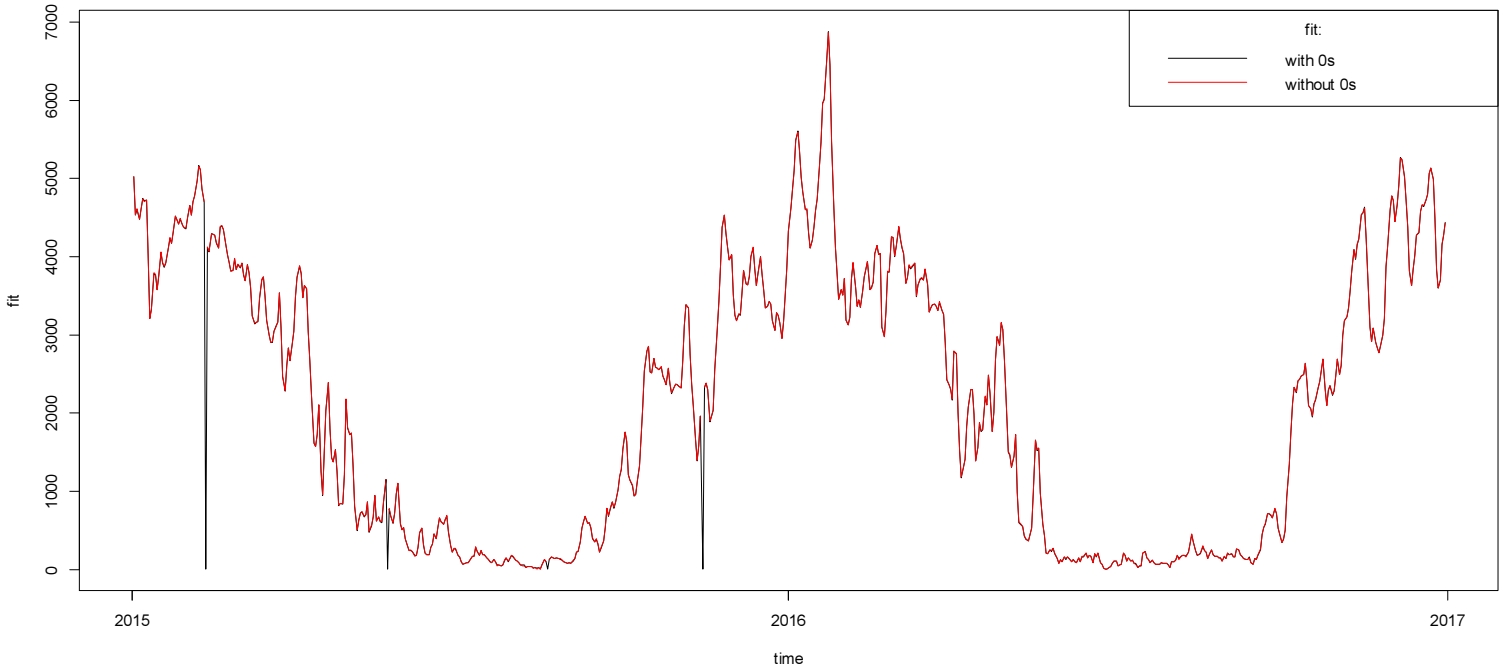


How does this simple model fit real data?



How is the model sensitive to outliers?

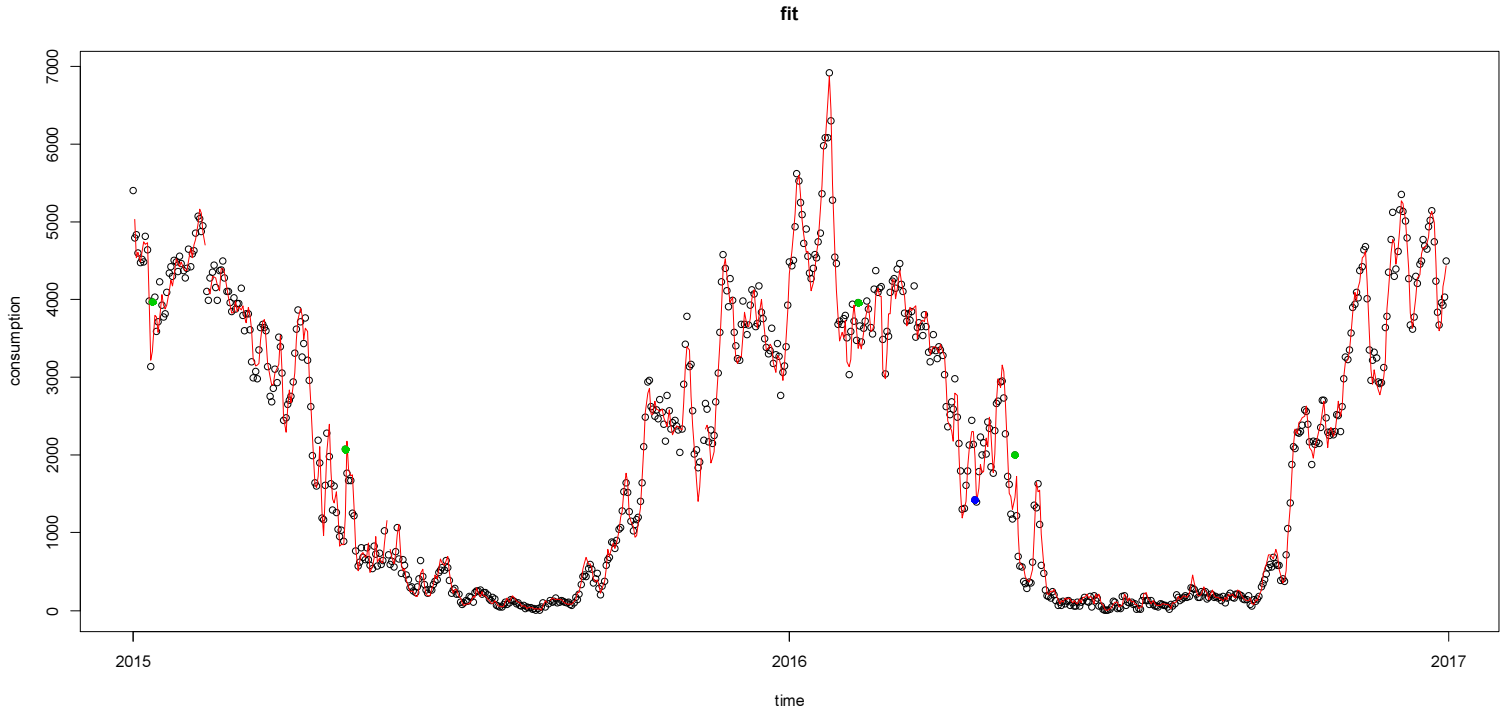
(few artificial zeros)



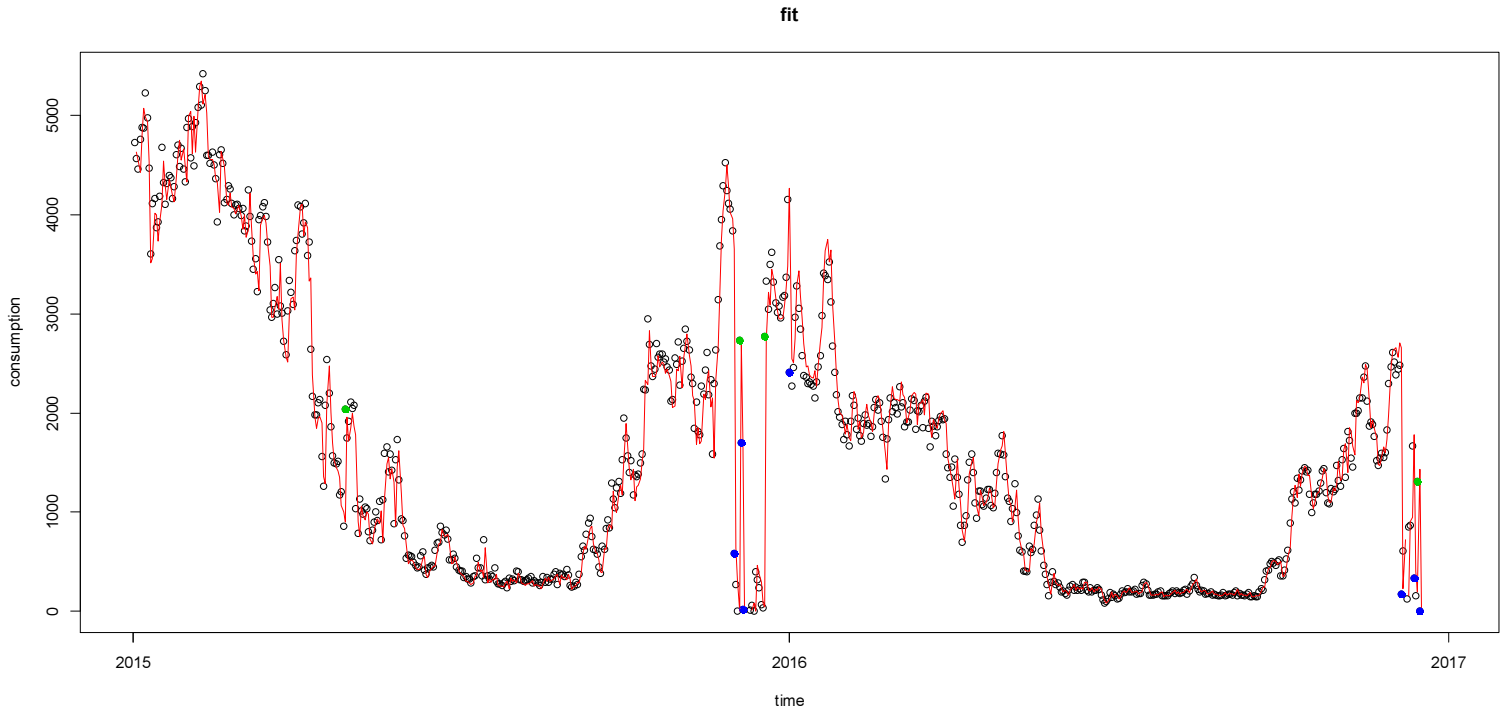
Practical task – scan for atypical days

e.g. Shewhart-like procedures on residuals

example of “OK” situation



Example of
a regulation station that is “not-OK”



Another practical task: interpolate to replace consumptions for erroneously recorded days

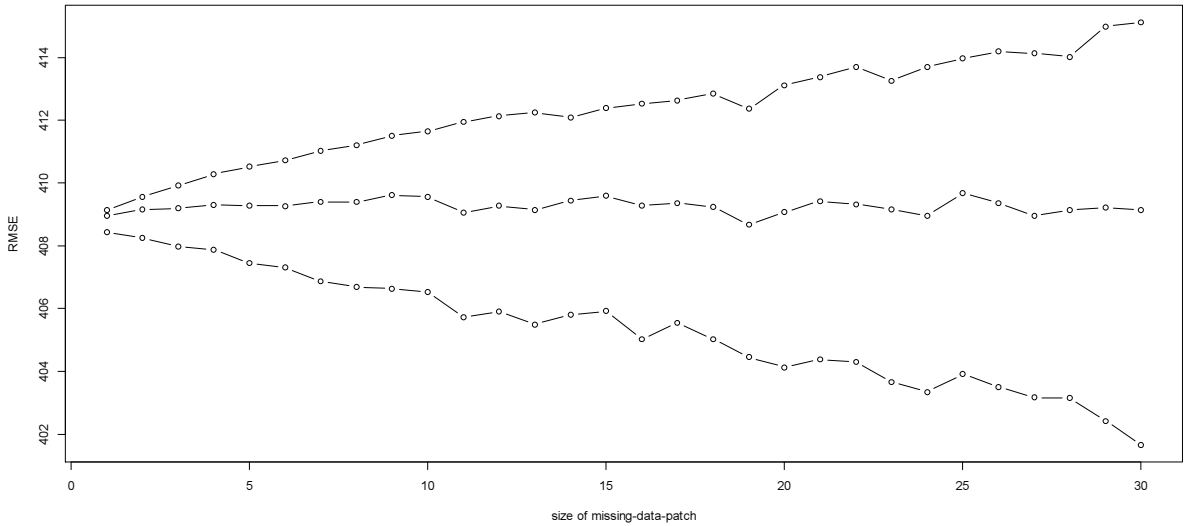
- In the routine records, occasional zero consumption days occur
- By cross-checking against service logs, network operator can find which of the 0's correspond to
  - real gas outage (correct zero)
  - flow measurement failure (erroneous zero, EZ)
- Erroneous zeros spoil computations of annual sums and hence of various balances needed for both accounting and other purposes
- Interpolate the model to replace EZ's

Quality of the EZ replacement

- Simulation experiment
- Use a simplified model (without AR term)
- Randomly place *r* clusters of *s* zeros
- Replicate 1000 times
- Compute statistical summaries
- E.g. bias, RMSE, MAE etc.

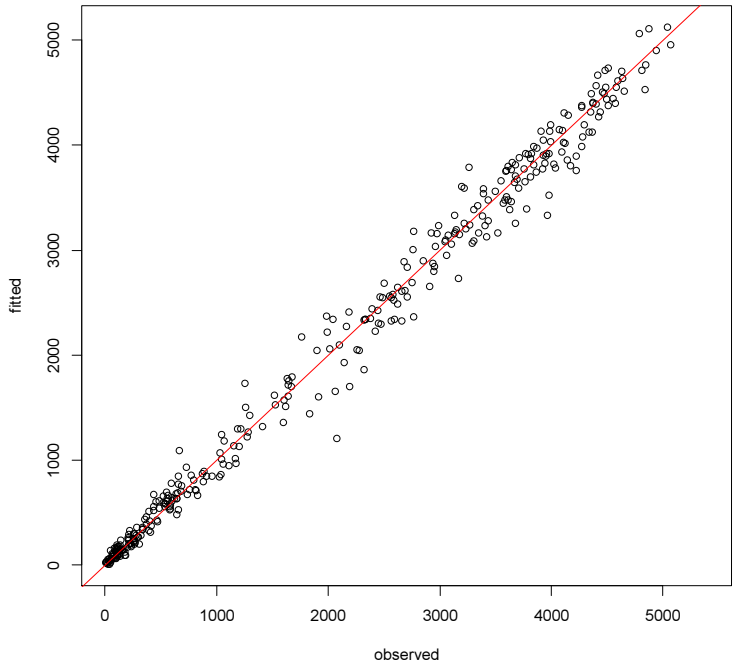
1 to 30 clusters of two Ezs

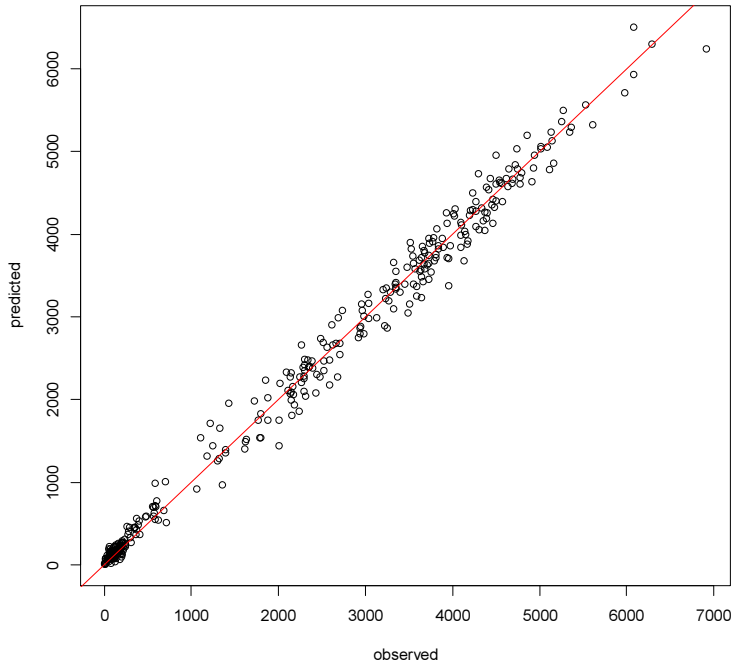
(25, 50, 75-th percentiles of RMSE for the interpolated points)



How does the model generalize to independent data?

(fit on 2015 and use on 2016)





Characteristic	Fit (2015)	Use (2016)
Bias	-6.9	7.1
RMSE	163.3	172.0
MAE	113.3	123.4

What about interaction between T and D?

- Because of the exponential form,
  M1 has temperature (T) and
  temperature-difference (D) terms that are:
  - interactive on the original scale
  - but additive on the log scale

- What about interaction even on the log scale?

Interactive GAM model, M2

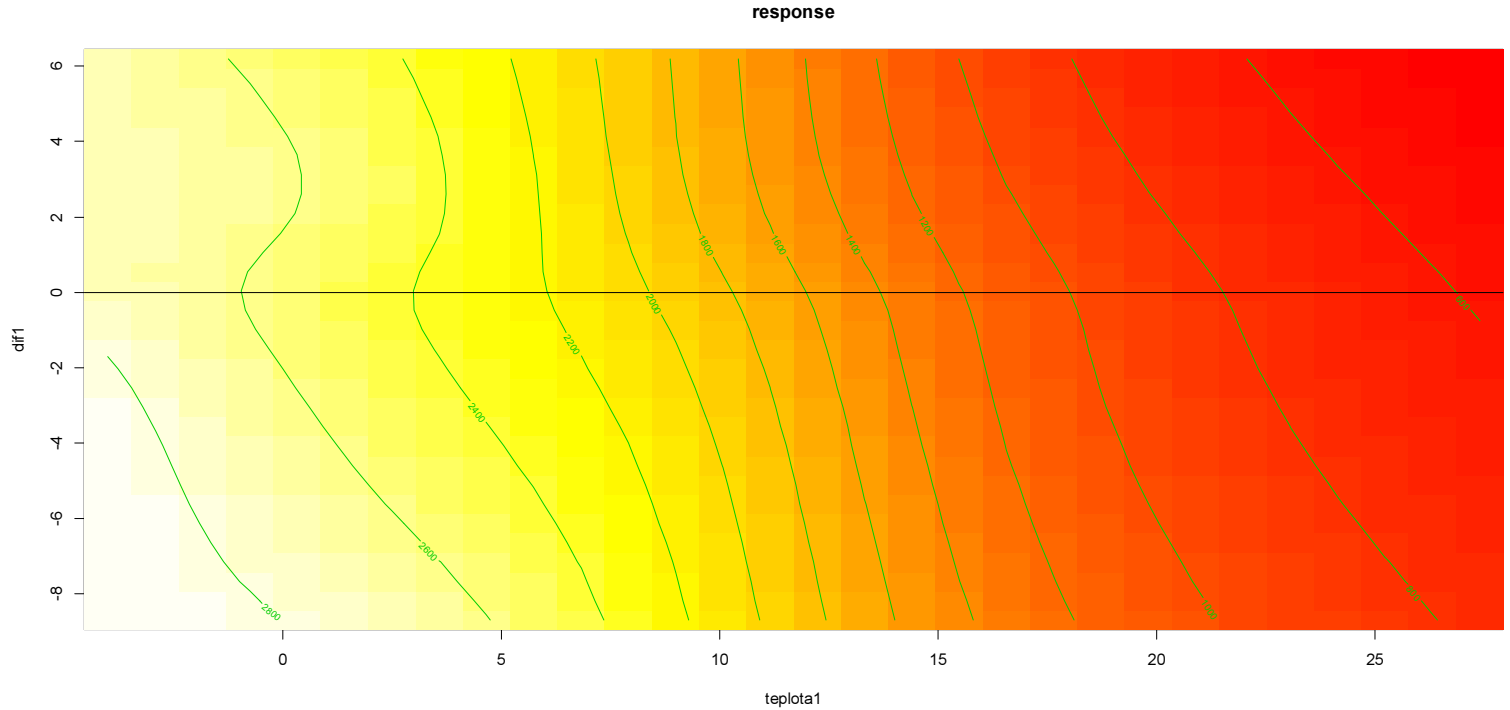
$$Y_t = exp\left(\beta_0 + \beta_1 . log(Y_{t-1}) + s_T(T_{t-1}) + s_D(T_t - T_{t-1}) + s_{TD}(T_{t-1}, T_t - T_{t-1})\right) + \varepsilon_t$$

with

- parsimonious interaction formulation
  (obviously not the full interaction in the ANOVA model sense)
- via penalized tensor-product spline term

We can:

- test the presence of interaction formally
- compare prediction performance of M1,M2 on independent data not used for fitting


response
dif1
6
4
2
0
-2
-4
-6
-8
0
5
10
15
20
25
teplota1
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800

A typical regulation station

Wald-like tests based on Wood (2013)

Term	EDF	Chi-square	p-value
S_T	3.6	718.6	<0.0001
S_D	1.0	17.4	<0.0001
S_{TD}	6.3	49.2	<0.0001

Performance on independent data		
Characteristic	M1	M2
Bias	19.6	4.1
RMSE	172.0	173.1

Interactive model is not overwhelmingly better from practical perspective
even though it is favored by formal tests

Selection of atypical regulation stations

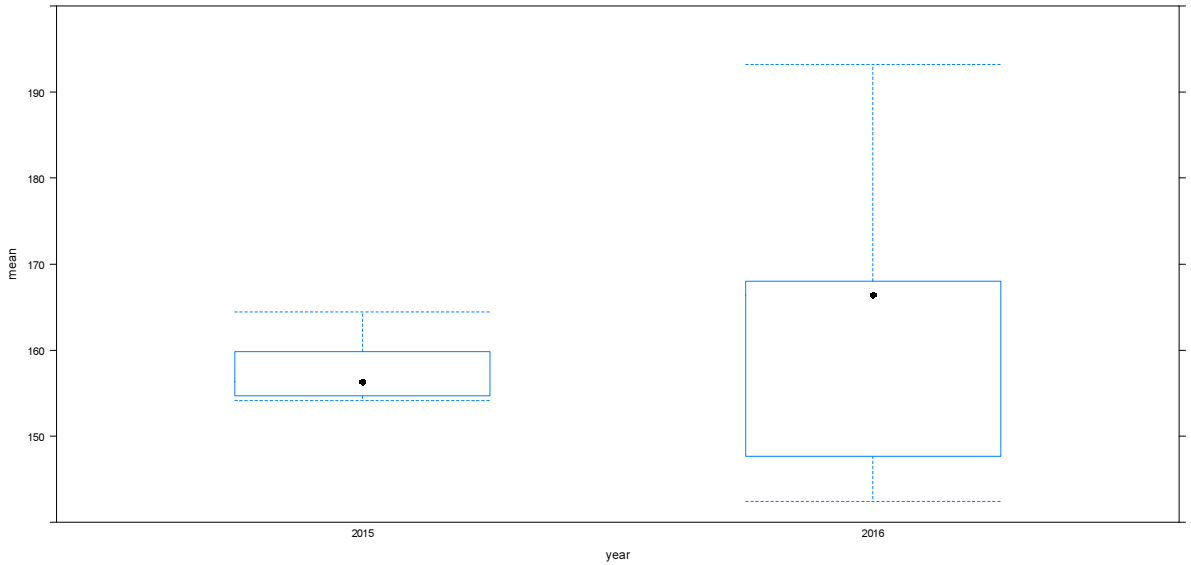
- The M1 model is useful both for atypical **days**
  (compared to the consumption-to-temperature relation typical for a given regulation station)

- But it can be useful for picking **regulation stations** that show atypical consumption-to-temperature relationship
  (compared to the behavior of typical regulation station)

- If one thinks about a distribution of annual consumption into daily consumptions, $p_d = \frac{Y_d}{\sum_{l=1}^{365} Y_l}$

- the fit of M1 model can be used to estimate $\widehat{p_d} = \frac{\widehat{Y_d}}{\sum_{l=1}^{365} \widehat{Y_l}}$ and compute:
  - moments $\widehat{M_1} = \sum_{d=1}^{365} \hat{p}_d . d$, $\widehat{M_2} = \sum_{d=1}^{365} \hat{p}_d . (d - \widehat{M_1})^2$ ...
  - entropy $\widehat{En} = -\sum_{d=1}^{365} \hat{p}_d . log_2(\hat{p}_d)$,
  - etc.

1st moment,
distribution over 31 regulation stations,
changes between years



Time-varying coefficient model, TVAR

- Linear, but time-varying temperature model formulation is alternative
- to the previous nonlinear temperature model with time-invariant structure

- The linear temperature coefficient trajectory is relatively easy
  - to interpret (local temperature sensitivity)
  - and to compare among regulation stations

TVAR as a GAM model

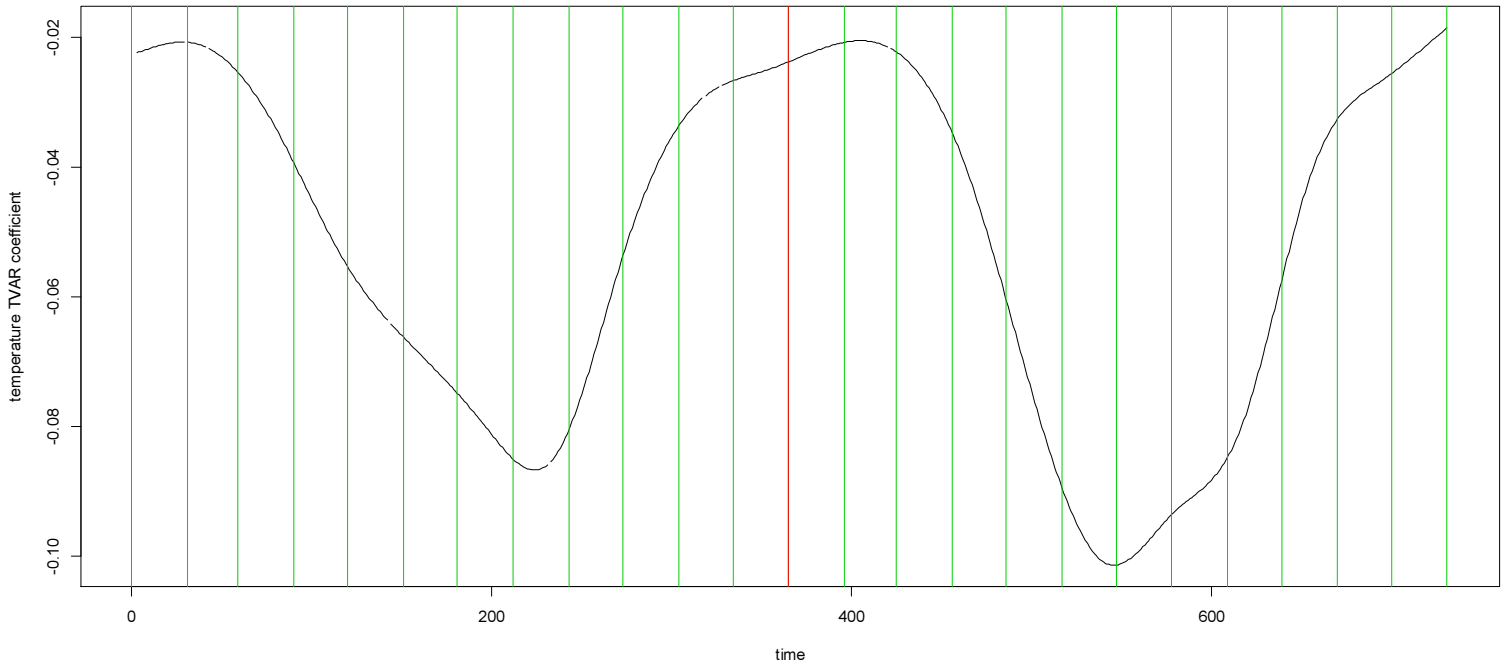
$$Y_t = exp(\beta_0 + \beta_1 . log(Y_{t-1}) + \beta_t . T_{t-1}) + \varepsilon_t$$

with coefficient trajectory
modeled via penalized spline

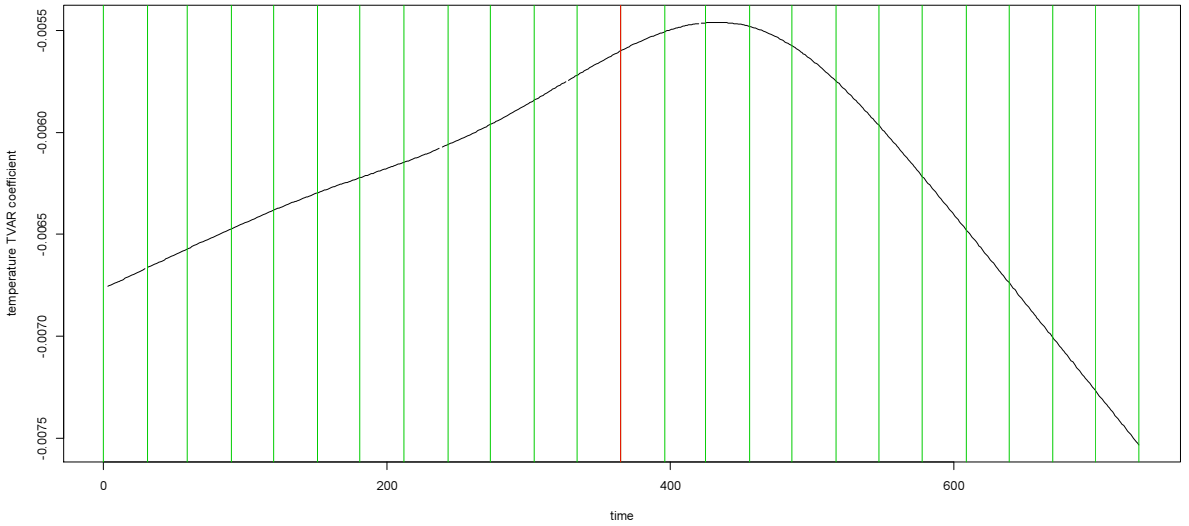
$$\beta_t = s(t) = \sum_{k=1}^{K} a_k . b_k(t)$$

$$a \sim N\left(0, \frac{1}{\lambda} . P^{-1}\right)$$

Example of a typical coefficient trajectory



Atypical behavior …



Alternative (state-space) formulation of a TVAR model

- Observation equation: $Y_t = \mu_{1,t} + \beta_t . T_{t-1} + \epsilon_t$

- State equations: $\mu_{1,t} = \mu_{1,t-1} + \mu_{2,t-1}$

  $\mu_{2,t} = \mu_{2,t-1} + \nu_t$

  $\beta_t = \beta_{t-1} + \omega_t$

- With $\epsilon_t \sim N(0, \sigma_\varepsilon^2)$

  $\nu_t \sim N(0, \sigma_\nu^2)$

  $\beta_t \sim N(0, \sigma_\beta^2)$

- Initialization from overdispersed $\begin{pmatrix} \mu_{1,0} \\ 0 \\ \beta_0 \end{pmatrix}$

  obtained from OLS

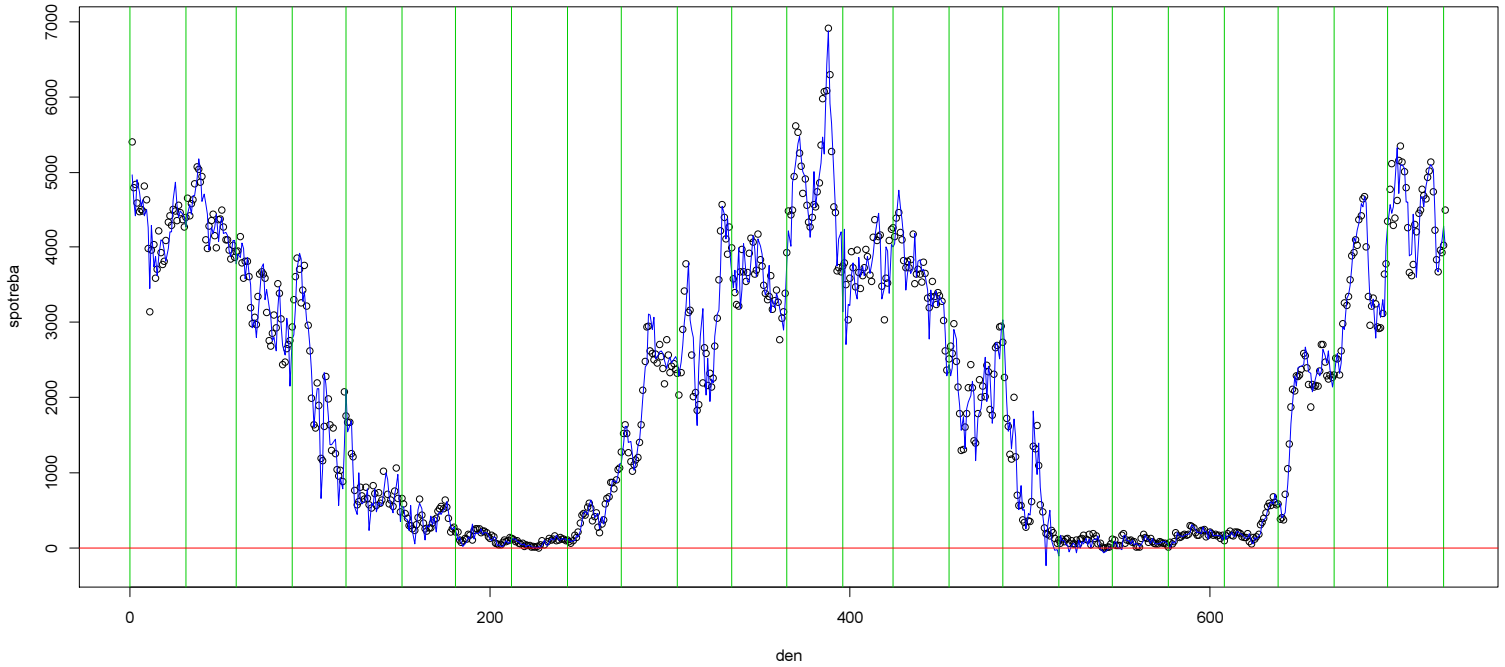
Model identification and state estimation

- Model identification (estimation of the structural parameters) can be achieved very effectively via prediction error decomposition (Harvey, 2014) and Kalman filter
- Once the state parameters are fixed, the state estimation (filtering, prediction) is achieved easily via Kalman filter

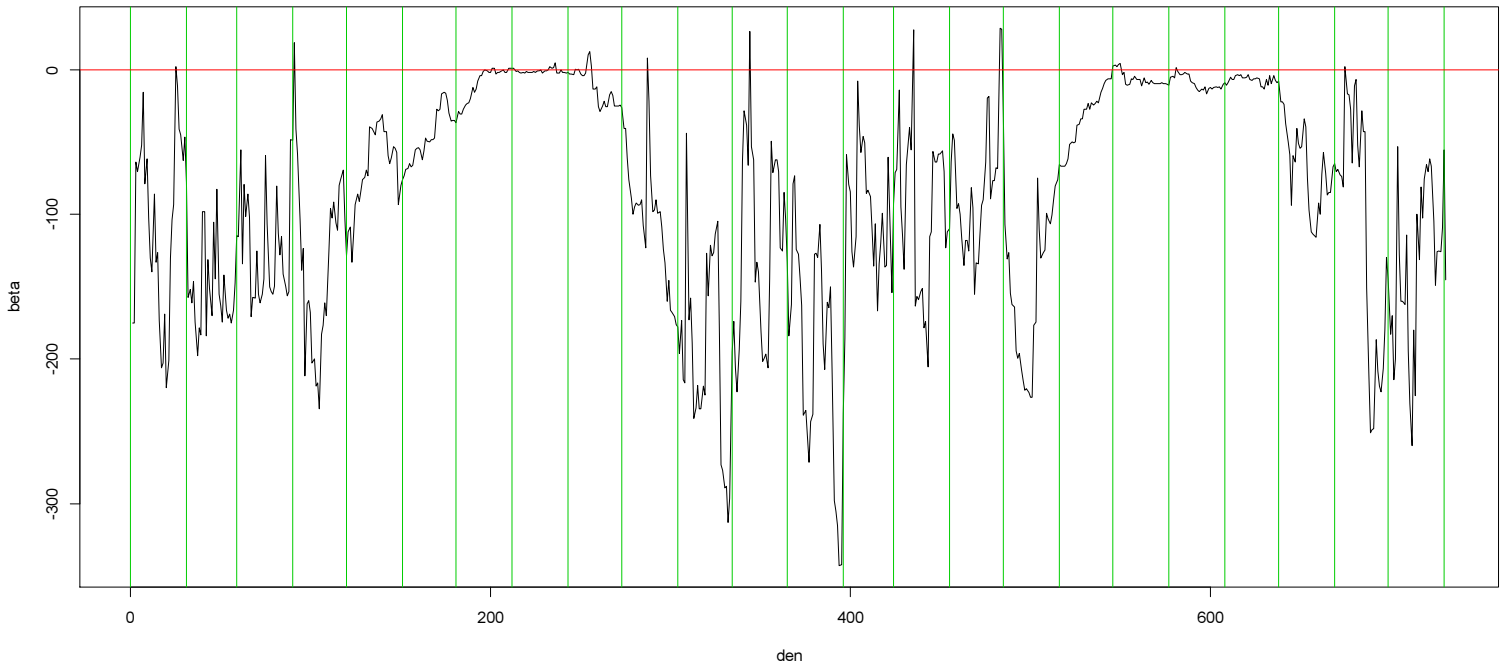
- The two steps together can be viewed as a particular example of the EM algorithm

Consumption trajectory

1-step ahead predictions



Temperature coefficient ($(\hat{\beta}_t)_{t=1}^{731}$) trajectory



Further look at the temperature dependence

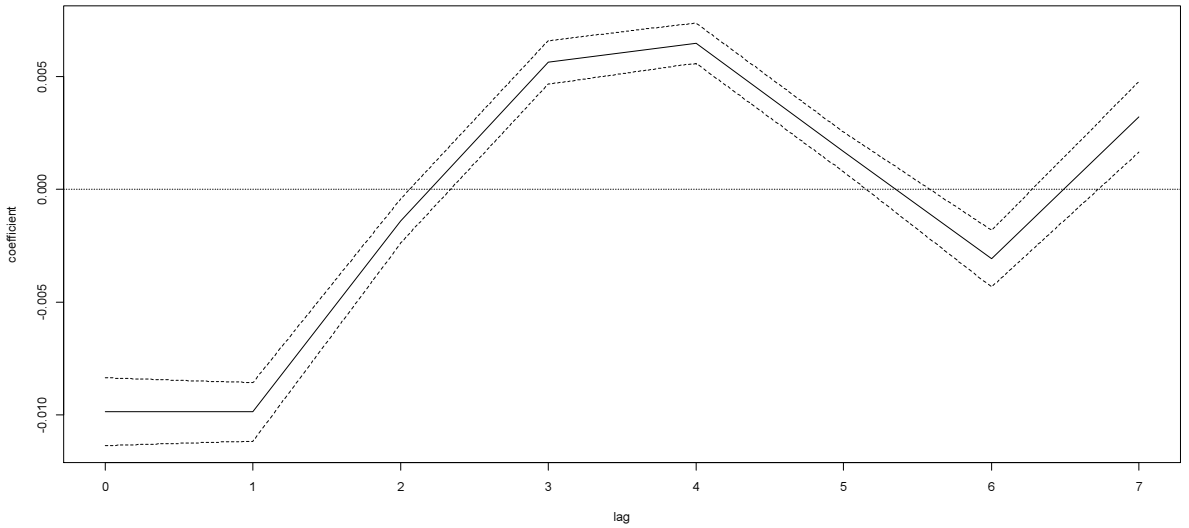
- So far, we have used models (GAM, state-space) with only one or two temperature lag (based on previous AIC selection)
- Is this all? Can we gain more insight by considering more lags?
- Obviously, if we consider several temperature lags simultaneously, we face collinearity which complicates or even precludes direct approach

Almon lag approach

- Almon (1965) approach regularize a multi-lag model component, say $\sum_{l=0}^{L} \delta_l . T_{t-l}$
- by imposing a constraint on the lag coefficients originally, it is lower order polynomial constraint $\delta_l = \sum_{k=1}^{K} \alpha_k . l^k$ with $K < L$ (choose *K* e.g. via AIC)
- the main advantage/beauty is that this is a linear transformation of parameters, $X.\delta = X.W.\alpha = \tilde{X}.\alpha$

  the linear for of the predictor is preserved and so is the GAM model class

Almon polynomial (L=7, k=4) estimate of the time-invariant linear temperature filter



Form of the Almon-type restriction

- Previous estimate uses k with the best AIC
- Even that does not look entirely realistic

  heavy weight on the zero lag does not correspond to the real temperature response behavior

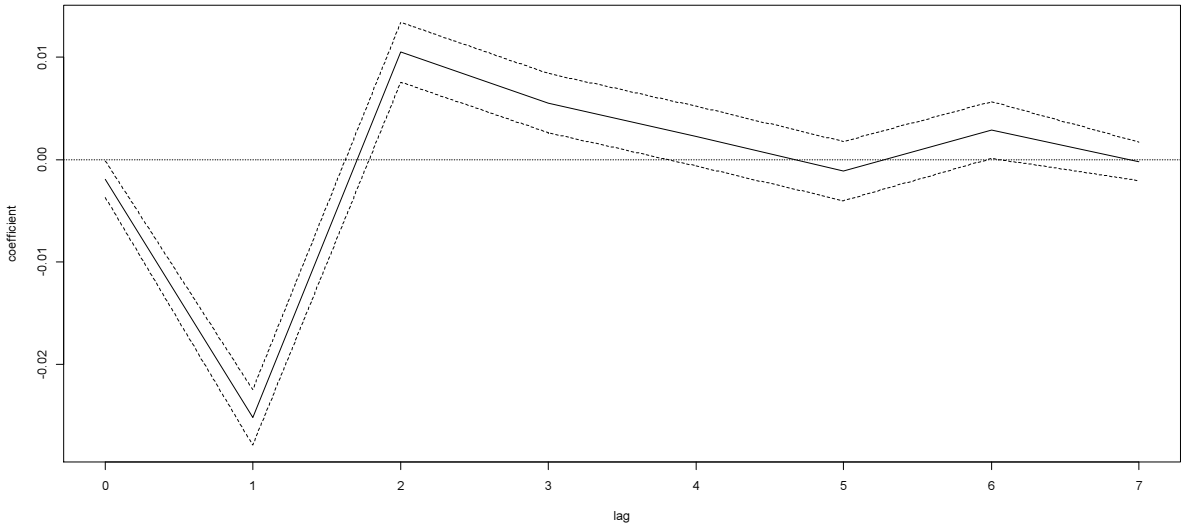
  in reality, there is a substantial difference between lag0 and lag1 response, in favor of lag1

- Polynomial restriction is simply not flexible enough to describe the temperature correctly

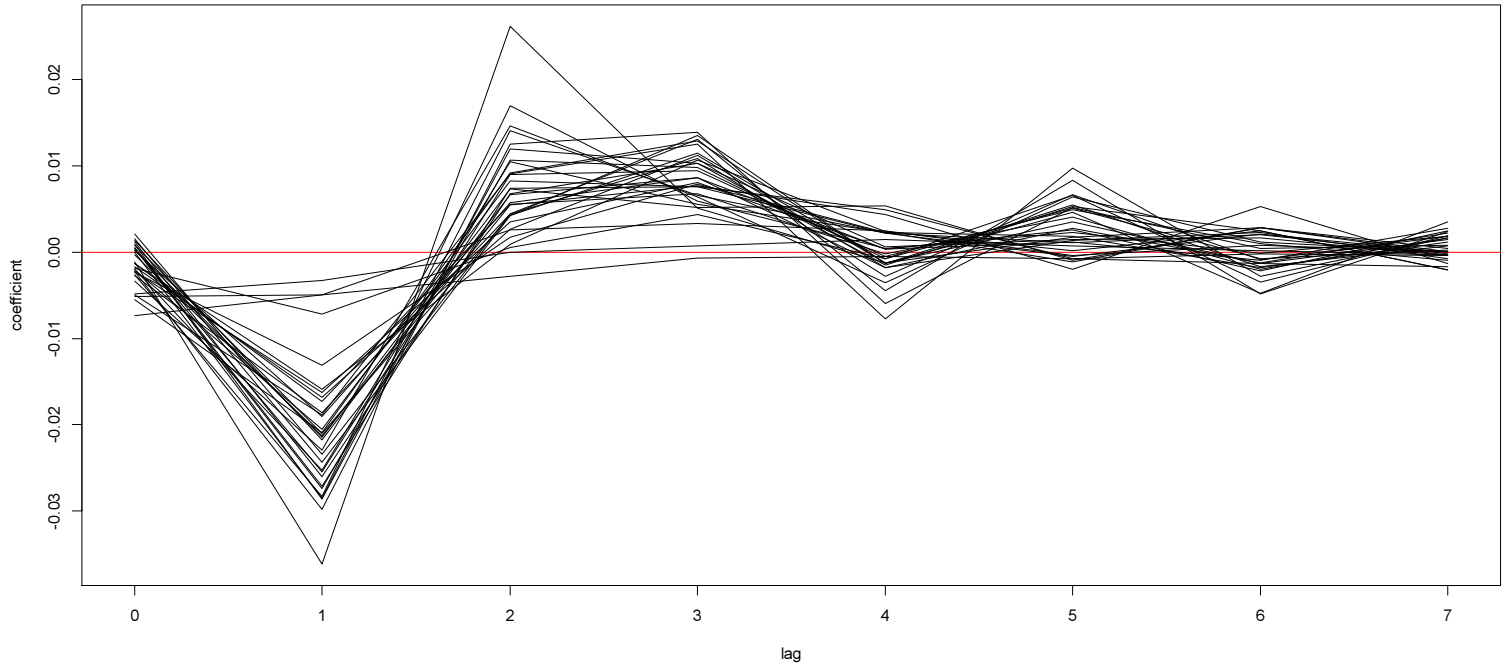
- We can generalize a bit the approach and use a more general basis expansion restriction $\delta_l = \sum_{k=1}^{K} \alpha_k . b_k(l)$

  B-spline (with penalized coefficients), in particular

Penalized B-spline restriction



Pattern in many different regulation stations



Morale

Shape of the time-invariant filter obtained from the more flexible model suggests that:

- picture obtained from the polynomial restriction is distorted and shifted to higher lags

In fact:

- lag0 contribution is small
- main driver is the lag1 (*as expected*)
- the temperature response considers a contrast between lag1 and average of lag2, lag3

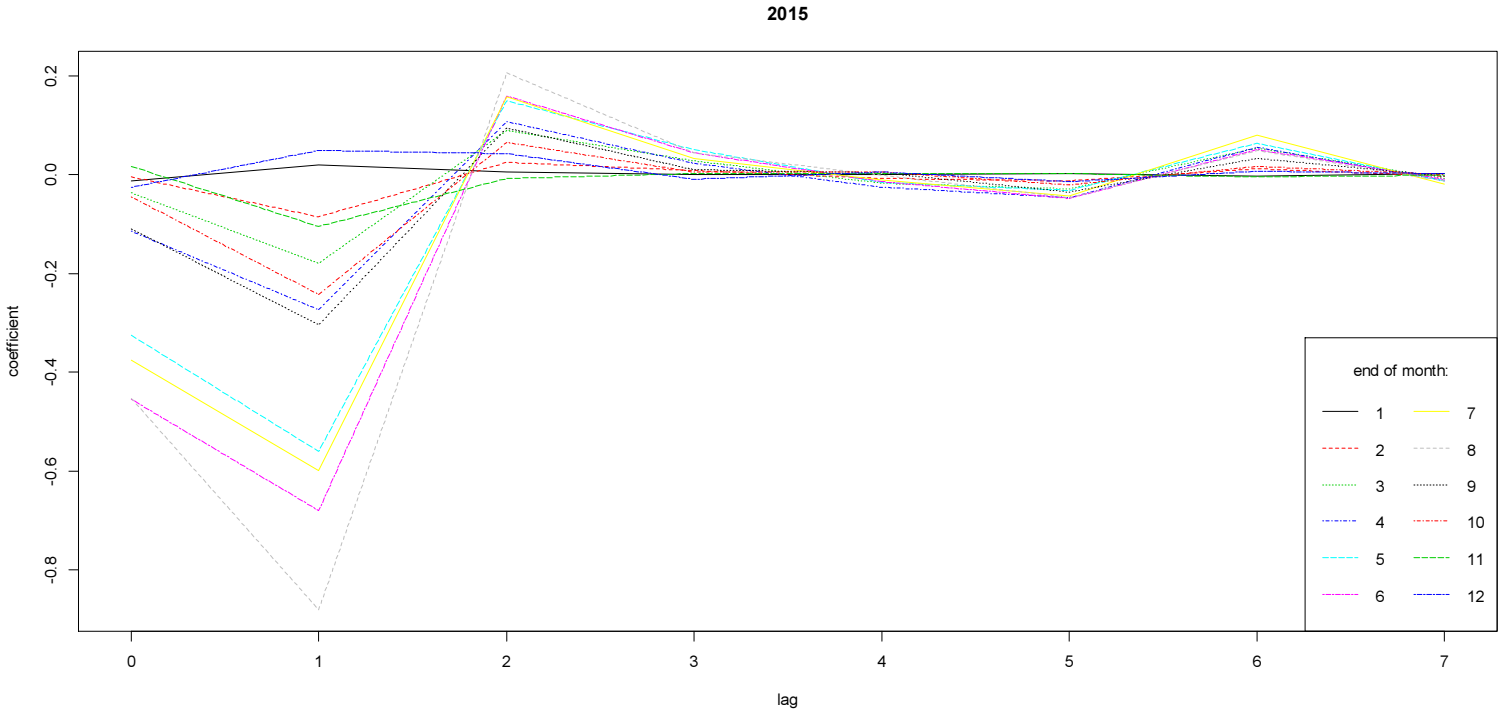
  this is different from the
  traditional lag1 only or (lag1),(lag0-lag1) approaches
  favored by practitioners

Deeper look: time-varying version of the generalized Almon model with penalized B-spline restriction

- Distributed lag sub-model $\sum_{l=0}^{L} \delta_{l,t} . T_{t-l}$
- Coefficient restriction $\delta_{l,t} = \sum_{k=1}^{K} \alpha_{k,t} . b_k(l)$
  B-spline
- Time-varying α-parameter $\alpha_{k,t} = \sum_{m=1}^{M} \gamma_m . \tilde{b}_m(t)$
  cubic spline or B-spline
- This is still linear in the γ‘s

One regulation station, 2015

trajectories run in days, here we show just month end days



Conclusions, I

- Natural gas distribution monitoring at a low level of aggregation presents interesting practical and methodological challenges

- Several statistical approaches were presented for finding atypical days within one regulation station and atypical stations

  nonlinear models (Shewhart-type procedures, moments, entropy)

  time-varying coefficient models

Conclusions, II

- Main driver of the consumption is obviously the ambient temperature

- Temperature effects are non entirely trivial

  nonlinear, time-varying or both

- Temperature effect is not concentrated to one lag but invariably distributed over several lags

  time-invariant/time-varying filter (linear on appropriate scale)
  with sometime surprising consequences

  they can be studied e.g. in Almon-type models in considerable detail