Modeling dynamics and short term prediction of complex processes

M. Brabec

Institute of Computer Science, Czech Academy of Sciences
mbrabec@cs.cas.cz

ENBIS15 post-confrence workshop:
“Modeling smart grids - Challenge for Stochastic and Optimization”
Praha, September 10, 2015

Energy-distribution/production-related data as a challenge

- The real-time data from energy industry are invariably complex and large
  - complex underlying processes
  - complicated hierarchical (longitudinal) structure
  - measurement errors are often not negligible
  - measurements might reflect the desired quantity only indirectly (only nontrivial functionals of the underlying process observable)
  - when pooling data from several sources, inconsistencies arise

- Substantial challenge for the analytical methods
  - simple statistical and data-analytic methods might easily yield confusing and even misleading results

Tools to meet the challenge: modern statistical methods

- Modern statistical modeling tools offer a solution
  - several classes of models usable in energy-industry context, here we will concentrate on the semi-parametric regression methods –
  - GLM, GAM and extensions (GAMLSS) with smooth (especially penalized spline) components
- Need flexible and structured approach
  - to cover non-standard situations
  - to have modular structure (implementation, checks, serviceability)
  - at least partially interpretable components (for model realism and qualitative control of its output) vs. black-box approach
  - fruitful hybrid of empirical (purely statistical) and theoretical models

Dynamic approach

- Many practical tasks in energy-distribution networks are related to prediction
  - predictions, their uncertainty (or full predictive distribution) are needed for decision-making
  - e.g. as formalized, economically motivated loss function optimization
- Markov-chain-based statistical models can provide framework for practical forecasting
  - parsimonious, hence efficient for parameter estimation and prediction
  - relatively easy implementation
  - can utilize endogeneous and exogeneous inputs
  - can provide uncertainty assessment in a rather unified way

Will illustrate the approach at examples of several energy applications

- Photovoltaic production
  - empirical+theoretical models fusion for prediction, calibration of the NWPs
- Natural gas consumption modeling
  - consumption trends from the space-time viewpoint
  - SLP profile development for official use
  - Bayesian calibration of the SLP using total customer pool info
- Wind energy
  - prediction of the wind-farm output
- Detailed analyses from Energy-meteorology
  - cloud dynamics from high-frequency data
  - clouds in motion, spatio-temporal field prediction

A typical semi-parametric statistical framework useful for modeling and prediction

GAM (Generalized Additive Model)

$$Y_{ijt} \sim Dist\left(\mu_{ijt}\right)$$

$$link\left(\mu_{ijt}\right) = \beta_0 + \sum_{p=1}^{P} \beta_p . X_{p,ijt} + b_i + \sum_{q}^{Q} s_q\left(Z_{q,ijt}\right) + s_{spat}\left(W_{1,ijt}, W_{2,ijt}\right)$$

$$b_i \sim N\left(0, \sigma^2\right), \text{ iid across } i$$

$$s_q(x) = \sum_{m=1}^{M_q} a_{q,m} B_{q,m}(x), \; q = 1, \ldots, Q$$

$$s_{spat}(x, y) = \sum_{m=1}^{M_{spat}} \sum_{n=1}^{M_{spat}} a_{spat,mn} B_{spat,mn}(x, y)$$

$$\underline{a}_q \sim N\left(\underline{0}, \lambda_q^{-1} . V_q\right), \; q = 1, \ldots, Q$$

$$\underline{a}_{spat} \sim N\left(\underline{0}, \lambda_{spat}^{-1} . V_{spat}\right)$$

Now available in many various SW, notably in R.

_ A simple model for wind dynamics, is it worthwhile to bother with GAM?

- Toy example: British hourly windspeed data from more than 2 years of one measurement location
- AR $Y_t = \beta_0 + \sum_{l=0}^{L} \beta_l . Y_{t-l} + \varepsilon_t$

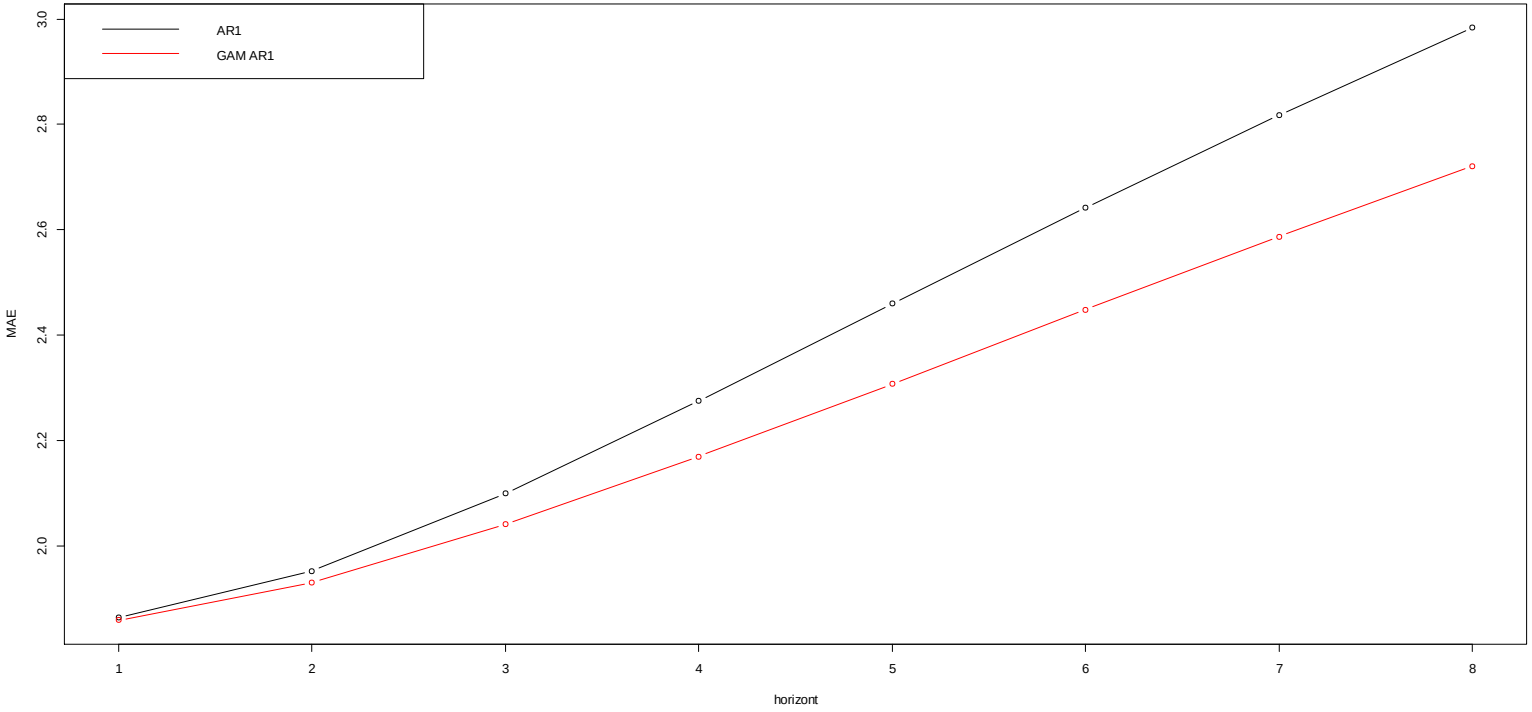
  $\varepsilon_t \sim N(0, \sigma^2)$, iid

  - AR(1) selected $Y_t = \beta_0 + \beta_1 . Y_{t-1} + \varepsilon_t$
- GAMAR

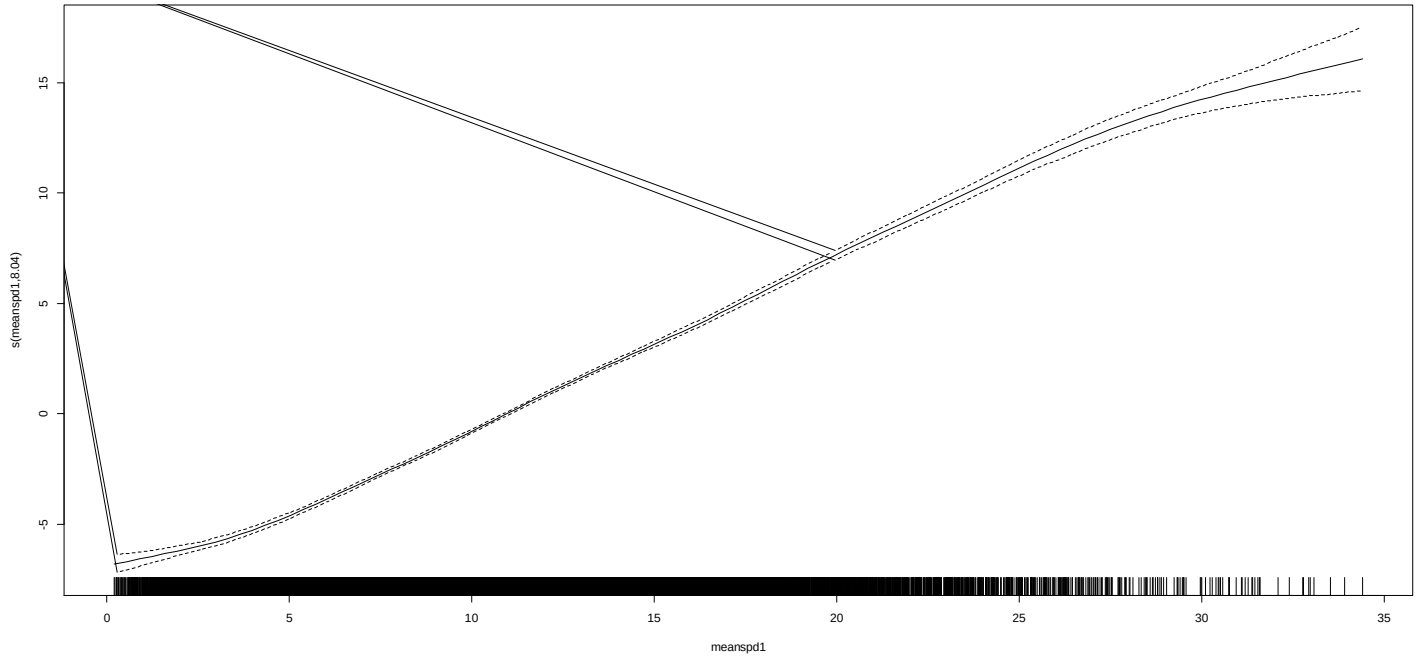
  - nonlinear AR, estimated nonparametrically

  $Y_t = \alpha + s(Y_{t-1}) + \varepsilon_t$

AR1 and GAMAR1



What is the reason for the success of the nonlinear model?



The story is very similar when we model and predict a windfarm output for short horizons

- Krystofovy Hamry windpark
  - about 97 GWh electricity produced in 2009, about 21 turbines

- A bit larger nonlinear autoregressive model was selected by AIC for the farm energy output model: $Y_t = \beta_0 + s_1(x_{t-1}) + s_{dif}(x_{t-2} - x_{t-1}) + \beta_1 y_{t-1} + \beta_2(y_{t-2} - y_{t-1}) + \varepsilon_t$

- Similar sigmoidal shape of the smooth functions s_1, s_{dif}

- RMSE for 1h prediction
  - is 243.9 (vs. 302 for GAMAR1)

Once we have a well identified model …

It can be used:

- to analyze the structure of the problem
  - test hypotheses about parameters and functional parameters

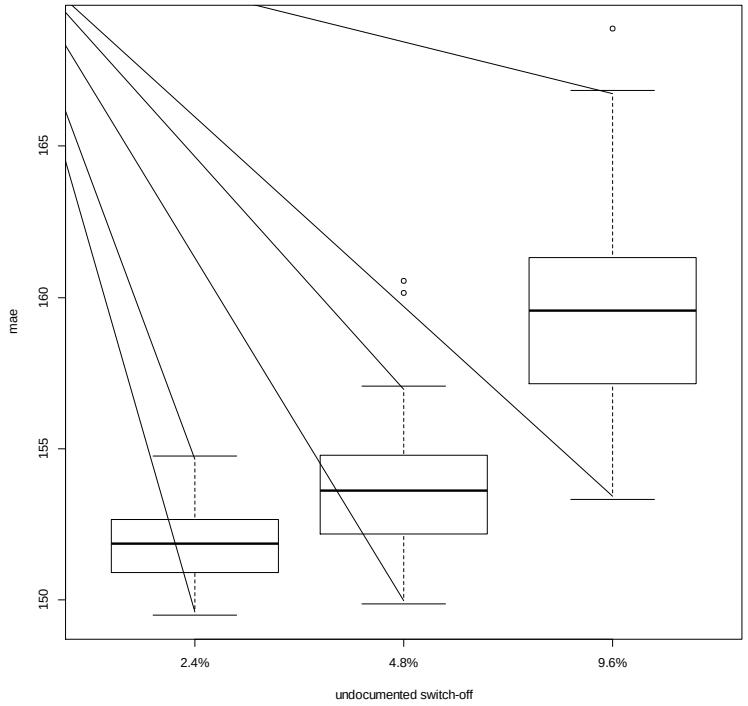
- as a basis of forecasting procedure
  - upon SW implementation of model estimated on the training data
  - possibly subject to periodic updates

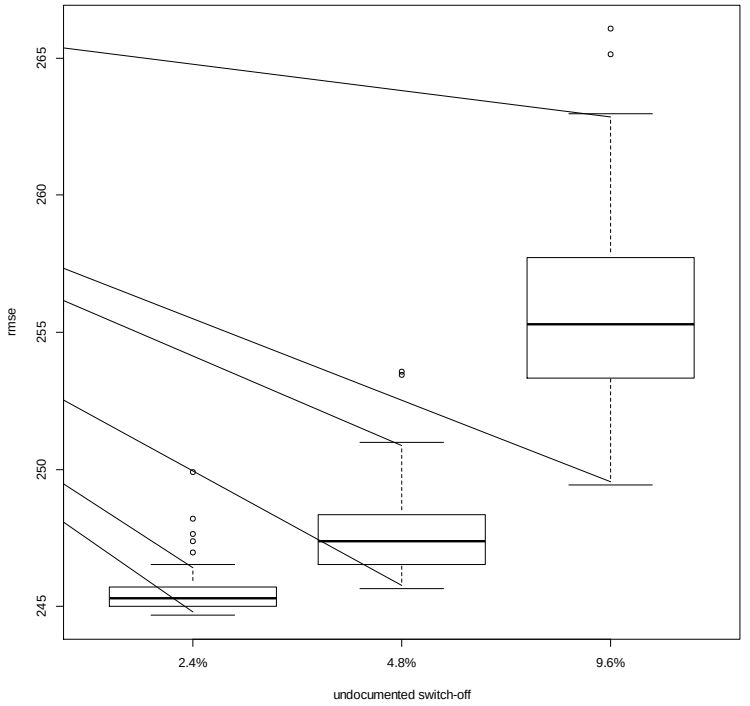
- for simulations
  - aiming at assessment of (otherwise difficult) tasks of substantial practical interest

Example

- How much it matters if we have windfarm data of lower quality?
  - undocumented switch-off
  - maintenance or security relate
  - in plain language: one column (with the switch-off indicator) is missing in the database
- Certainly, it makes the prediction results worse
  - if the model is trained without taking the indicator into account
- But is it worthwhile to pay more money for improving them?

Prediction performance
net effect of practical value to be compared to the price of better data

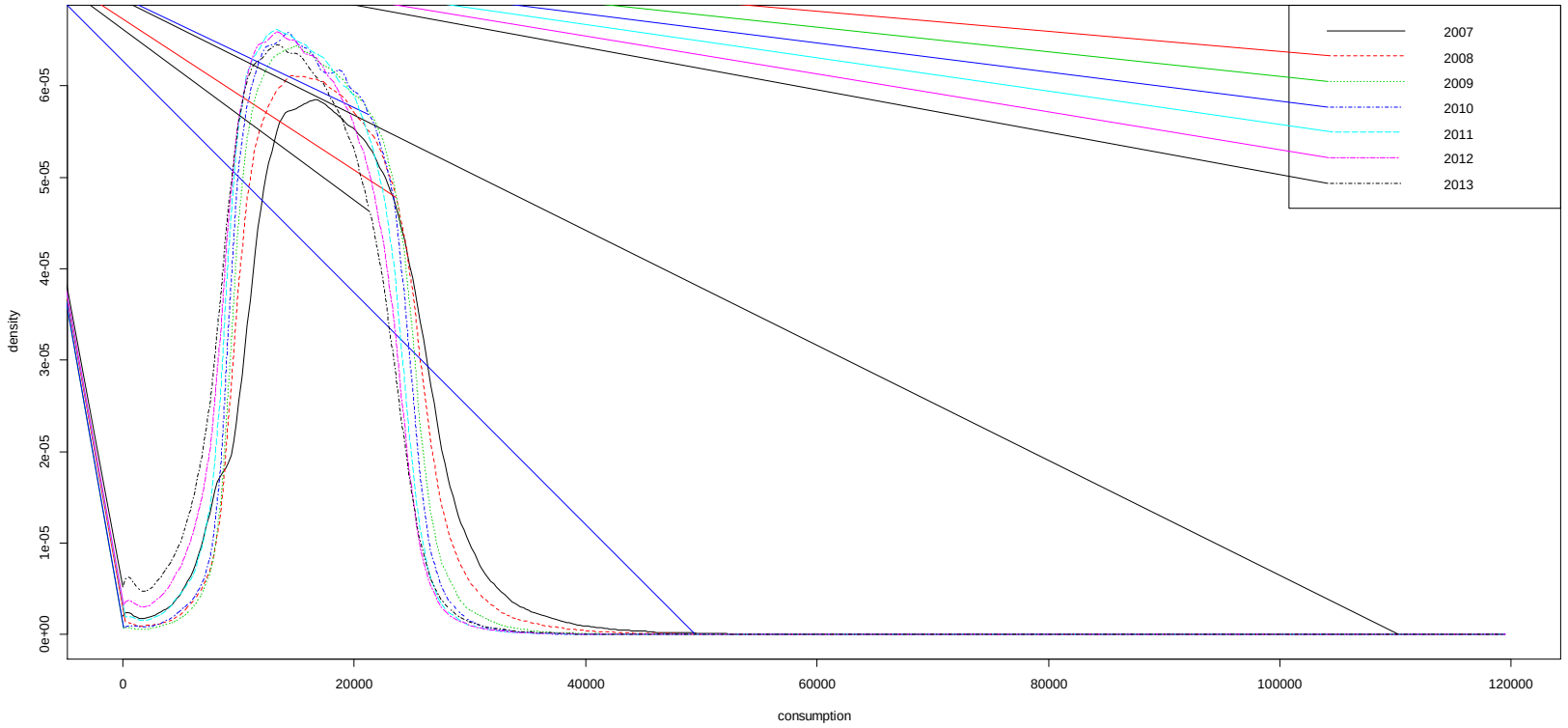




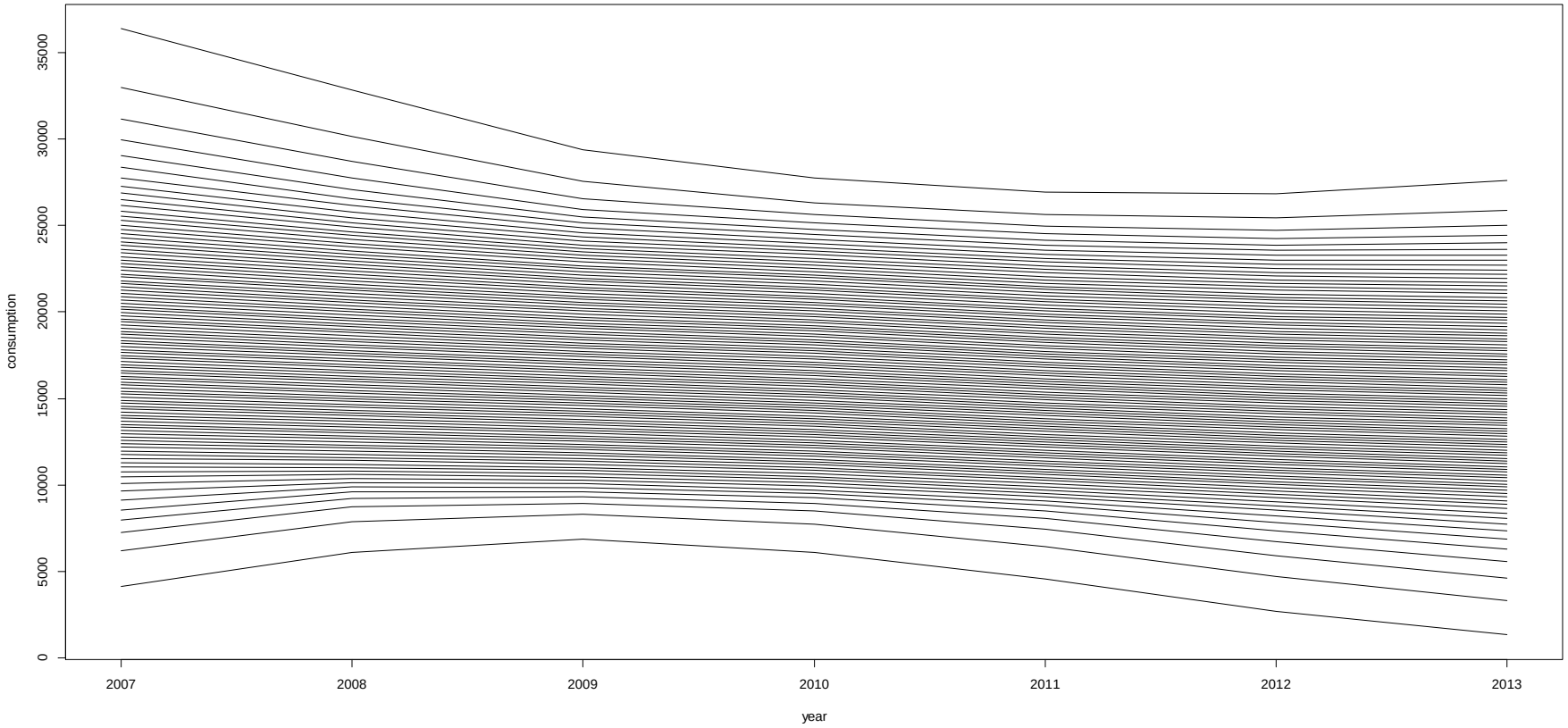
_ Business intelligence extracted from data

- details of spatio-temporal trends in natural gas consumption needed to guide planning
- 2007-2013 RWE individual household customer annual consumption data (corrected for temperature and calendar effects via normalization by the official SLPs) in kWh
- It is known that the overall consumption trend is decreasing, the interest lies in whether slope of the linear trend is spatially homogeneous

Marginal consumption distribution
(averaged over space)



Linearity?
(empirical quantiles computed year by year)



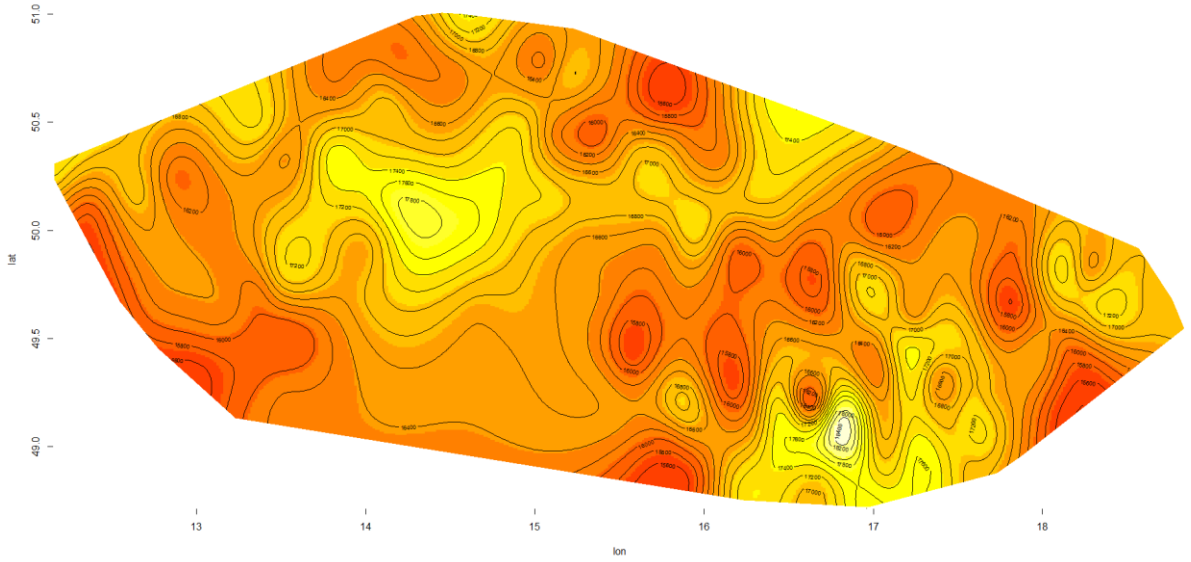
Complicated (spatial, autocorrelated) data

- Typically, for standard spatial analyses of continuous data (like the linear trends), geo-statistical methods like kriging (with estimated covariogram or variogram) are used
- Here, we will illustrate that the GAM with 2-dim spline basis can be used to model the data in a very efficient and easy-to-grasp manner

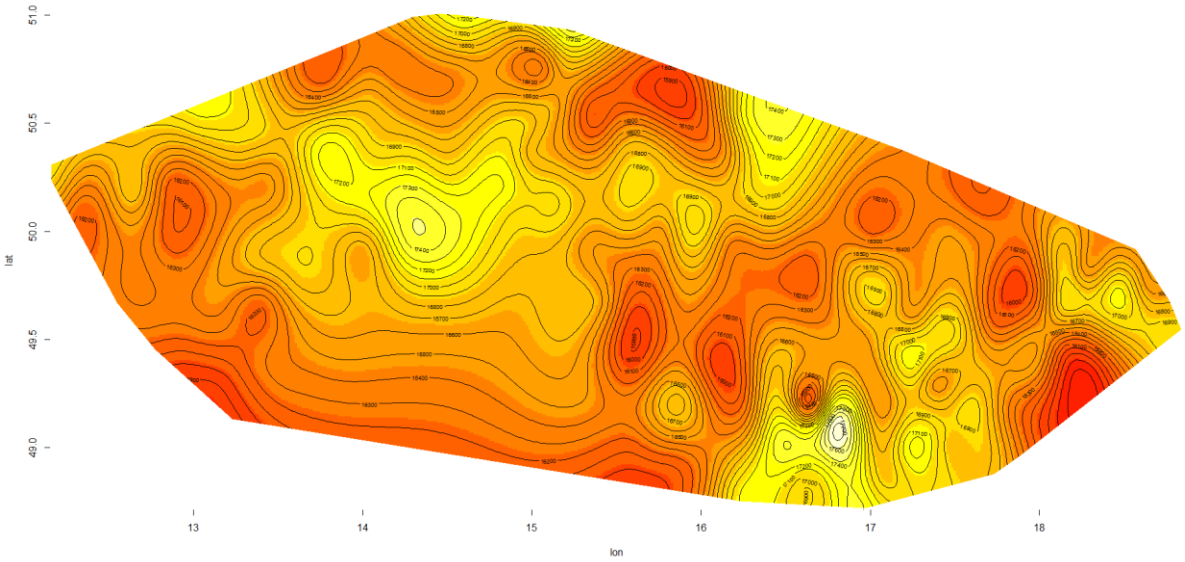
V_{spat} from $\underline{a}_{spat} \sim N\left(\underline{0}, \lambda_{spat}^{-1}.V_{spat}\right)$ specification of $s_{spat}\left(W_{1,ijt}, W_{2,ijt}\right)$ component supplies the spatial variance-covariance component in a way somewhat similar to the geostatistical modeling (via variogram estimation), but it allows for (smooth) nonstationarities (advantage!)

$$Y_{ijt} = \beta_0 + s_{spat}\left(W_{1,ijt}, W_{2,ijt}\right) + \varepsilon_{ijt}$$

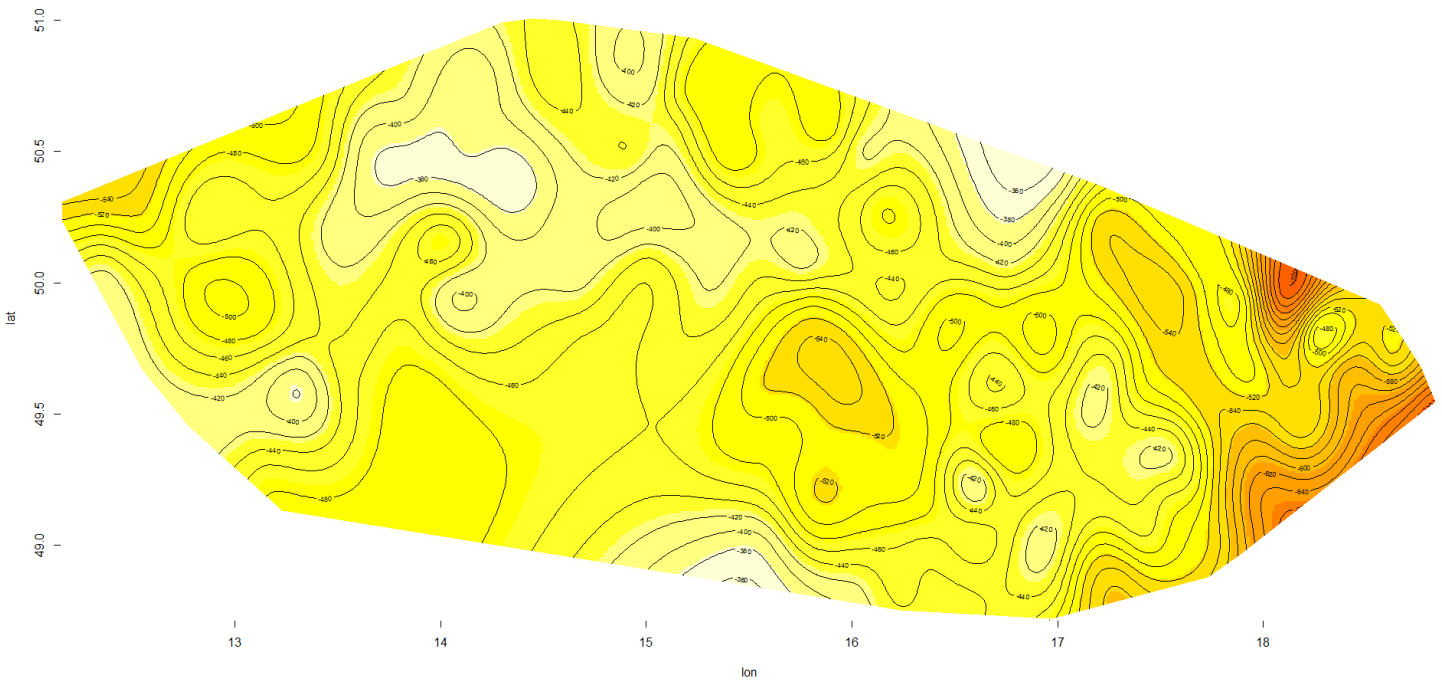
Map of the mean consumption
(mean over 2007-2013 computed individually and then modeled by GAM)



Robustness?
(10% trimming)

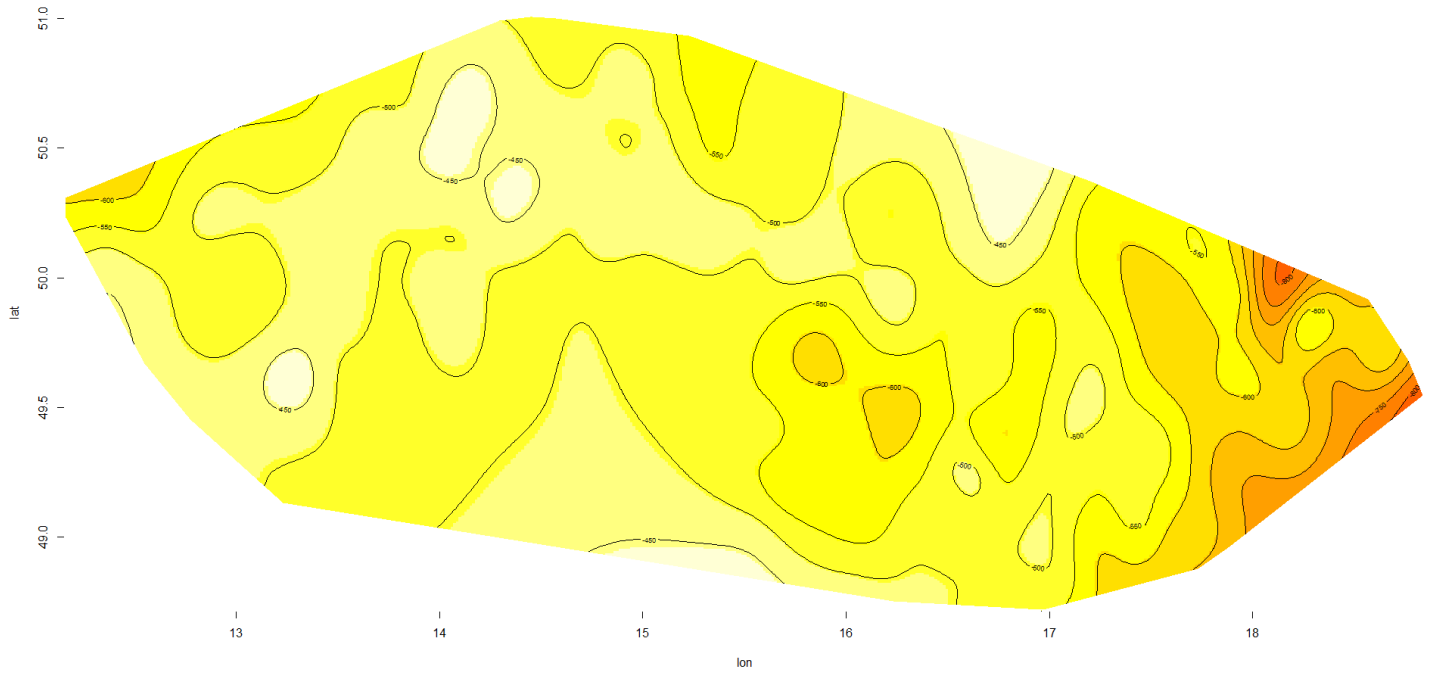


Map of the slope of 2007-2013 linear trend



Other, more in-depth views possible

(individual median of negative inter-annual change, smoothed spatially)



_ Typical scheme for NWP use in PV power forecasting

- NWP model output used as input for stat. model
  (whose outputs are eventually used as the PV power predictions)
- Main NWP variable used for PV is the GHI
  other met variables can be used as well (temperature, pressure, …)
- Statistical modeling can be seen as a calibration of the NWP outputs
  typically, it needs to be nonlinear and/or time-varying
- Sun/panel geometry and possibly other info can be used as well
  e.g. power production history

$$\hat{Y}_{t+h} = f\left(G_{t+h}; \ldots, Y_t; \theta\right)$$

- The ultimate goal is to look at the quality of the PV power forecast

- Nevertheless, it is useful to look first at the quality of the NWP prediction of the solar irradiation itself

Motivation:

- check one component of the prediction system
- kind of “upper bound” on the power prediction model performance quality

  other problems make the prediction task more complicated:
  details of panel PV production (efficiency - temperature, dust, snow)
  tilted panel complexities (geometry, direct/diffuse light)

Spatial locations of the 15 CHMI official measurement sites

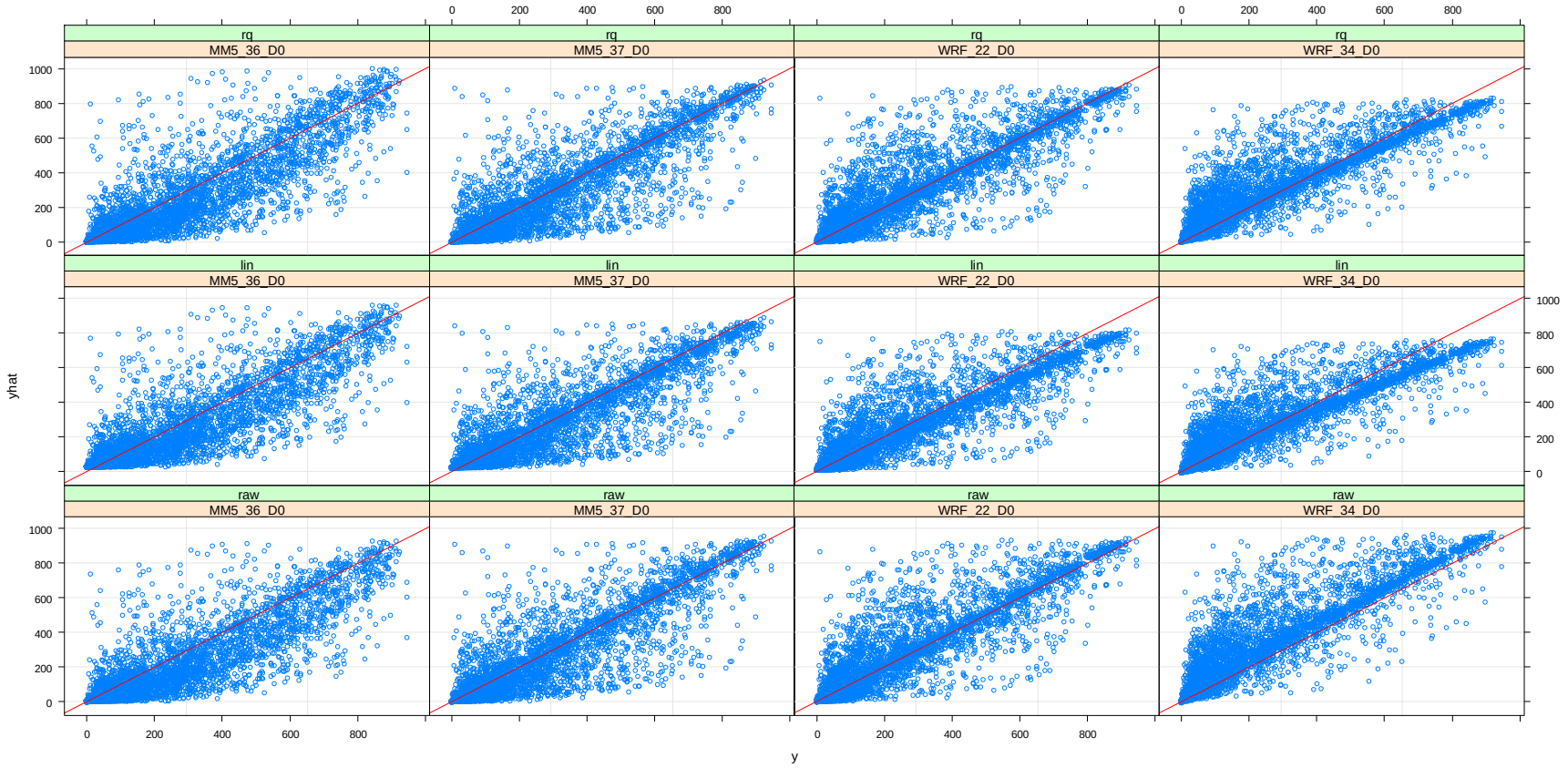
Look at:

- Different NWP models
  MM5 v 3.6, 3.7 and WRF v 2.2, 3.4
  predictions for D0 horizons

- Different “post-processing of NWP”:
  - raw NWP
    assessment of how good NWP is per se
  - two versions of simple calibration:
    assessment of how corrigible NWP is
    - linear regression,
    - quantile (L1) regression

Performance of different NWP models

(one site: Hradec Kralove)



Consequences for PV power modeling and prediction

- NWP as a predictor of the main PV power output is far from being perfect

$$Y_t = f(x_t) + \varepsilon_t$$
$$\tilde{x}_t = x_t + B_t + \nu_t$$

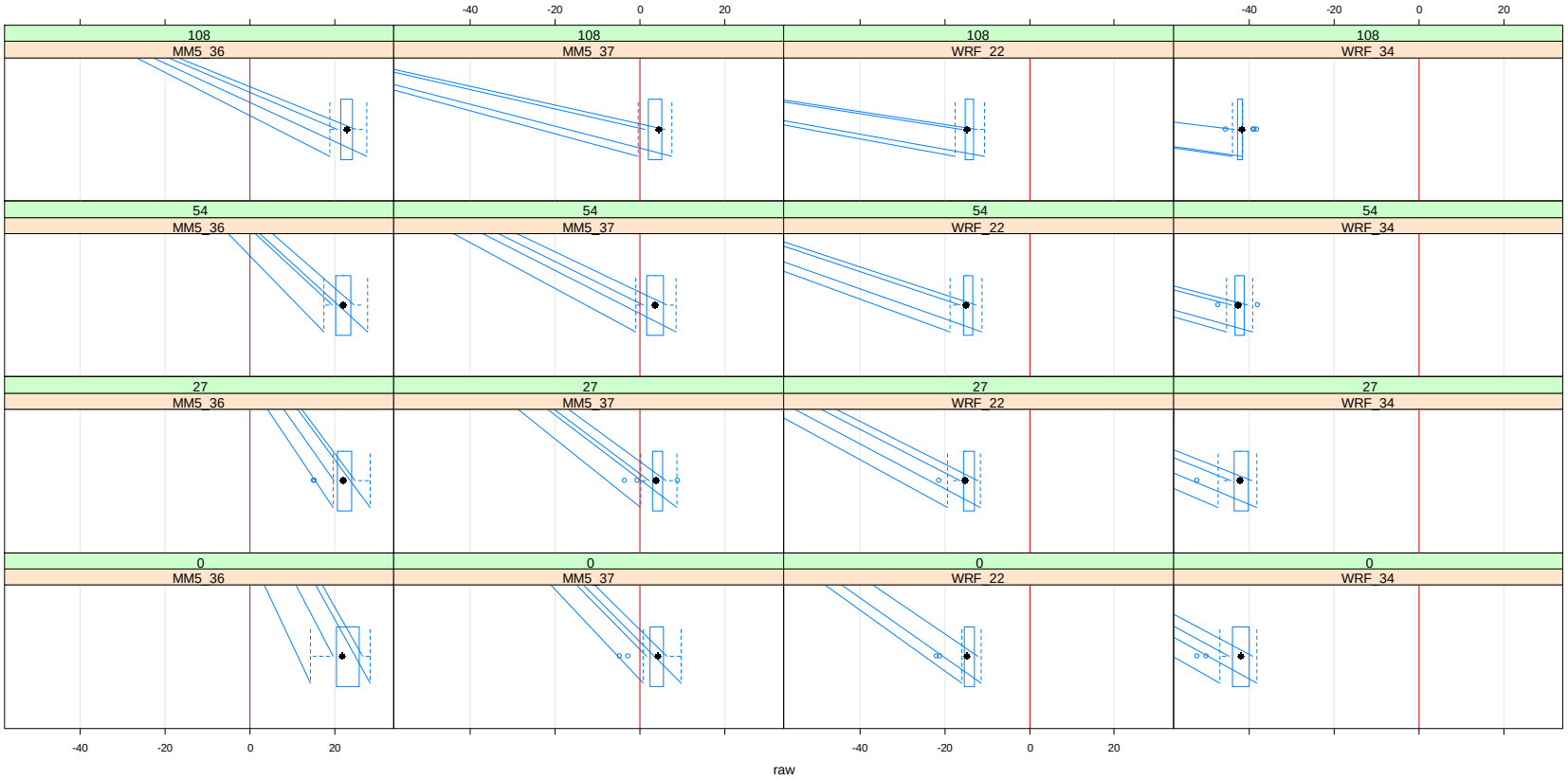
- It is quite noisy
  dealing with an errors-in-variables-problem
  NWP-related problem is specific, more complicated version of EVP

- Complicated statistical properties induced by L/U bounds
  bias, heteroscedasticity, time-varying skewness …

- Different NWP models behave differently,
  both in terms of random variability
  and systematic errors

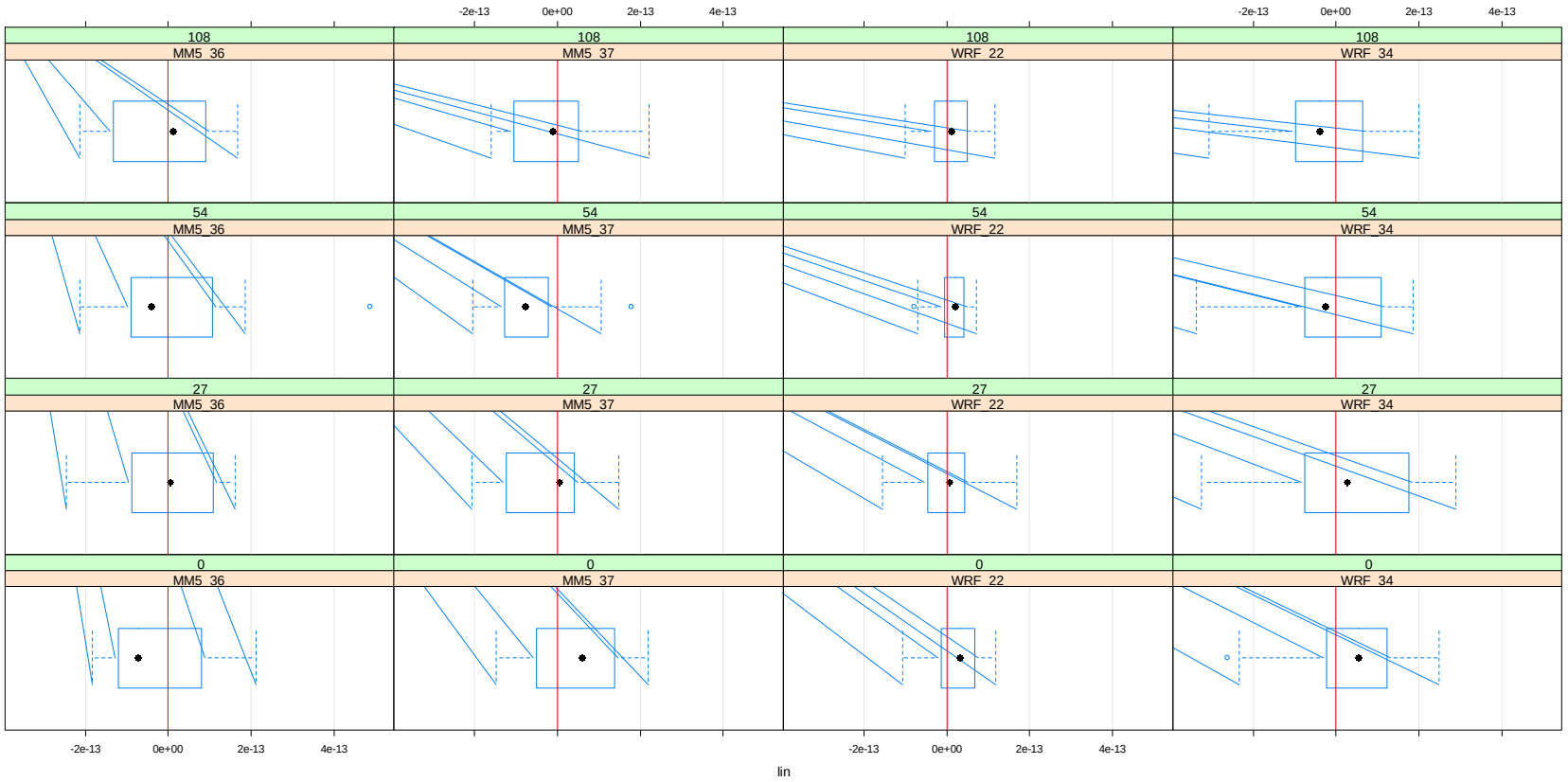
Negative bias, raw (non-calibrated) NWP, D0

effect of spatial pre-smoothing



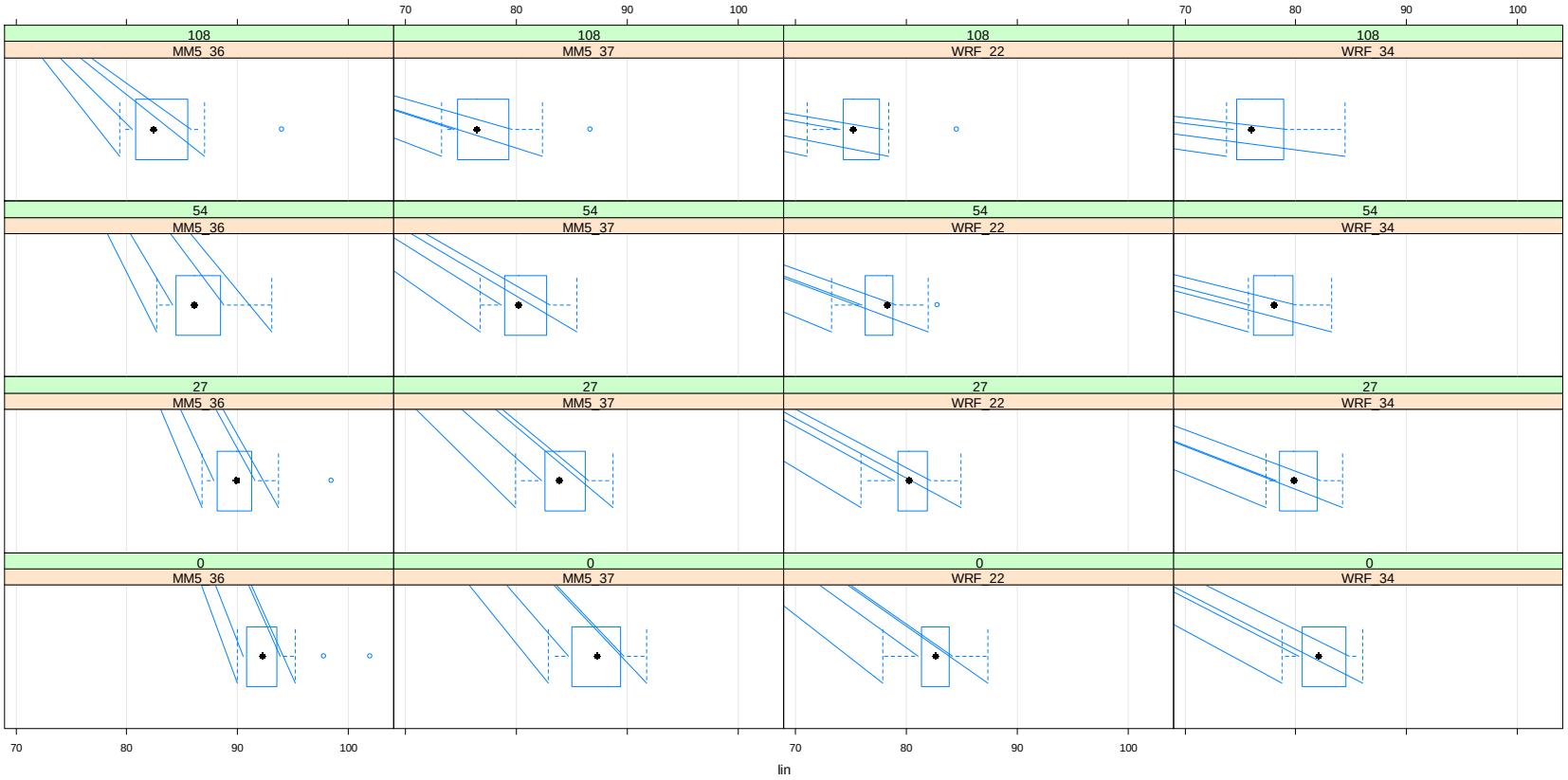
Negative bias, quantile regression calibrated NWP

effect of spatial pre-smoothing

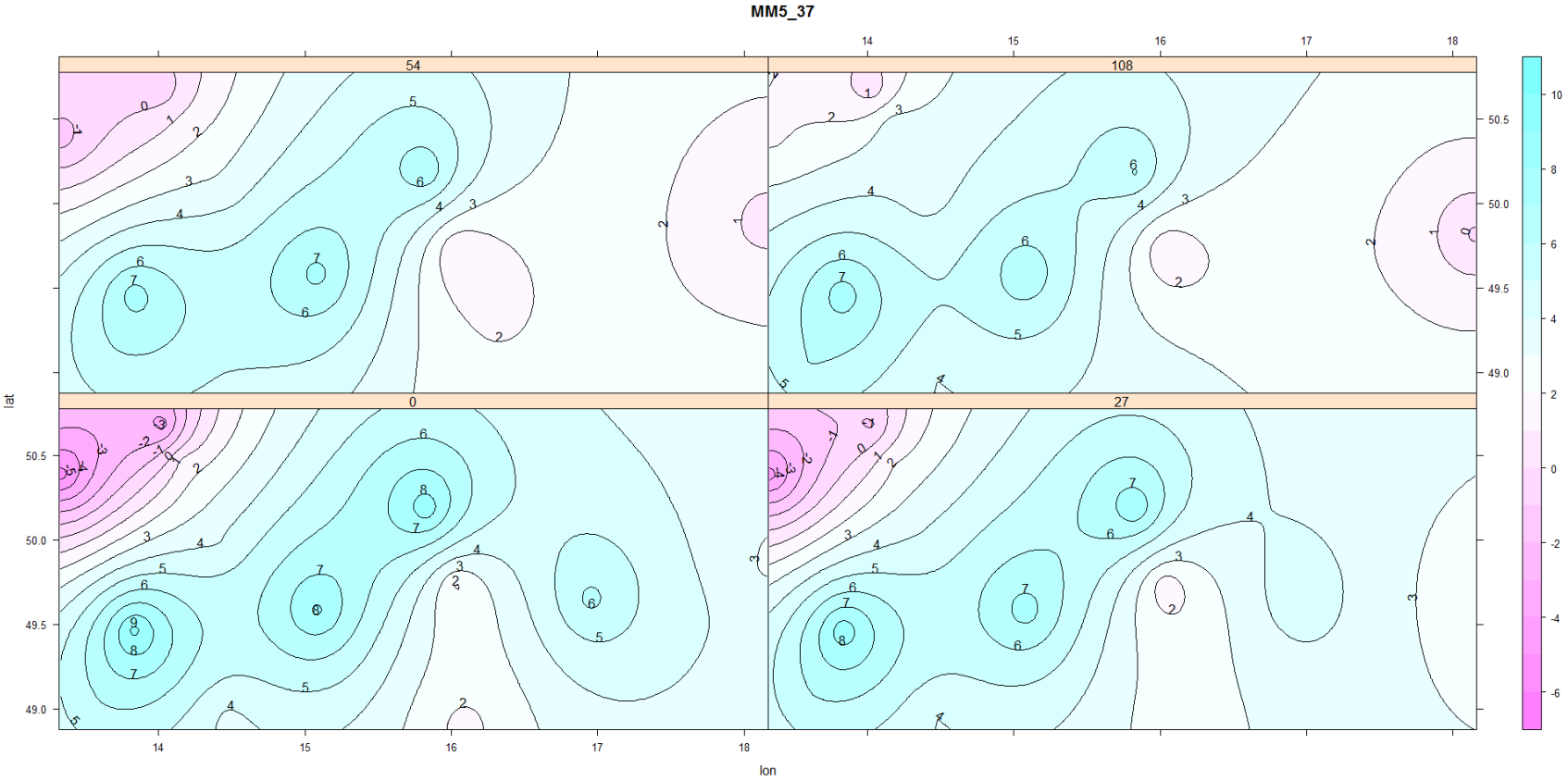


RMSE, quantile regression calibrated NWP

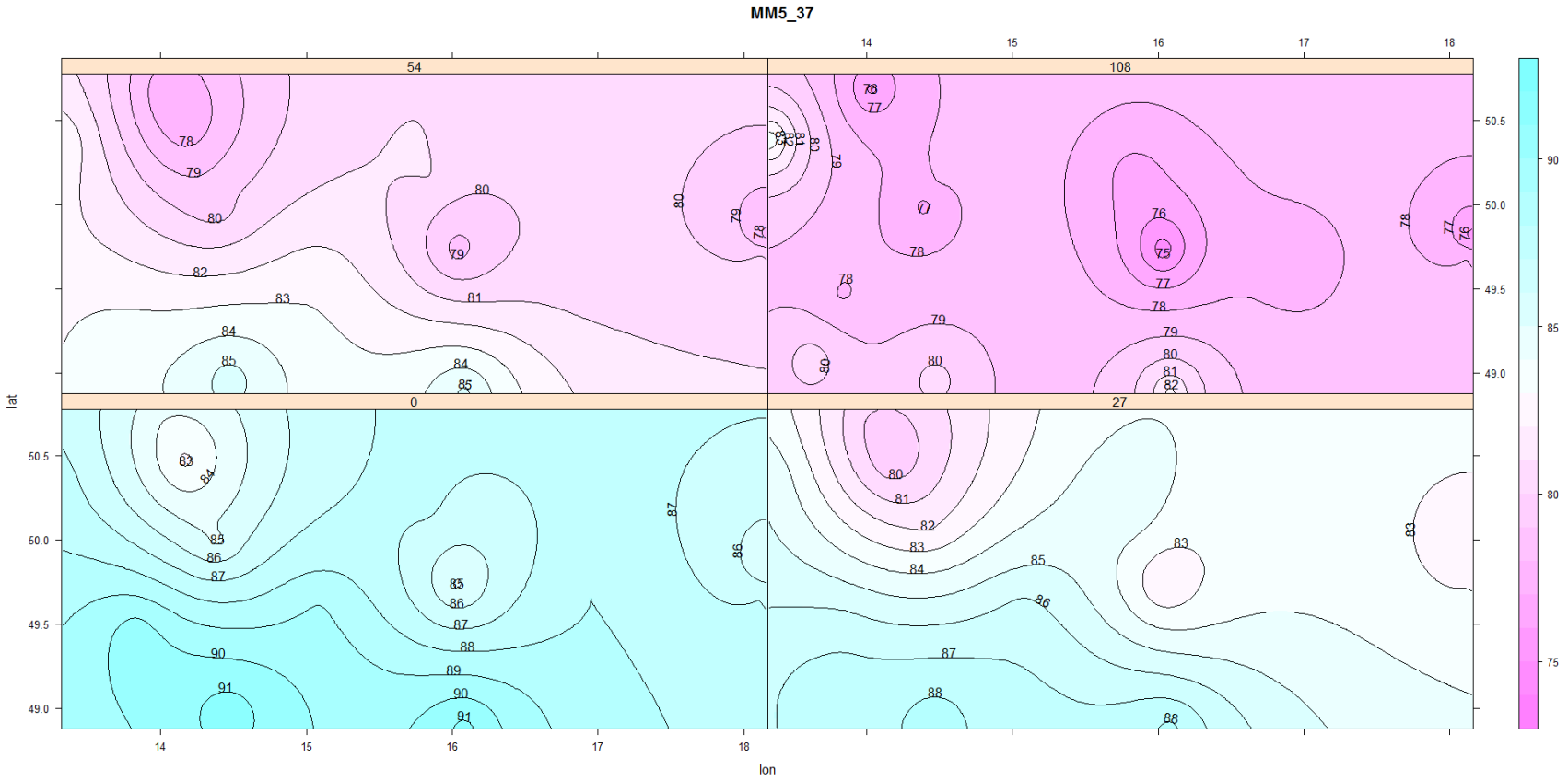
effect of spatial pre-smoothing



MM_37, negative bias, raw NWP



MM_37, RMSE, linearly calibrated NWP



Focus of the practically meaningful predictions quality assessments

- So far, we looked at traditional measures like RMSE, MAE, bias

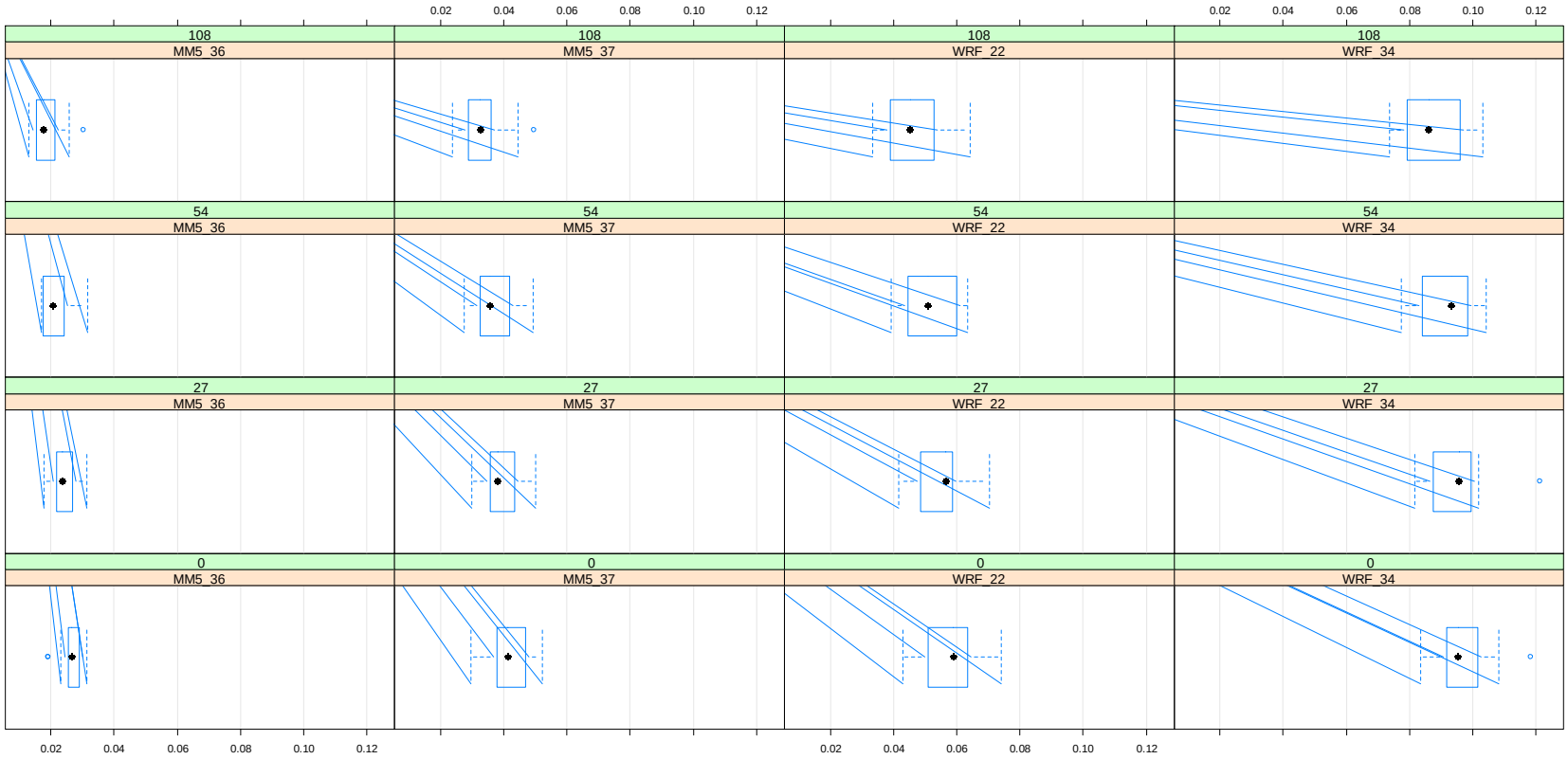
- This is typical, but does it capture everything?

  it is very common among practitioners (and also in solar-energy-related journals) to base model selection on such overall performance measures

- Perhaps, one should treat the “gross errors” differently?

FP rate

large (larger than 90th percentile of the measurements)



A take-home message

- Non-negligible part of the NWP model behavior is in fact a noise

  Pre-smoothing is advisable.

  It needs some care and effort. In particular, it **should not** be done just along the output time trajectories.

- The errors show several systematic features

  systematic biases

- Systematic deficiencies should be corrected by a “calibration” – via statistical model, based on long-term data behavior

From GHI to power

- Power predictions are more complicated than the GHI predictions

  this is despite the fact that theoretically the PV output is more or less linear in the (true) incoming light intensity

- Additional tilted panel irradiation computations

  - true panel irradiation is not easy to get
  - geometry, solar and panel angles
  - diffuse and direct irradiation components behave differently but typically, only GHI is available from NWP (climatologically-based decompositions are typically used)

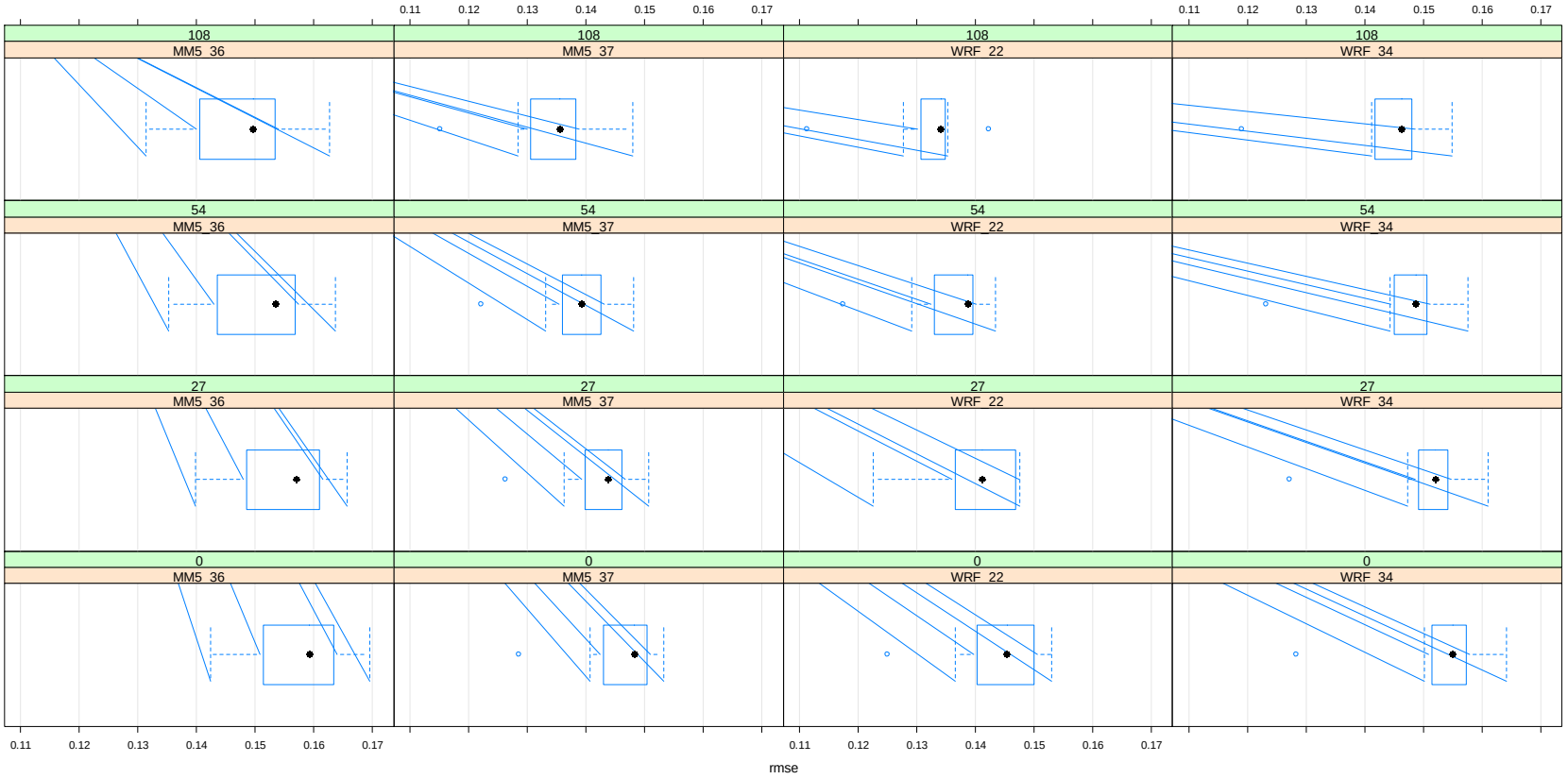
- Additional level of complexity added by the process of PV conversion

  efficiency depends on environmental variables (dust, snow, temperature,…)

  day-to-day operational issues etc.

Effect of spatial averaging and NWP model, RMSE

QR output, D0, across farms



Additional smoothing

- We saw that spatial pre-smoothing of the NWP output tends to be beneficial

- What about other, more focused smoothing of the NWP output

  GAM models with P-splines and roughness penalties

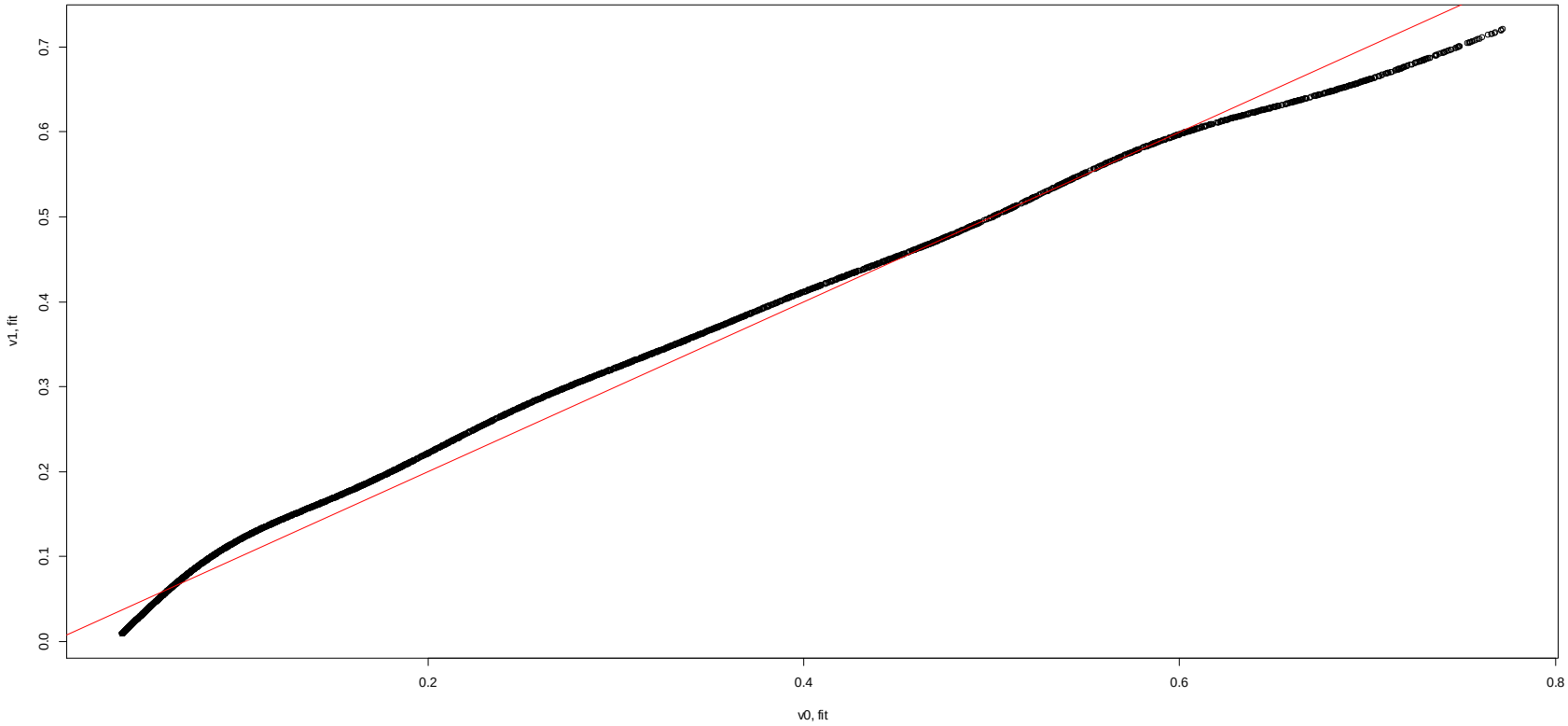
  smoothing mean power response w.r.t. the NWP as a covariate

  penalty with coefficient determined via crossvalidation

$$Y_t = \sum_{k=1}^{K} \beta_k . b_k\left(I_t\right) + \varepsilon_t$$

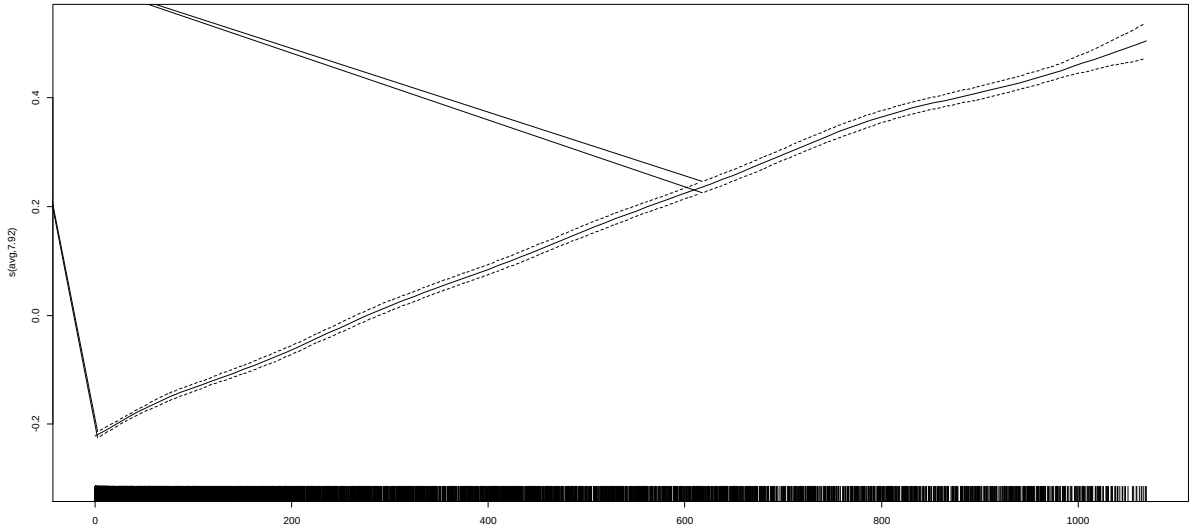
Effect of smoothing single NWP input

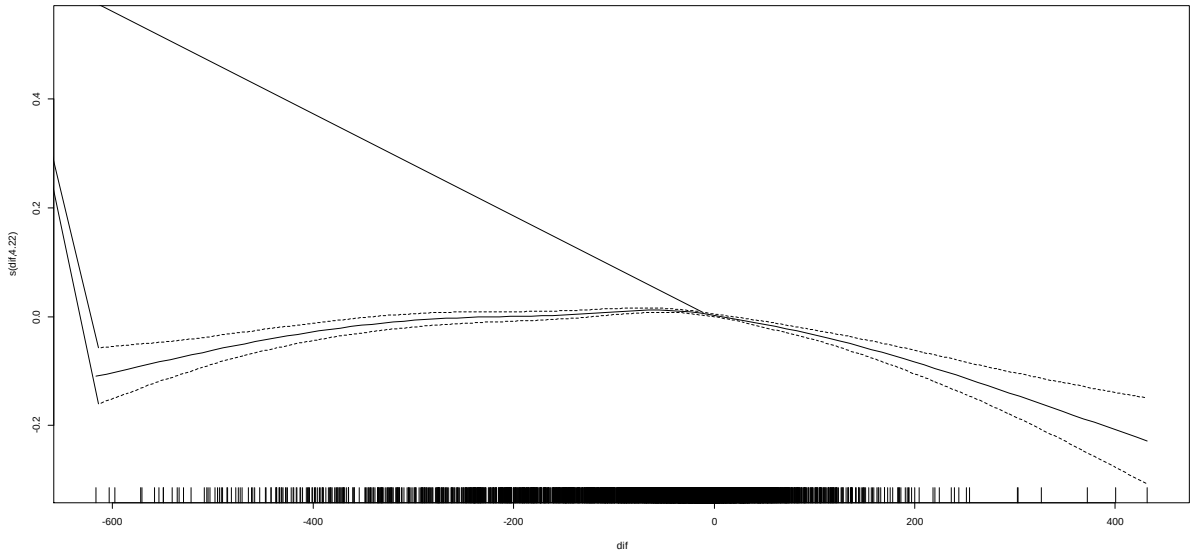
s(MM5_37)



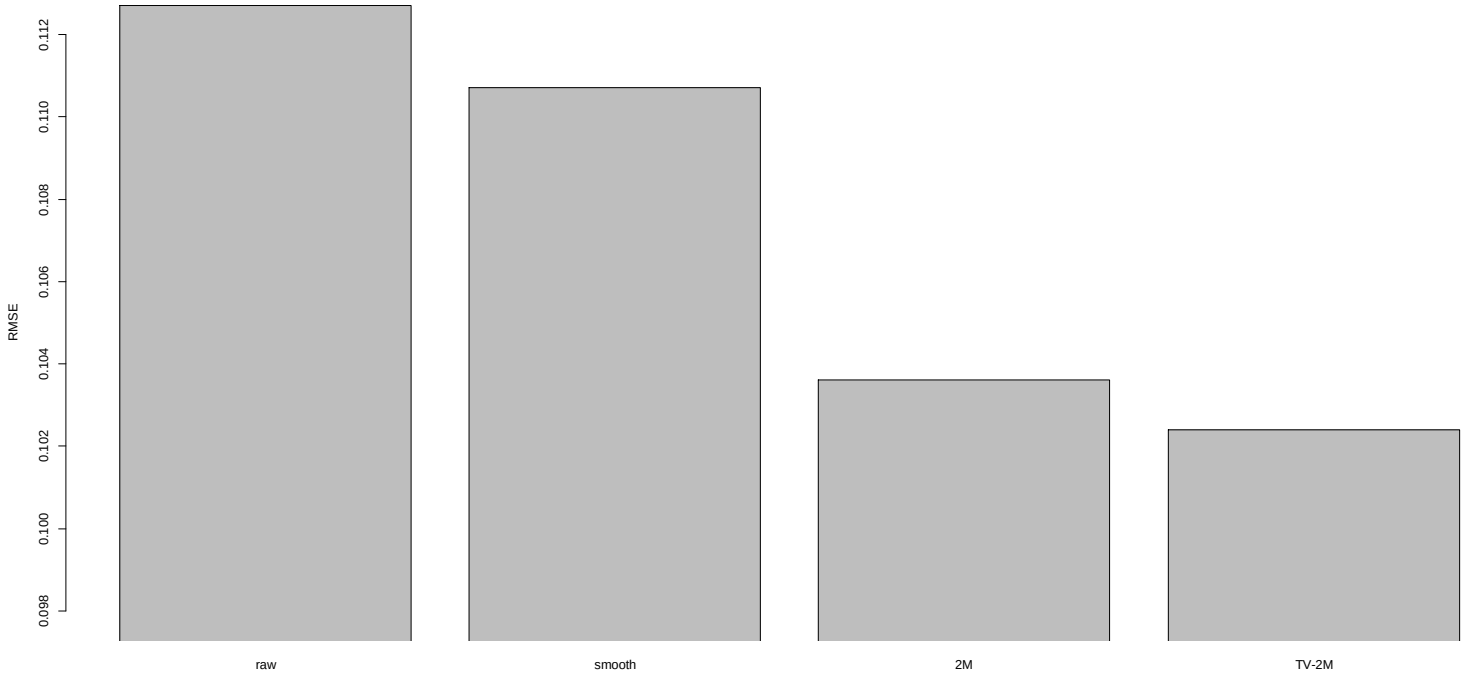
Smoothing two NWPs (MM5_37, WR_22)

$$P_t = \alpha + f_1\left(\frac{MM5\_37_t + WRF\_22_t}{2}\right) + f_2\left(MM5\_37_t - WRF\_22_t\right) + \varepsilon_t$$





Effect of NWP calibration



_ COST WIRE prediction competition

- COST ES 1002, WIRE

  Weather Intelligence for Renewable Energies

- Solar farm, Catania, Italy 2010-2011

  total nominal power 2.1 kW

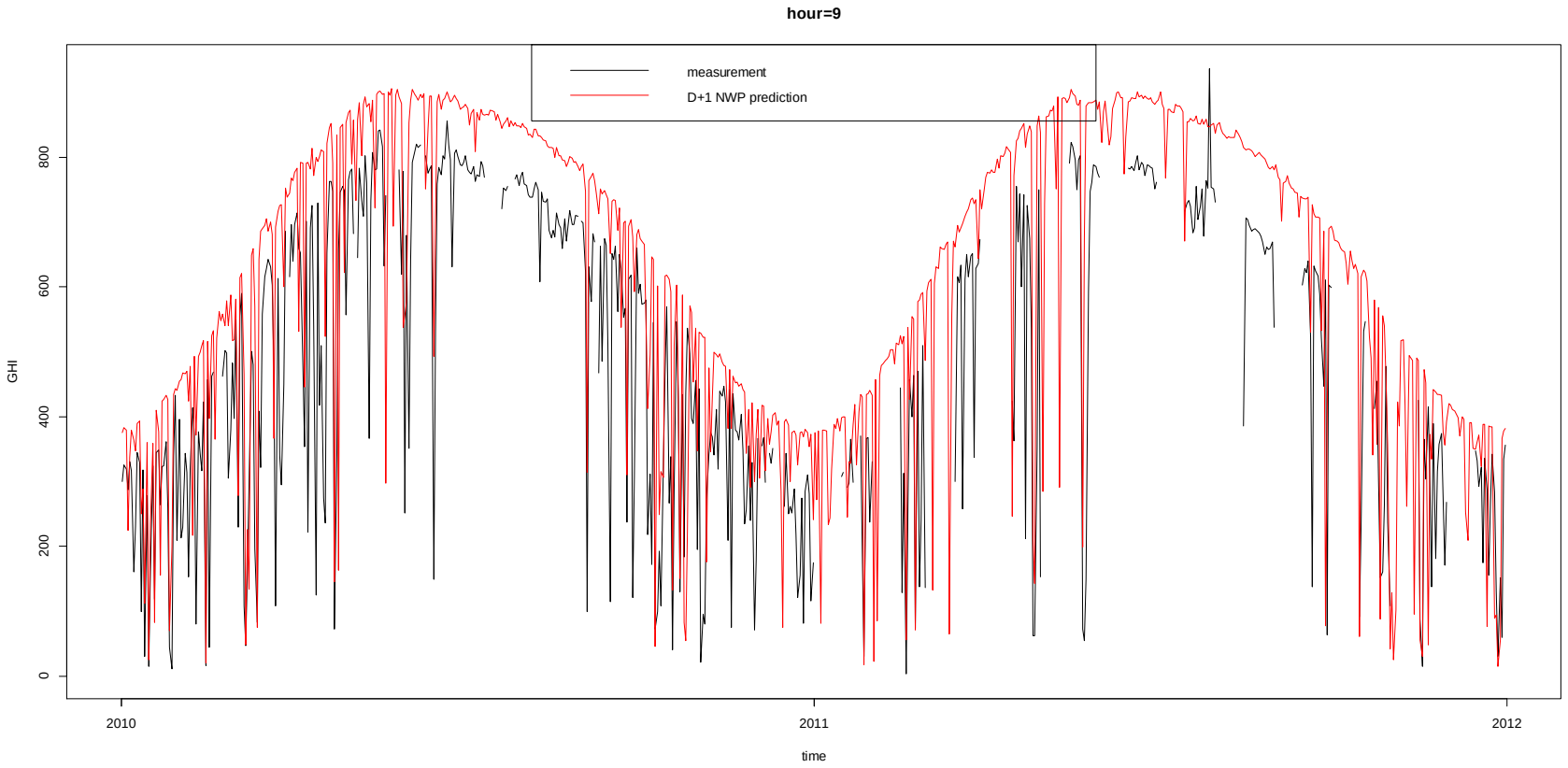
- Teams from 19 countries

  employing various prediction techniques

- Scenario: “train” (estimate) on 2010 and use (evaluate) on 2011 data

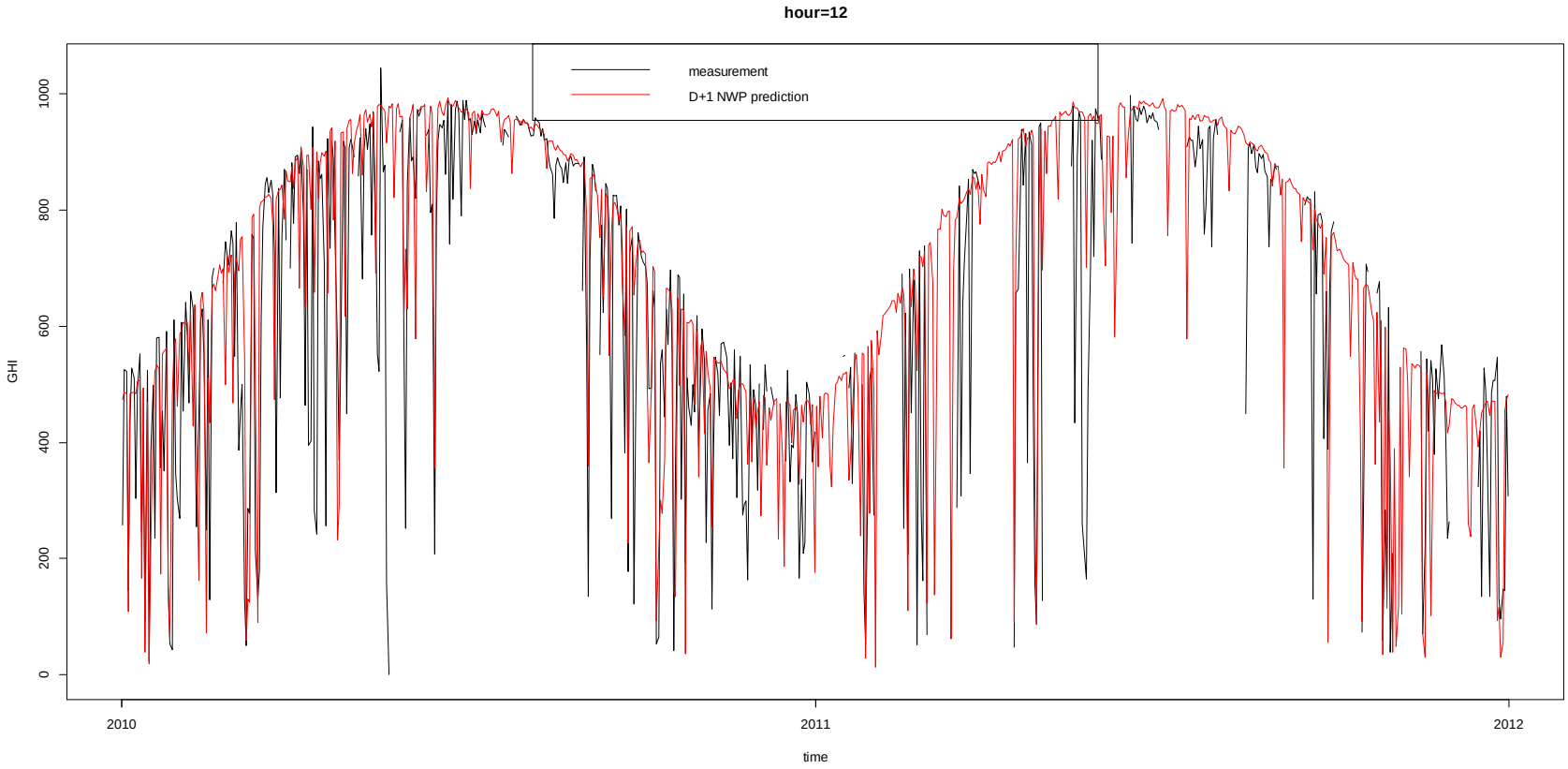
Raw NWP is far from being perfect …

RAMS, D0 horizon



Bias has a nontrivial temporal structure …

RAMS, D0 horizon



Model for power prediction

- Gaussian GAM

  with both linear and spline components motivated physically

- Using two NWP model outputs as inputs for the statistical prediciton model

  WRF, RAMS

- Light components enter model separately

  direct, diffuse

- Penalized difference between the NWP's

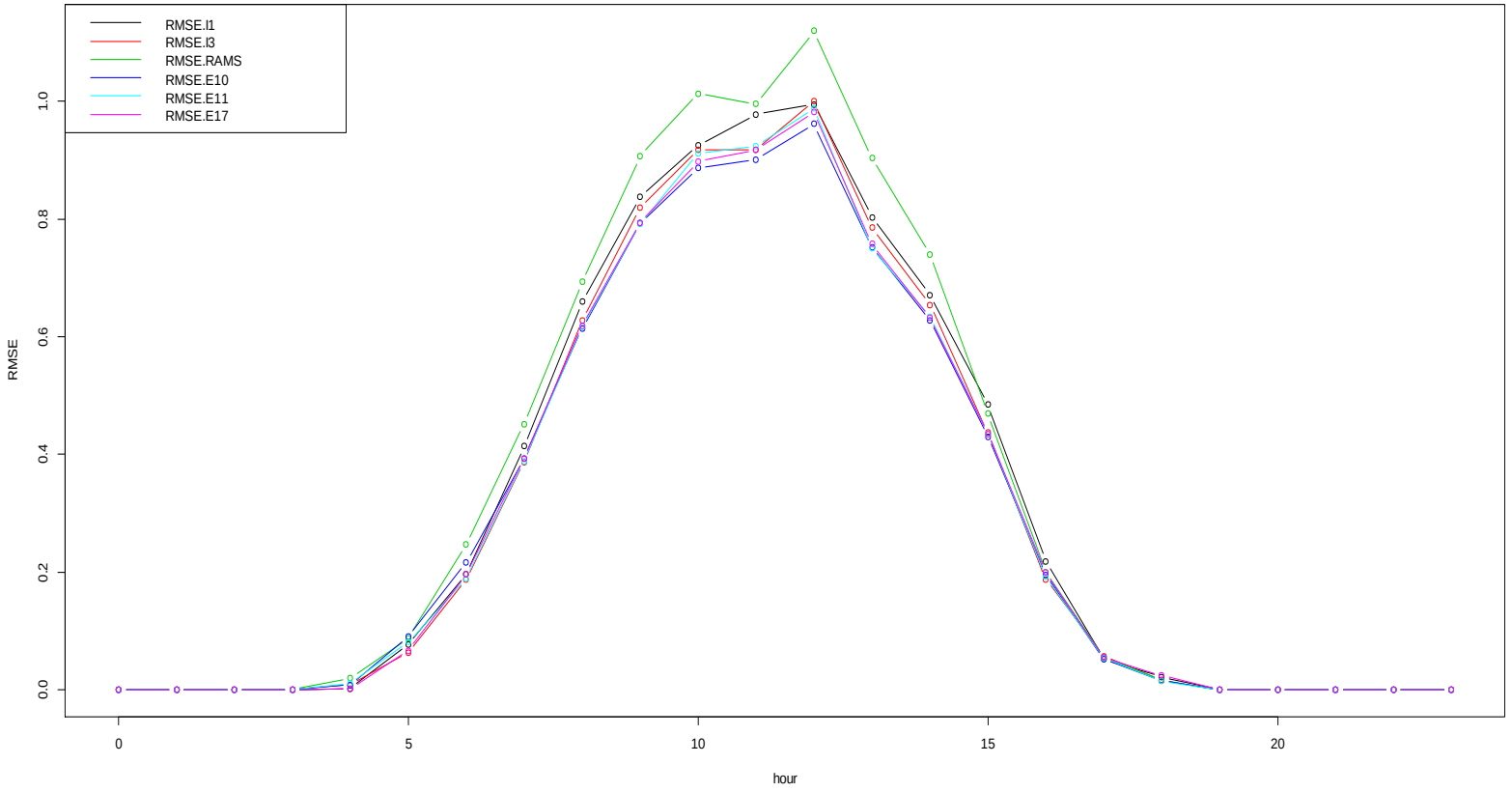
- Interaction between NWP, cos of zenith angle

Model
with Bayesian and shrinkage motivation

$$Power_t = \beta_1.DIR.WRF_t + \beta_2.DIFF.WRF_t +$$
$$\beta_3.DIR.WRF_t.\cos(z_t) + \beta_4.DIFF.WRF_t.\cos(z_t) +$$
$$s_1(DIR.WRF_t - DIR.RAMS_t) + s(DIFF.WRF_t - DIFF.RAMS_t) + \varepsilon_t$$

RMSE, MAE

out-sample



_ Spatio-temporal prediction of cloud coverage obtained from a satellite

- Satellite data are used to improve short-term forecasts of cloudiness and hence of photovoltaic production, their use becomes widespread and almost routine

- GISAT project

  - regular grid over Europe and more
  - domain lat 30.02 to 64.98, lon -9.97 to 29.98
  - squares of 150 arcsec, corresponding roughly to 4.6x4.6km square
  - data available each 15 min
  - prediction to the next hour (and more)

Model, I

- Uses past “optical flow” estimated by satellite experts from comparison of consecutive images

  (from immediate past before making the prediction)
- Markov spatio-temporal model

  working with kernel operating on close spatial neighborhood of a predicted point
- Kernel is deformed

  according to the optical flow vector (size and angle)
- Model is trained on 15min transition and then propagated in the Markovian style

(t-1)

$i-2, j+2$	$i-1, j+2$	$i, j+2$	$i+1, j+2$	$i+2, j+2$
$i-2, j+1$	$i-1, j+1$	$i, j+1$	$i+1, j+1$	$i+2, j+1$
$i-2, j$	$i-1, j$	i, j	$i+1, j$	$i+2, j$
$i-2, j-1$	$i-1, j-1$	$i, j-1$	$i+1, j-1$	$i+2, j-1$
$i-2, j-2$	$i-1, j-2$	$i, j-2$	$i+1, j-2$	$i+2, j-2$

(t)

Model, II

- Spatial position i,j and time t
- Optical flow *O*

$$Y_{ijt} \sim Bernoulli\left(\pi_{ijt}\right)$$

$$\log\left(\frac{\pi_{ijt}}{1-\pi_{ijt}}\right) = \beta_0 + s\left(O_{size,i,j}\right).Y_{i,j,t-1} +$$

$$\sum_{k,l \in B} \beta_{k,l}.Y_{i+k,j+l,t-1} + \sum_{u,v \in S} s_{u,v}\left(O_{size,i+u,j+v}, O_{angle,i+u,j+v}\right).Y_{i+u,j+v,t-1}$$

Model, III

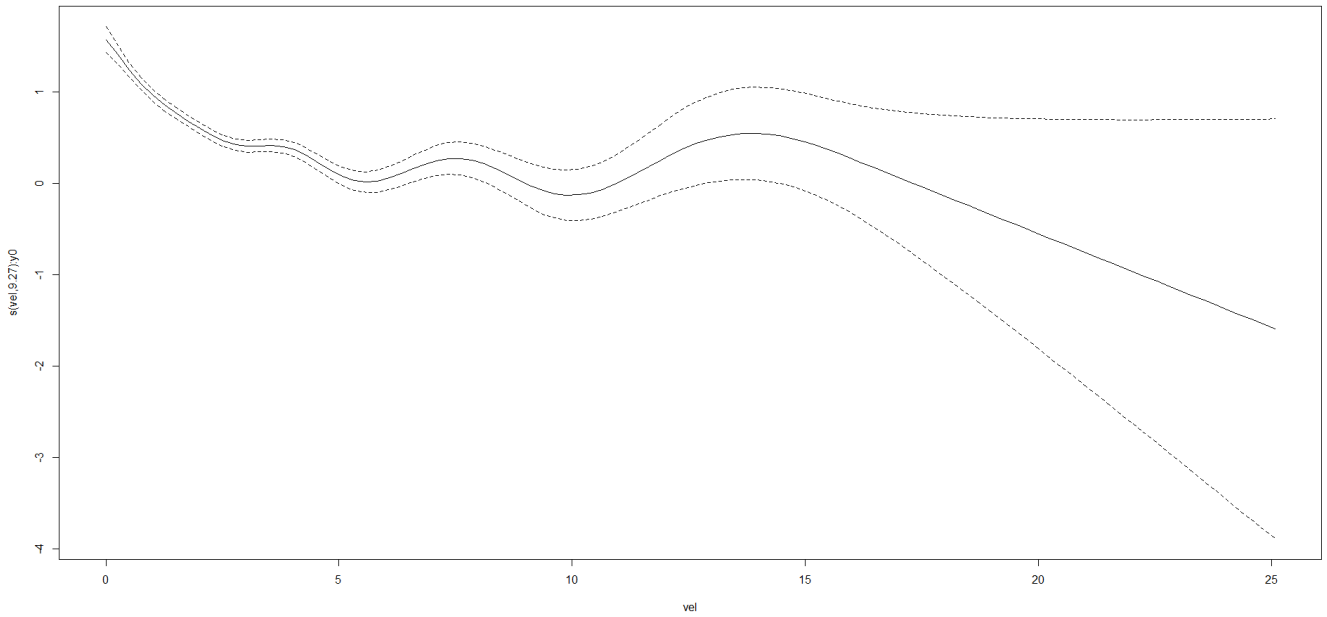
S	B	B	B	B
S	S	B	B	S
S	S	S	S	B
S	S	B	S	B
B	S	B	B	B

Examples of how one can

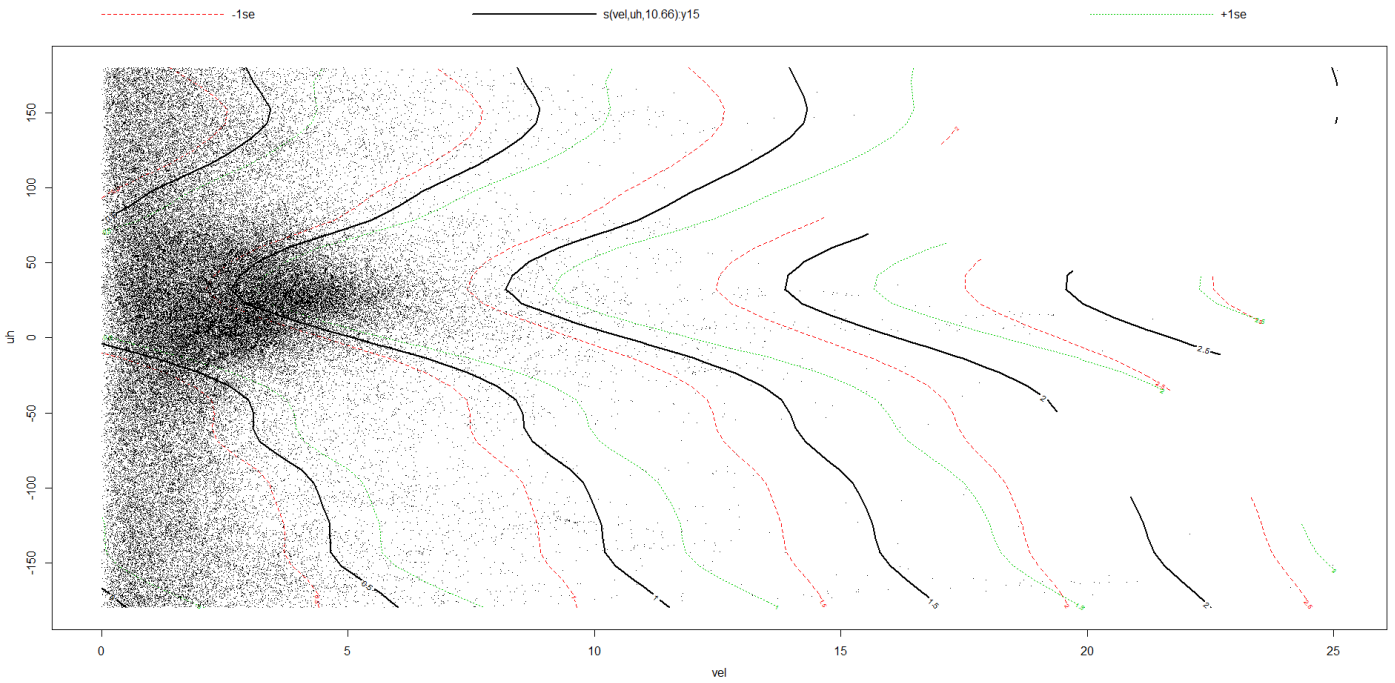
- “read” the model components

- and check them against physically motivated ideas about the structure of influence

Coefficient of the same pixed in the past, depending on the flow vector size



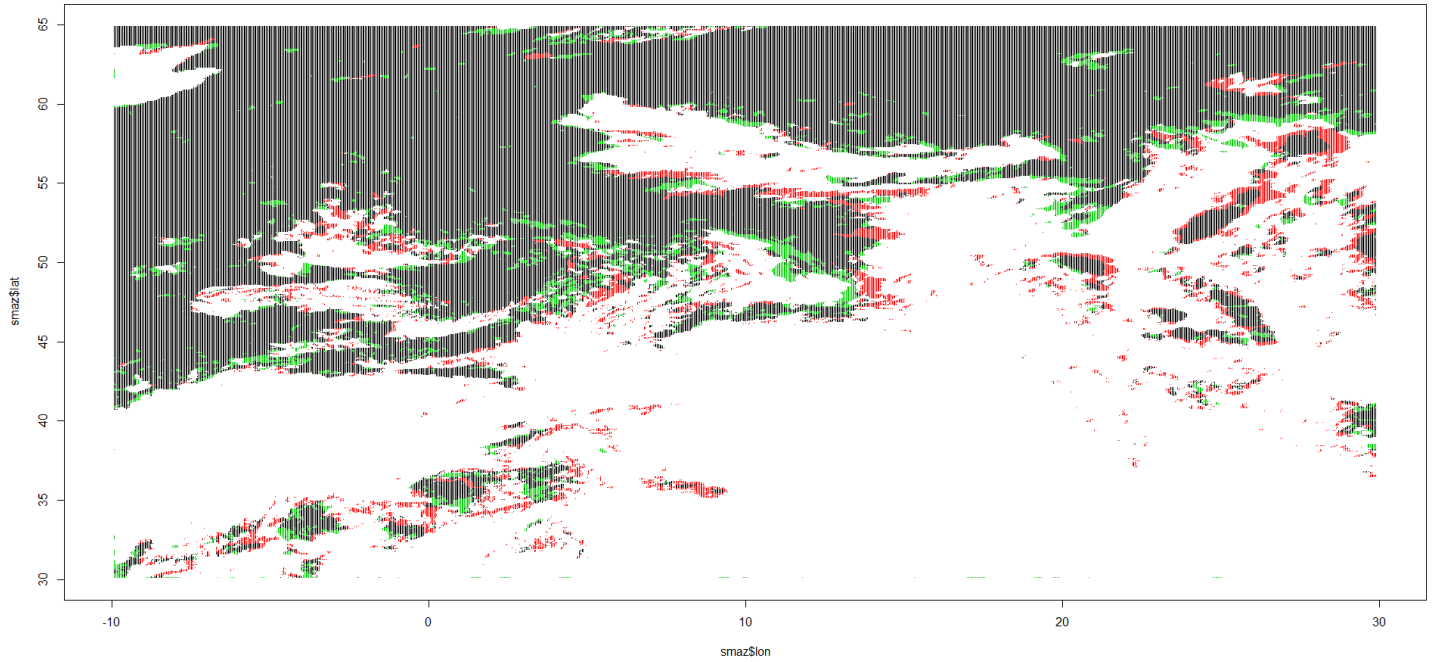
Coefficient of the pixel SW from predicted location, dependence on magnitude and angle of the flow vector



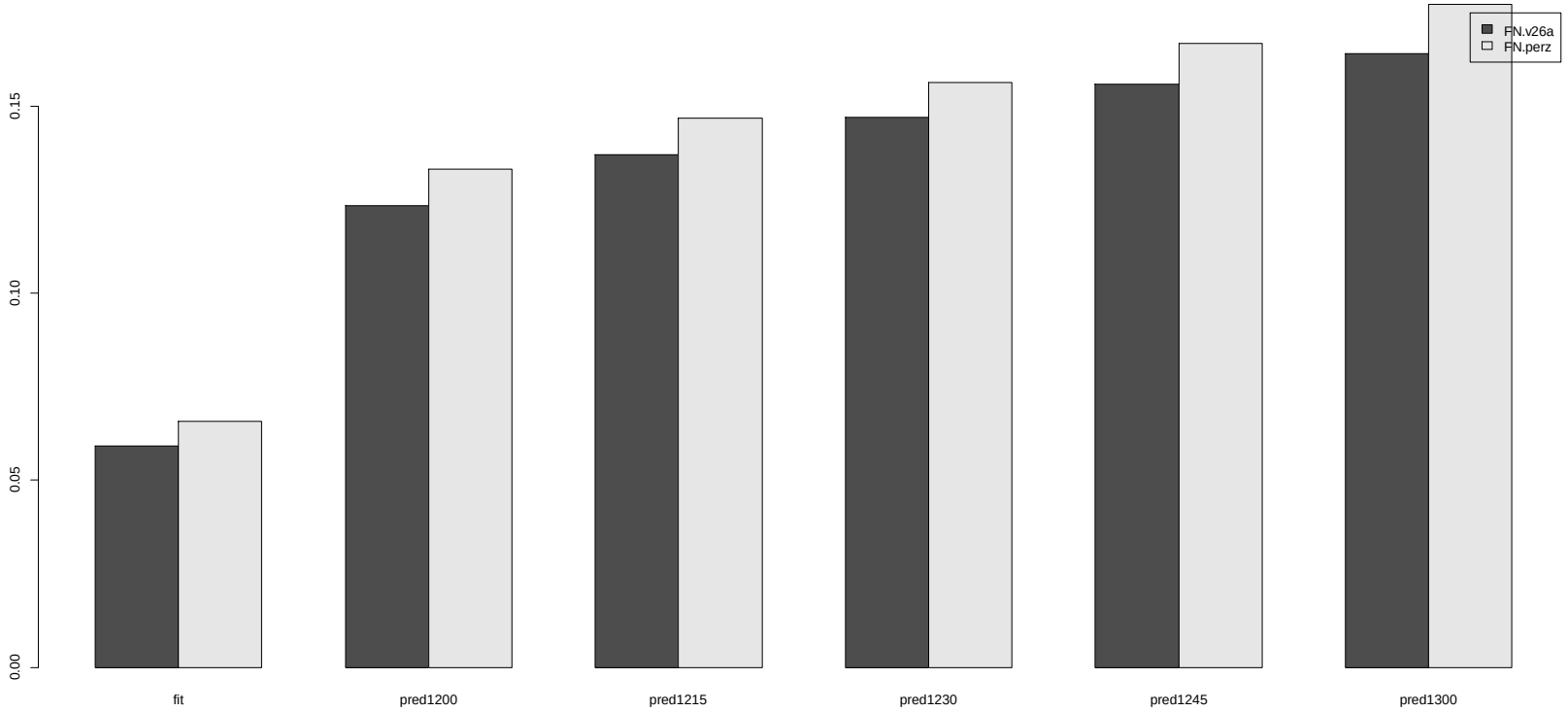
Summaries

	Total misclassification	FP (for cloudiness)	FN (for cloudiness)
persistence	0.10362	0.08498	0.13303
Markov model	0.09112	0.07368	0.11863

Model predictions



Different horizons



_ Statistical model for Standardized Load Profiles (**SLP**)

Experience from two SLP official projects

- Czech Republic and Slovakia
- in CR, the SLP model is now a part of gas-regulation legislative

The SLP model has several expert-motivated interpretable/checkable components

- stratification upon customer type segments
- multiplicative components
- effect for previous (long-term) consumption (offset-type)
- correction for calendar effects (weekday type, Christmas, Easter)
- correction for long-term trends (insulation, change of heating …)
- temperature effects on two time scales (immediate and weeklong)

Stratification

- Model is built separately for various *segments* of customer pool (stratification)
- HOU and SMC
- Using info about natural gas appliances and broad consumption level brackets
- E.g. HOU1 – cooking, HOU4 – space heating

Data for SLP modeling

- Sample of hourly measured customers
  - of cca 1000
  - empirical data suffer from occasional measurement and other errors
- Individual data
  - historical consumption
  - consumer segment type
- Exogeneous explanatory variables
  - temperature
  - calendar and other time effects

- Aggregated (routine) data for checking

GAM model

- Normal, with log link, offset and smooth (spline) components
- Important interaction between short-term temperature effect and day type
- Stratified on customer segment (HOU, SMC)

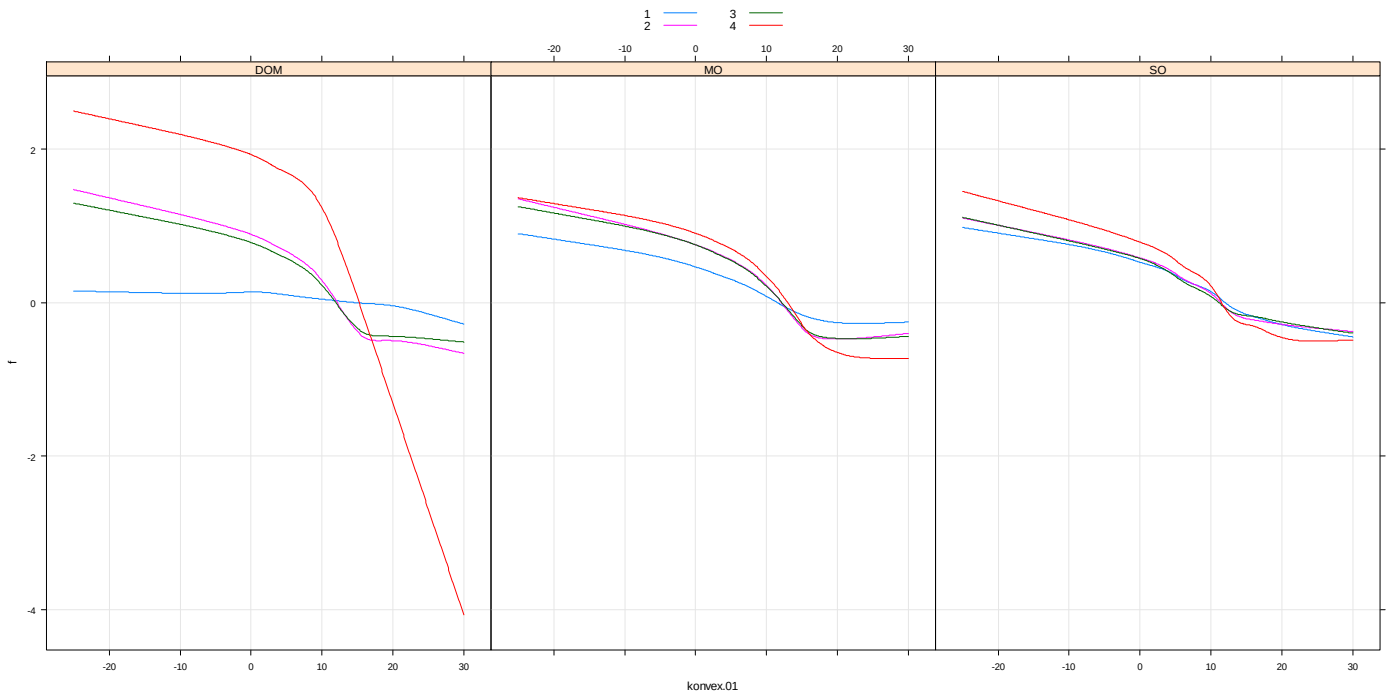
$$Y_{ikt} \sim N\left(\exp\left(\mu_{ikt}\right), \sigma_k^2\right)$$

$$\mu_{ikt} = \log\left(p_{ik}\right) + \sum_{j=1}^{5} \alpha_{j,k}.I\left(\text{day } t \text{ is of type } j\right) + s_{trend,k}\left(t\right) +$$

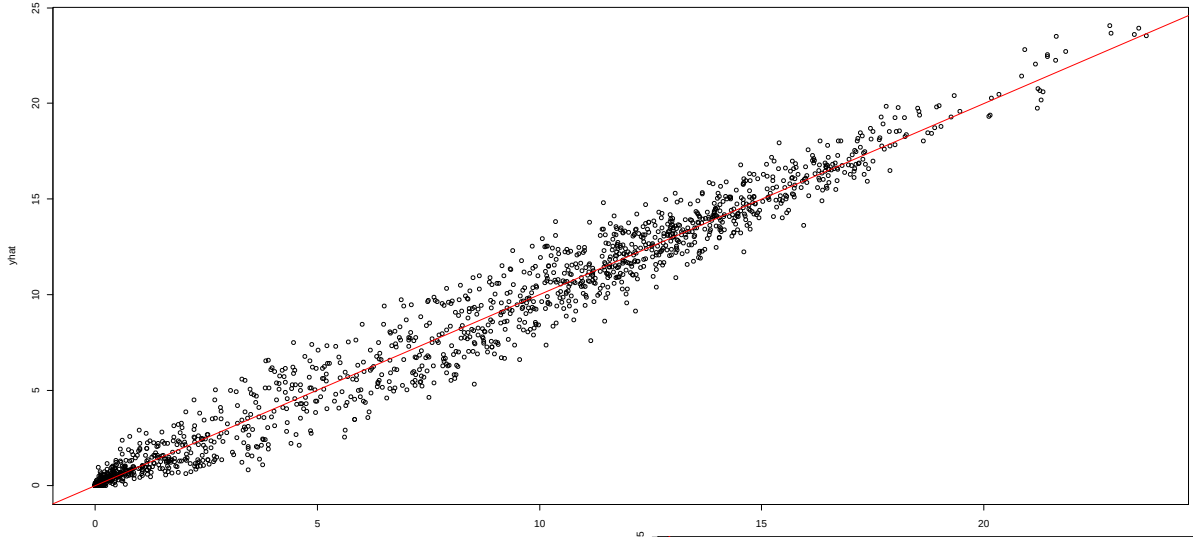
$$\beta_{Christmas,k}.I\left(\text{day } t \text{ is within the Christmas period}\right) + \beta_{Easter,k}.I\left(\text{day } t \text{ is within the Easter period}\right) +$$

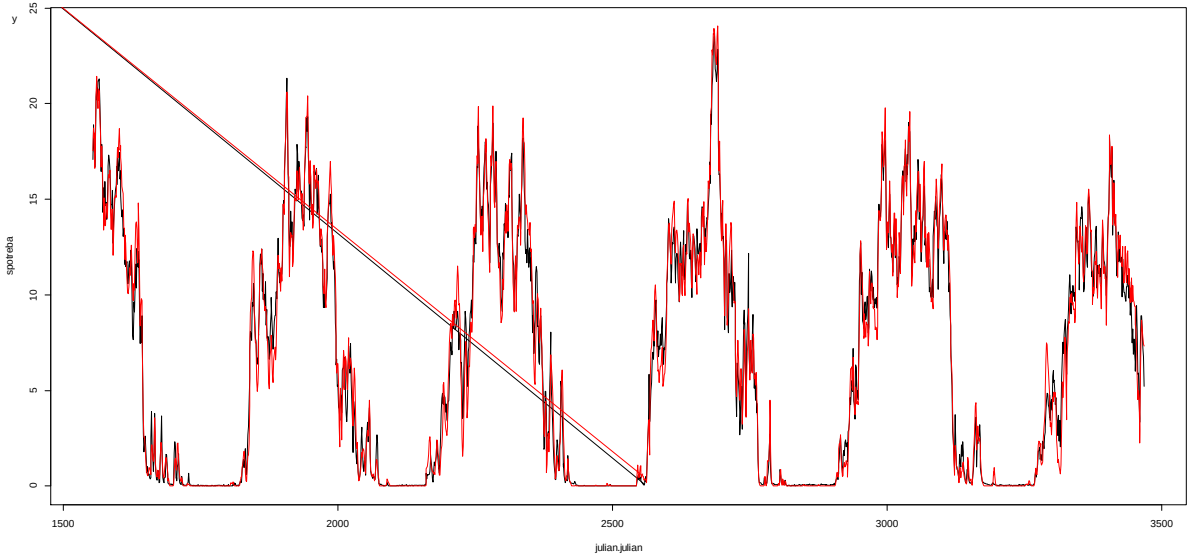
$$\sum_{j=1}^{5} \gamma_{j,k}.s_{short,k}\left(w.T_t + \left(1-w\right).T_{t-1}\right).I\left(\text{day } t \text{ is of type } j\right) + s_{week,k}\left(\frac{\sum_{j=0}^{6} T_{t-j}}{7}\right)$$

Example of a short-term temperature effect smooth funciton, $s_{short,jk}(.)$

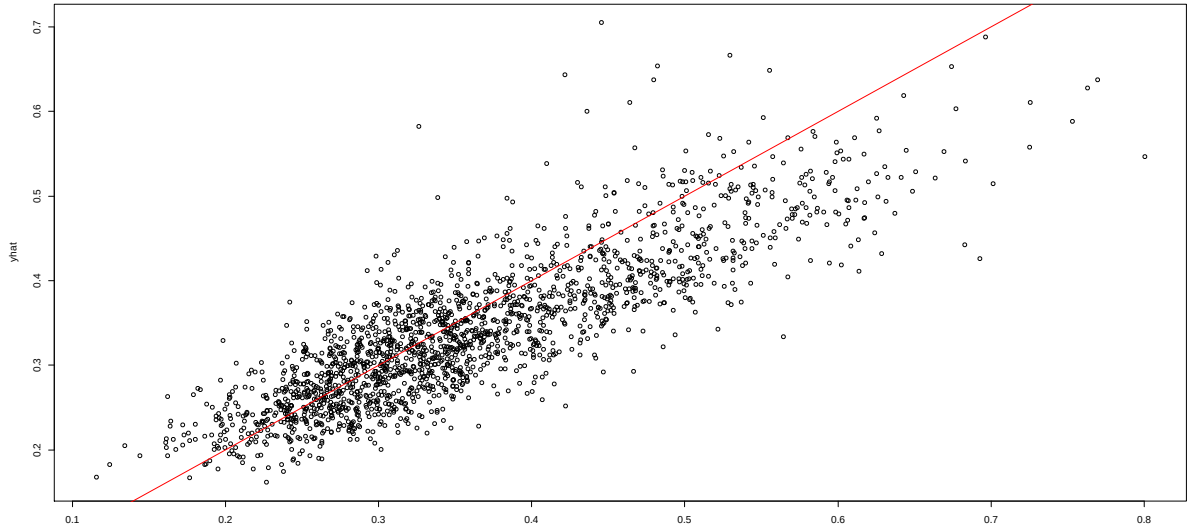


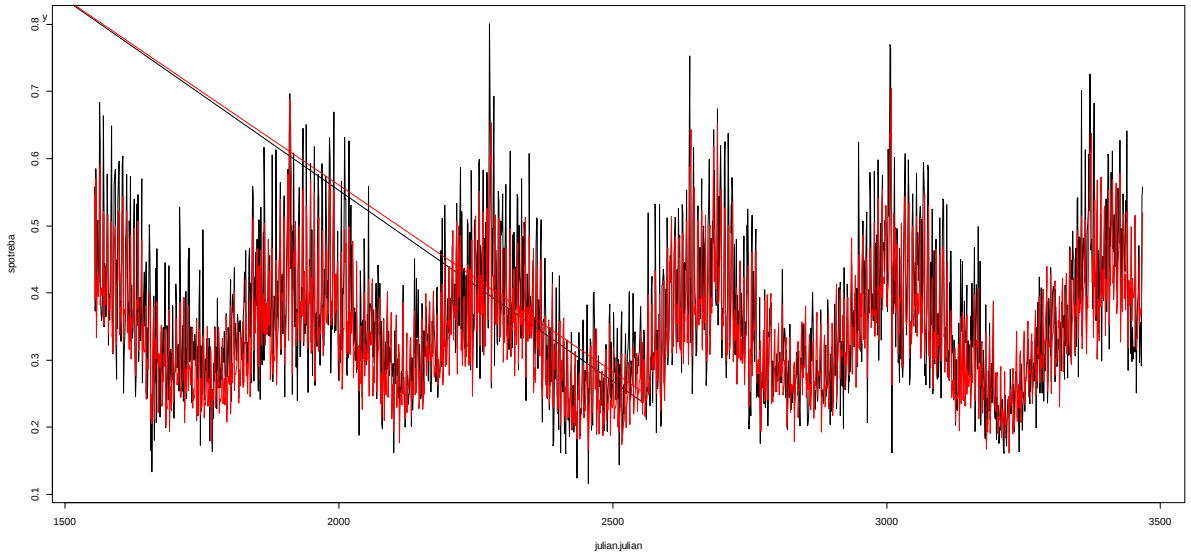
Fit on the continuously measured data
HOU4





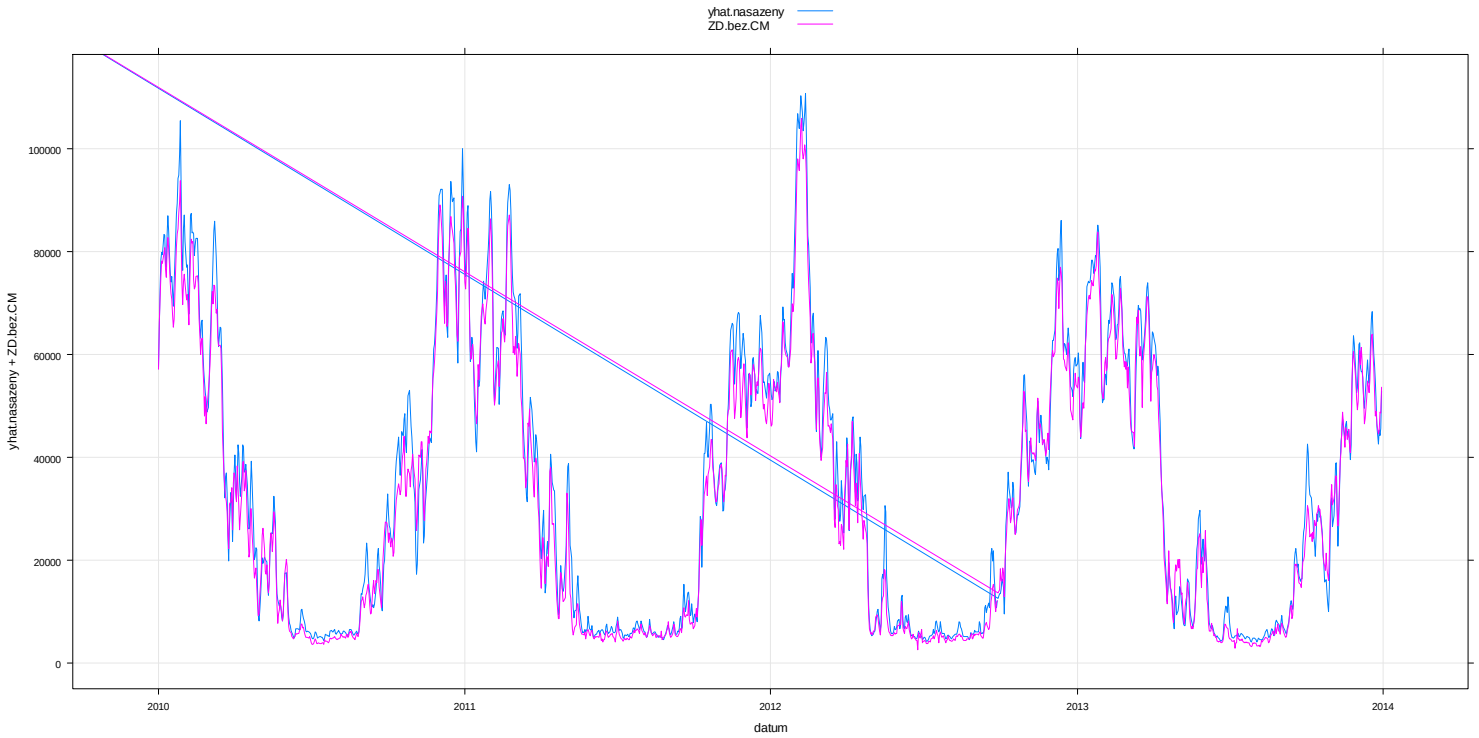
HOU1


yhat
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


y
spotreba
julian.julian
1500
2000
2500
3000
3500
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Out-of-sample performance, GasNet

large HOU+SMC pool of customers (more than 1 mil.)
(followed/read-off routinely – i.e. approx annually)



_ Aggregation

- In the energy context, it is common to meet various forms of aggregation-disaggregation-reaggregation problem

- Motivated by many practical needs

  - e.g. network balancing for technical and financial purposes

- Data are obtained/modeled at one aggregation level and needed at another

  - both finer and coarser might be needed

  - aggregation over time, space, individual customers etc.

Situation

- A nonlinear time series model:

  (with a given function $f_t \equiv f(t)$ and transformation $g(.)$)

$$Y_t = f_t . g(\mu_t) + \varepsilon_t$$
$$\mu_t = \rho . \mu_{t-1} + \nu_t$$

- Observed repeatedly for many (independent) individuals,

- yields (relatively standard) *longitudinal* data

$$\{Y_{it};\ i = 1,\ldots,n;\ t = 1,\ldots,T\}$$

A complication

- Individual observations Y_{it}
  are not accessible, however

- Individual sums over time $Y_{i;r,s} = \sum_{t=r+1}^{s} Y_{it}$
  (individual interval sums over $(r, s]$),
  are available, instead

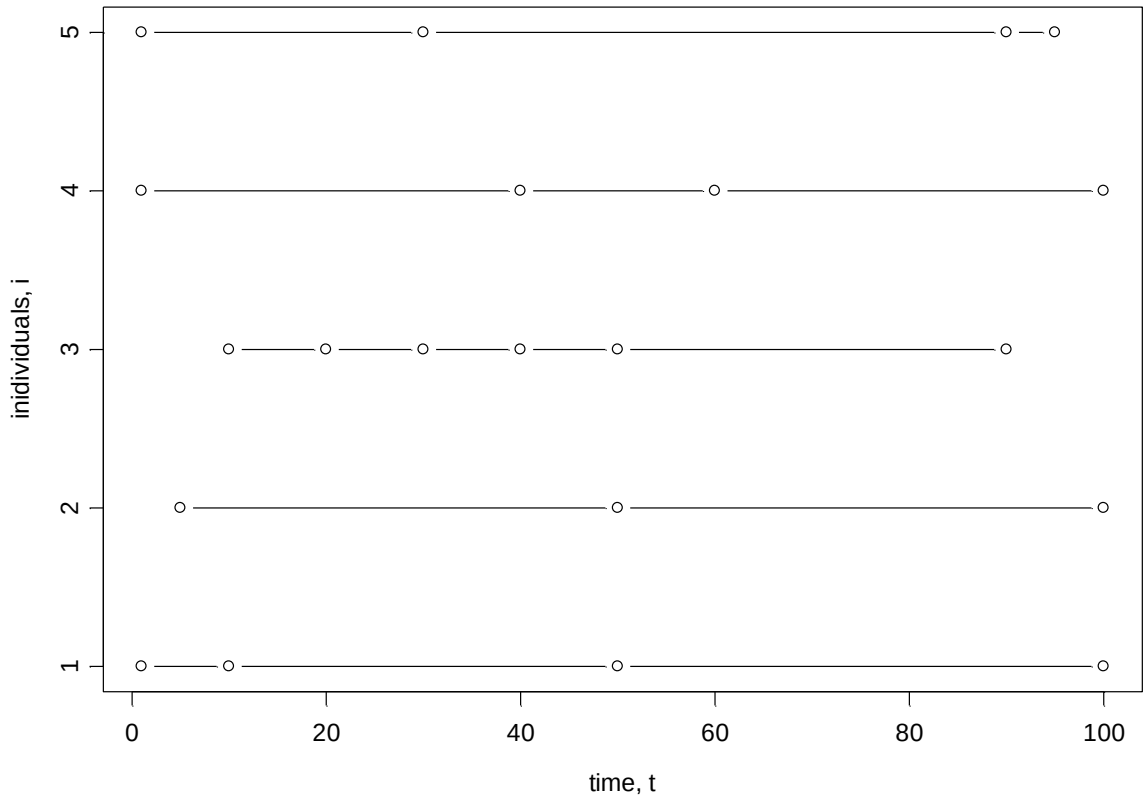
Irregular data

- Lengths and positions of the intervals are available are generally different for different individuals.

- This yields a more complicated data structure
$\left\{ Y_{i;t_{ij},t_{ij+1}};\ i=1,\ldots,n;\ j=1,\ldots,m_i \right\}$

- Different number of observations per individual, in different timing $\left(t_{i_1},t_{i_2}\right],\left(t_{i_2},t_{i_3}\right],\ldots,\left(t_{i_{m_i-1}},t_{i_{m_i}}\right]$

Data as interval sums for individual collection of intervals



Formal problem, I

- In fact, we are dealing with “integrated” (or aggregated) observations

  (integration is done over time for each individual separately)

- But meaningful inference is needed for different levels of aggregation than that available in data. This concerns both:
  - regularity (and interpretability of results)
  - different requirements for time-resolution

Formal problem, II

- This is similar to estimating derivatives of a curve when “integral observations” are available
  (e.g. in growth curves, where inference for growth velocities are based on total length measurements)
- But more complicated
  (observations of process functionals are used to estimate other functionals)
- Similar to MAUP
  (Modifiable Aerial Unit Problem) of Spatial Statistics, Cressie (1996)

For various purposes, utility company needs:

- individual estimates of daily consumption

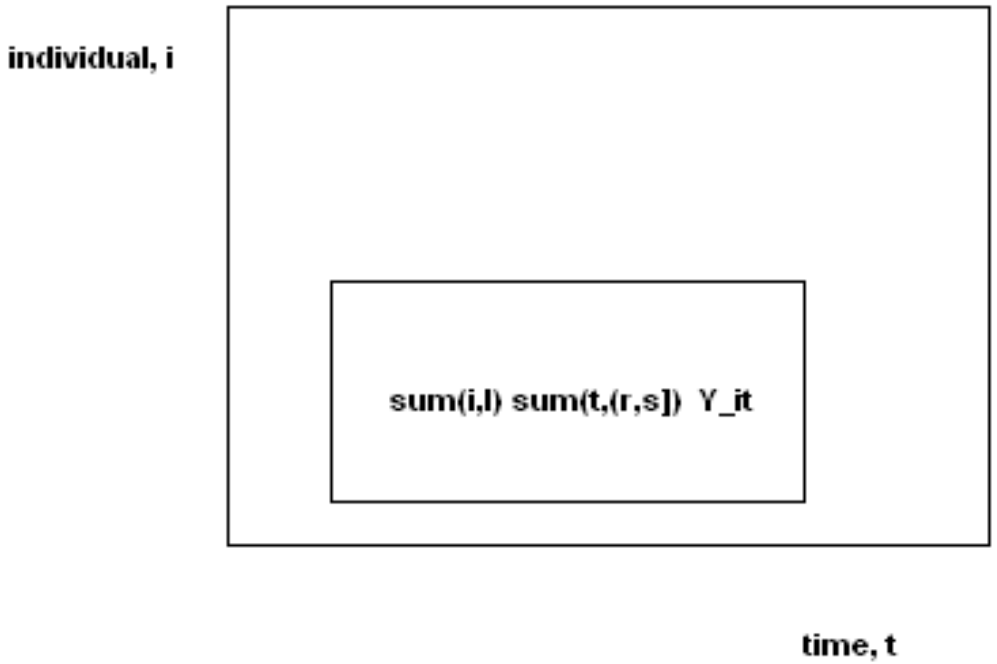
  (finest, daily, or *t*-resolution)

- estimates of double sums both across time and customers, like $C_{s,t} = \sum_{t=r+1}^{s} Y_{it}$

  (in practice, this is generally even more important)

  (regular accounting, price changes, planning transportation capacity, redistributing discrepancies between amount of gas ordered and consumed during a given period, etc.)

Desired data aggregation



State-space model in fine time-resolution

$$g(Y_{it}) = G_i . f_{it} . \exp(\mu_{it}) + \varepsilon_{it}$$

$$\mu_{it} = \rho . \mu_{i,t-1} + \nu_{i,t}$$

with $f_{it} = r_t . \exp(-\gamma . [\min(T_t, 14) - \min(N_t, 14)]) + p$

and a given r_t (yearly periodic, $r_t = r_{t+365}$)

- Independence across *i* and *t*

  $\nu_{it} \sim N(0, \sigma_\nu^2)$, $\varepsilon_{it} \sim N(0, \sigma^2)$, $G_i \sim LN(\mu_G, \sigma_G^2)$

- Initial conditions, $\mu_{i0} = 0$

- Structural parameters $\underline{\theta} = (\sigma^2, \sigma_\nu^2, \rho, \mu_G, \sigma_G^2, \gamma, p)'$

TS view

For a given individual *i*, this is a:

- seasonal (r_t and N_t),
- non-stationary (f_t through T_t),
- nonlinear ($\exp(.)$ transformation of state)

state-space model

Mixed model view

Individual (perhaps not too long, but quite numerous) time series are bound together:

- Individual scaling factors (G_i) are not completely free, but tied by the assumed common distribution ($G_i \sim LN(\mu_G, \sigma_G^2)$)

  (they come from a common population of individuals)

- Nonlinear mixed effects (NLME) type model.

  (producing desirable shrinkage for G_i estimation, among other things)

Estimates from the model

(when structural parameters $\underline{\theta}$ are known)

- Online estimates of de-noised data version are readily obtained by application of
  Extended Kalman Filter (EKF)

- Approximate
  - (one-day ahead) predictor $E\left[G_i . f_t . \exp(\mu_{it}) \mid Y_{i1}, Y_{i2}, \ldots, Y_{i,t-1}, G_i\right]$
  - Filter $E\left[G_i . f_t . \exp(\mu_{it}) \mid Y_{i1}, Y_{i2}, \ldots, Y_{i,t}, G_i\right]$

- Based on local linearization
- Occasional missings do not cause problems

  (just skipping update in EKF)

Estimation of $\underline{\theta}$

- Structural parameters $\underline{\theta} = \left(\sigma^2, \sigma_v^2, \rho, \mu_G, \sigma_G^2, \gamma, p\right)'$ are unknown and have to be estimated from data

  (this amounts to “filter training” and identification of the distribution for individual multipliers G_i)

- ML estimation is employed

- (Approximate) likelihood evaluation is rather easy, and efficient.

  (based on prediction error decomposition)

_ Energy Meteorology

- Assessment of photovoltaic potential uses as much information as possible
- Relative sunshine duration (or its complement, *cloud shade, CS*) is an important local feature to consider
- But, by far, it is not measured everywhere
- One possibility how to get closer measurements is to use (calibrated) point *cloudiness, PC*
- That is measured routinely at many professional meteorological stations

PC calibration to estimate CS

- Certainly, one can estimate CS (conditional) mean, given the PC – this is a regression task and it can be achieved rather easily by spline regression
- If one is interested in the conditional distribution (not just in the mean), the task is more difficult
- The conditional distribution might be needed e.g. for optimization of economically-based loss functions (in the risk asessment)
- Can use GAMLSS extension of the GAM to get the practical solution …

BEINF (beta inflated) model

$$CS_i \sim BEINF(\mu_i, \sigma_i, \nu_i, \tau_i)$$

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = f_\mu(PC_i)$$

$$\log\left(\frac{\sigma_i}{1-\sigma_i}\right) = f_\sigma(PC_i)$$

$$\log(\nu_i) = f_\nu(PC_i)$$

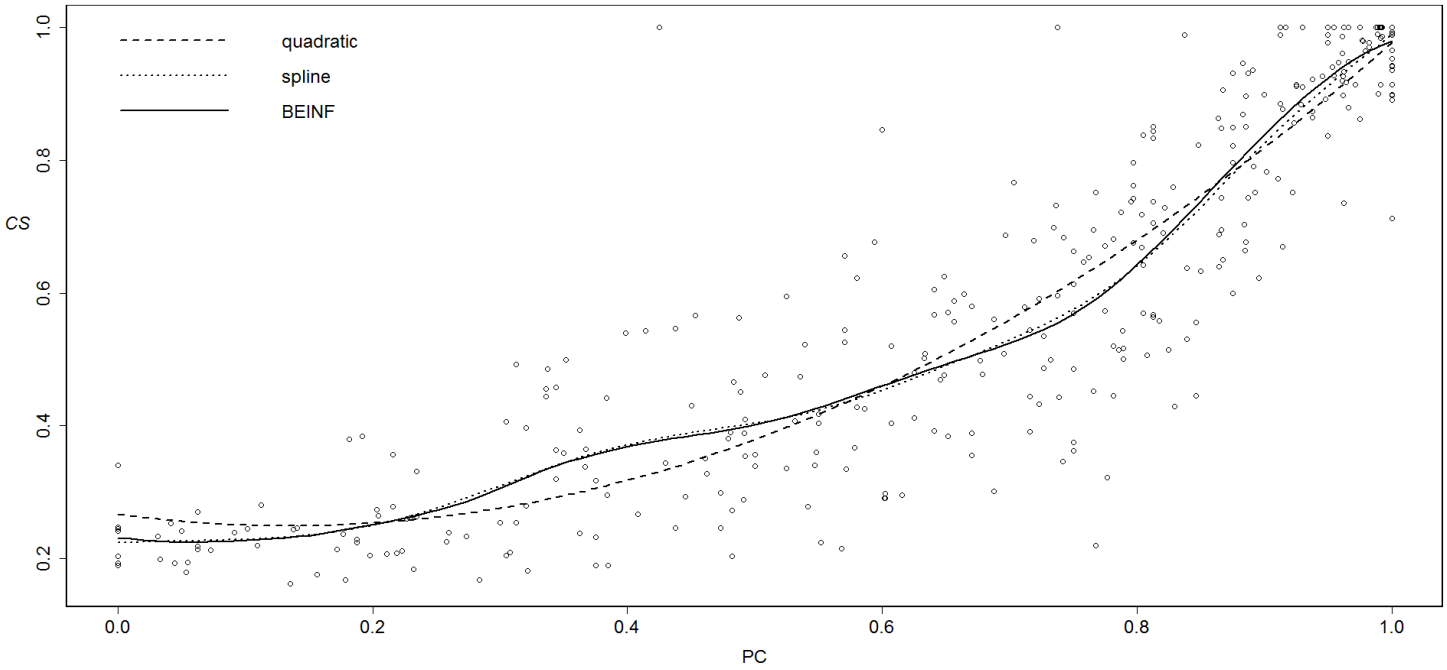
$$\log(\tau_i) = f_\tau(PC_i)$$

- i.e. $CS_i = Bernoull(\pi_{0i}) + Bernoull(\pi_{1i}) + (1-\pi_{0i}-\pi_{1i}).Beta(\alpha_i, \beta_i)$
- where

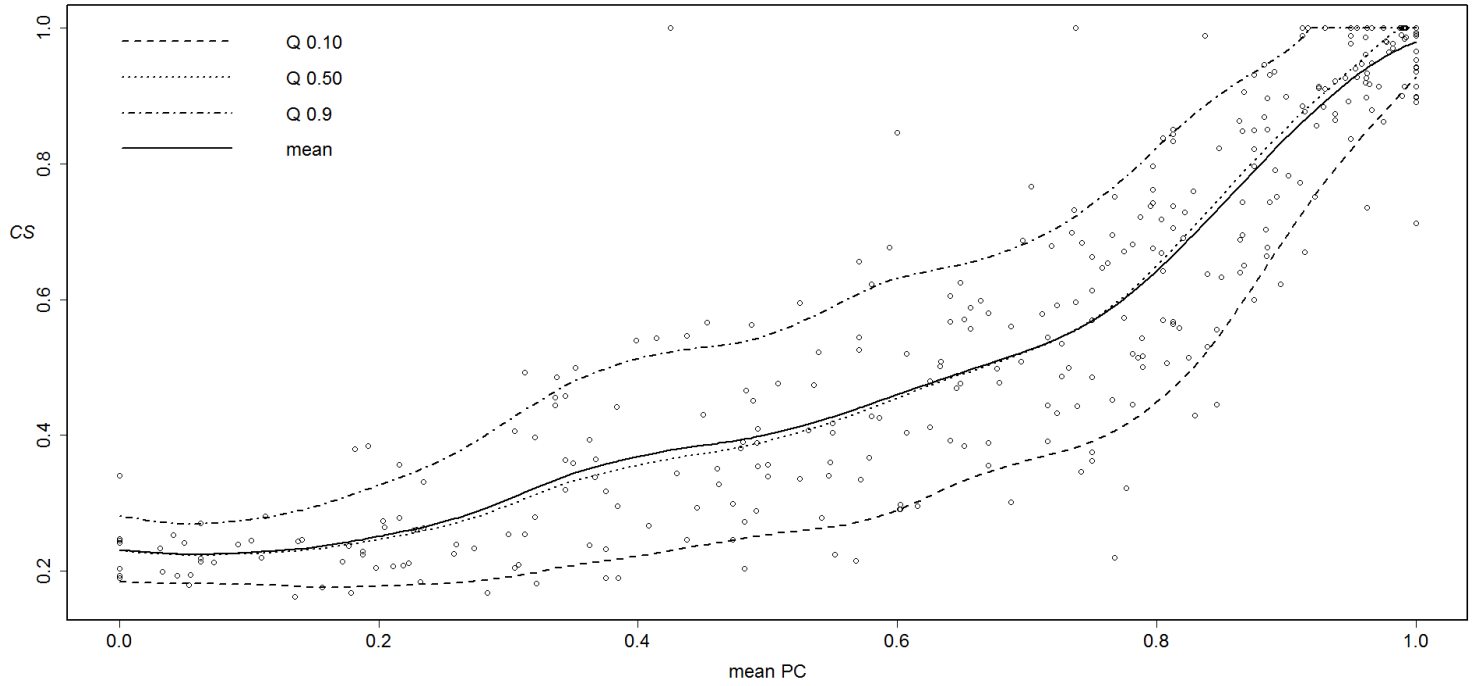
$$\alpha_i = \frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2} \qquad \pi_{0i} = \frac{\nu_i}{1+\nu_i+\tau_i}$$

$$\beta_i = \frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2} \qquad \pi_{1i} = \frac{\tau_i}{1+\nu_i+\tau_i}$$

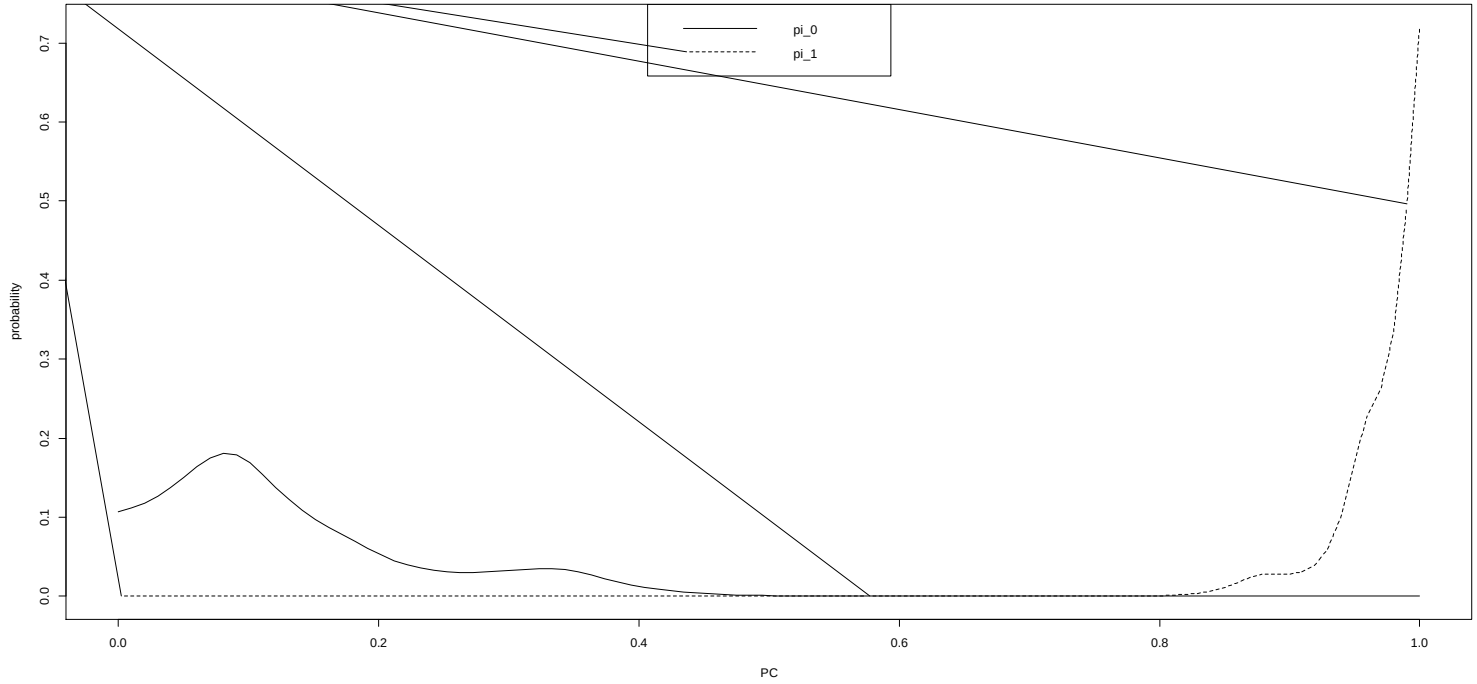
Estimated CS (conditional) mean



Conditional quantiles



More complicated features …
(all obtained consistently from the same model)



Morale

- Statistical modeling offers a unified and flexible methodology to cover many difficult practical tasks arising in energy industry
- The model has to come to the data and underlying problem and not vice versa
- The model should arise in close cooperation of statistician(s) and energy experts
- Purely empirical and purely “mathematical” modeling approaches can be united based on broader umbrella to the benefit of an end-user