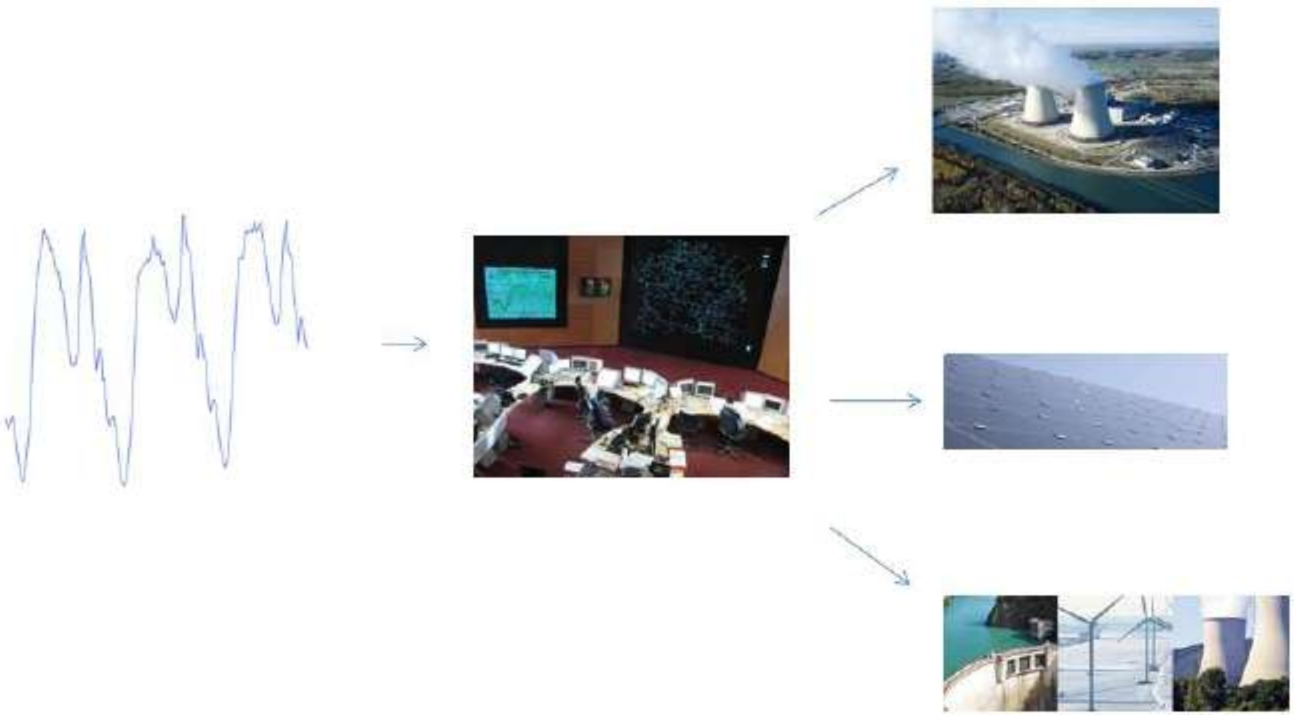
AGGREGATION OF EXPERTS FOR LOAD AND PRICE FORECASTING

Séminaire FIME, 20 Mars 2015, Institut Henri Poincaré

Yannig Goude	EDF R&D, Université Paris-Sud
Pierre Gaillard	EDF R&D, HEC Paris-CNRS
Raphaël Nédellec	EDF R&D
Gilles Stoltz	HEC Paris-CNRS

ELECTRICITY LOAD FORECASTING

- **Electricity consumption is the main entry for optimising the production units**



Focus here on Day+1 forecasts

HEAT DEMAND FORECASTING

- **Optimize the use (satisfy the heat demand, minimize production cost) of a co-generation (heat and electricity) plant in Poland**

Focus on h+1, …, h+72h forecasting

PROB. ELECTRICITY PRICE FORECASTING

- **competition GEFCOM 2014,** sponsored by IEEE Power and Energy Society
  - september 2014-december 2014
  - Probabilistic forecast (quantile 1%,…,99%) of hourly electricity prices in US based on:
    - Zonal/total electricity load forecast
    - Past prices
  - Online forecasting of 15 days
  - Performance evaluation: pin-ball loss

  - Participation (nb of teams): Load (333), Price (250), Wind (208), Solar (218)

Focus on h+1, …, h+24h forecasts

AGGREGATION OF EXPERTS

- **a dynamic field of research in the machine learning community**
- **empirical literature is large and diverse**
- **massive development of new forecasting methods**
  - implementation in open source softwares
  - Easier access to a large variety of forecasts
- **aggregating them is a natural ambition**
- **in many recent forecasting challenges aggregation is a key point that often makes the difference:**
  - the energy forecasting competition GEFCOM12, Hong, T.; Pinson, P. & Fan, S. *Global Energy Forecasting Competition 2012* **International Journal of Forecasting** , 2014, 30, 357 - 363
  - netflix competition Paterek, *A. Predicting movie ratings and recommender systems* - **a monograph** 2012

SEQUENTIAL AGGREGATION OF EXPERTS

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign weight to each expert and we predict

$$\widehat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t \quad \left(= \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \right)$$

Our goal is to minimize our cumulative loss

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{i=1,\dots,N} \sum_{t=1}^{T} (x_{i,t} - y_t)^2}_{\text{Loss of the best expert}} + \underbrace{R_T}_{\text{Estimation error}}$$

Good set of experts

Good aggregating algorithm

Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. **Cambridge University Press** (2006)

SEQUENTIAL AGGREGATION OF EXPERTS

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign weight to each expert and we predict

$$\widehat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t \quad \left(= \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \right)$$

Our goal is to minimize our cumulative loss

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{\boldsymbol{q} \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convex combination}} + \underbrace{R_T}_{\text{Estimation error}}$$

Good set of experts
As varied as possible

Good aggregating algorithm

EXPONENTIALLY WEIGHTED AVERAGE FORECASTER (EWA)

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

$$\widehat{p}_{i,t} = \frac{exp\left(-\eta \sum_{s=1}^{t-1} (x_{i,s} - y_s)^2\right)}{\sum_{j=1}^{N} exp\left(-\eta \sum_{s=1}^{t-1} (x_{j,s} - y_s)^2\right)}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

Our cumulated loss is upper bounded by

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{i=1,\dots,d} \sum_{t=1}^{T} (x_{i,t} - y_t)^2}_{\text{Loss of the best expert}} + \underbrace{\square \sqrt{T \log N}}_{\text{Estimation error}}$$

EXPONENTIATED GRADIENT FORECASTER (EG)

Each instance t

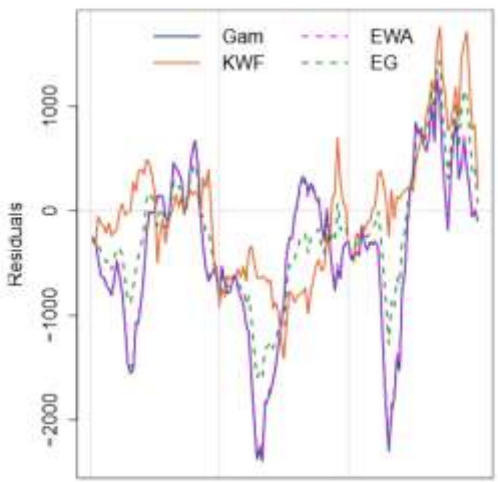
- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

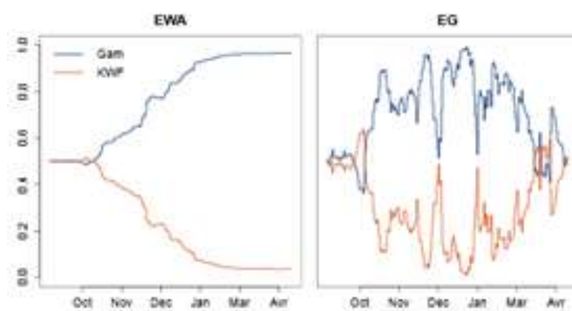
$$\widehat{p}_{i,t} \propto exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right) \qquad \text{where} \quad \ell_{i,s} = 2(\widehat{y}_s - y_s)x_{i,s}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

Our cumulated loss is then bounded as follow

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{q \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convexe combination}} + \underbrace{\square \sqrt{T \log N}}_{\text{Estimation error}}$$





MULTIPLE LEARNING RATE-POLYNOMIAL, RIDGE

Algorithm 1 The polynomially weighted average forecaster with multiple learning rates (ML-Poly)

Input: $h \geqslant 1$, horizon of prediction
Initialize: For $t \leqslant h$, $\boldsymbol{p}_t = (1/K, \dots, 1/K)$ and $\boldsymbol{R}_1 = (0, \dots, 0)$
for each instance $t = 1, 2, \dots, n-h$ **do**
0. pick the learning rates
$$\eta_{k,t} = 1 / \left(1 + \sum_{s=1}^{t} \left(\ell_s(\widehat{y}_s) - \ell_s(x_{k,s})\right)^2\right)$$
where $\ell_s : x \mapsto x(y_s - \widehat{y}_s)$.
1. form the mixture $\widehat{\boldsymbol{p}}_{t+h}$ defined component-wise by
$$\widehat{p}_{k,t+h} = \eta_{k,t} \left(R_{k,t}\right)_+ \; / \; \left[\boldsymbol{\eta}_t \cdot \left(\boldsymbol{R}_t\right)_+\right]$$
where $\boldsymbol{x}_+$ denotes the vector of non-negative parts of the components of $\boldsymbol{x}$
2. predict $\widehat{y}_{t+h} = \widehat{\boldsymbol{p}}_{t+h} \cdot \boldsymbol{x}_{t+h}$ and observe y_{t+1}
3. for each expert k update the regret
$$R_{k,t+1} = R_{k,t} + \ell_t(\widehat{y}_{t+1}) - \ell_t(x_{k,t+1})$$
end for

Gaillard, P., Stoltz, G., van Erven, T.: *A second-order bound with excess losses*, **COLT proceedings** (2014).

Automatic calibration works well in practice

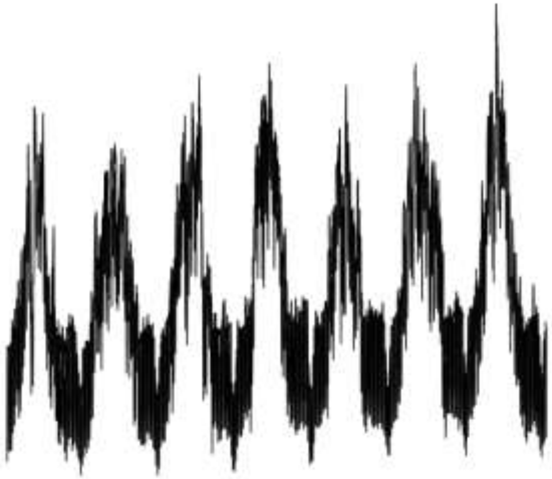
Fast tuning

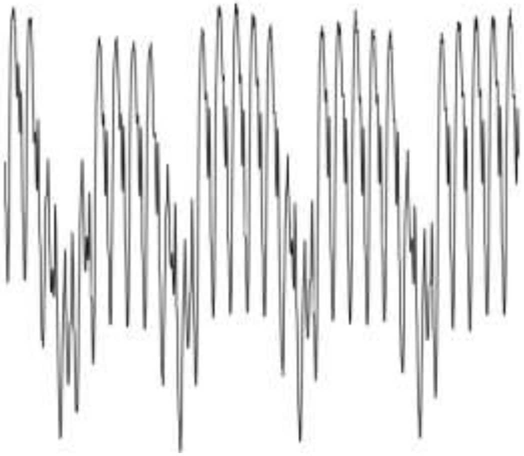
Algorithm 2 The ridge regression forecaster (Ridge)

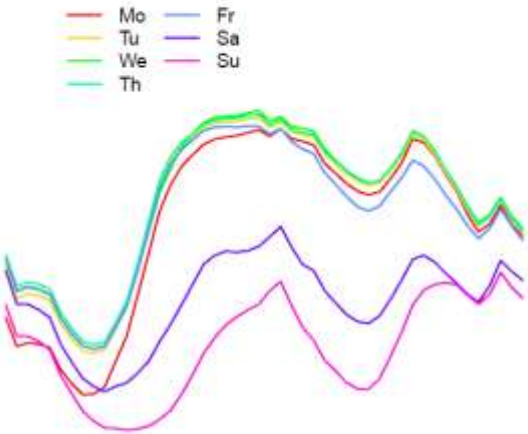
Input: $\lambda > 0$, learning rate; $h \geqslant 1$, horizon
Initialize: for $t \leqslant h$, $\widehat{\boldsymbol{p}}_t = (1/K, \dots, 1/K)$
for each instance $t = 1, 2, \dots, n$ **do**
1. form the mixture $\widehat{\boldsymbol{p}}_{t+h}$ defined by
$$\widehat{\boldsymbol{p}}_t = \underset{\boldsymbol{u} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \sum_{s=1}^{t} \left(y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s\right)^2 + \lambda \left\|\boldsymbol{u} - \widehat{\boldsymbol{p}}_0\right\|_2^2 \right\}$$
2. output prediction $\widehat{y}_{t+h} = \widehat{\boldsymbol{p}}_{t+h} \cdot \boldsymbol{x}_{t+h}$
end for

Stable weights

ELECTRICITY CONSUMPTION DATA

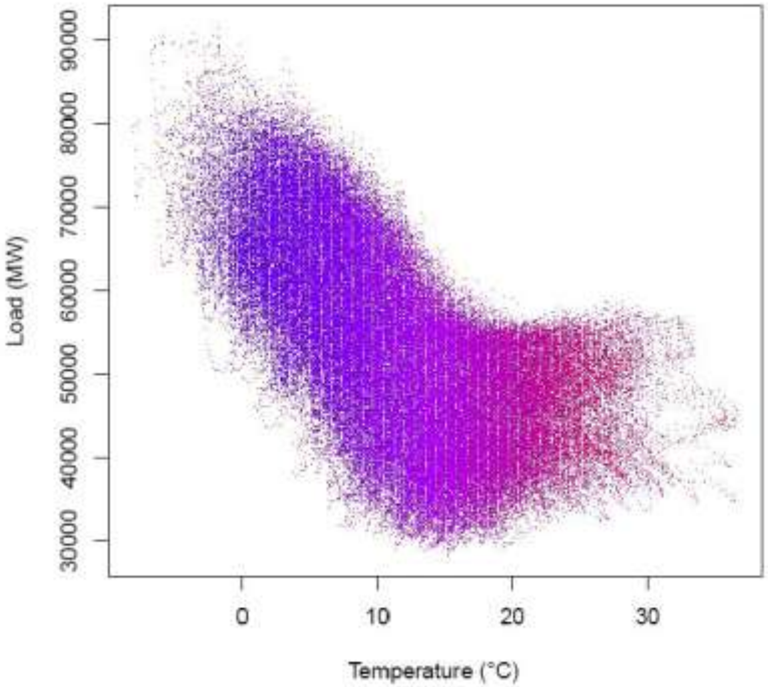


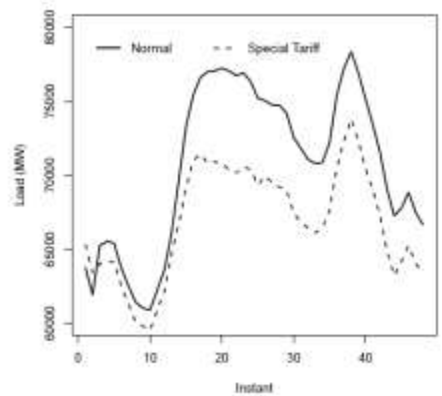


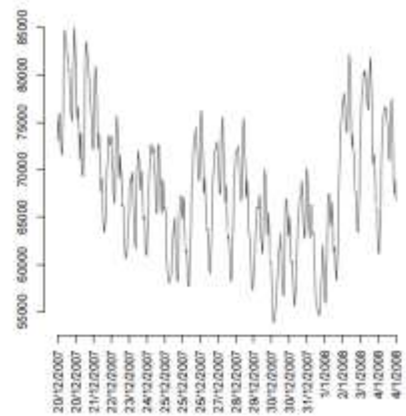


- **Trend**
- **Yearly, Weekly, Daily cycles**

ELECTRICITY CONSUMPTION DATA







- **meteorological events**
- **Special days**

GAM (GENERALIZED ADDITIVE MODELS)

- **A good trade-off complexity/adaptivity**

$$y_t = f_1(x_t^1) + f_2(x_t^2) + \ldots + f(x_t^3, x_t^4) + \ldots + \varepsilon_t$$

$$min_{\beta, f_j} ||y - f_1(x_1) - f_2(x_2) - \ldots||^2 + \lambda_1 \int f_1''(x)^2 dx + \lambda_2 \int f_2''(x)^2 dx + \ldots$$

- **Publications**
  - Application on load forecasting
    - A. Pierrot and Y. Goude, *Short-Term Electricity Load Forecasting With Generalized Additive Models* **Proceedings of ISAP power**, pp 593-600, 2011.
    - R. Nédellec, J. Cugliari and Y. Goude, *GEFCom2012: Electricity Load Forecasting and Backcasting with Semi-Parametric Models*, **International Journal of Forecasting** , 2014, 30, 375 - 381.
    - S.N. Wood, Goude, Y. and S. Shaw, *Generalized additive models for large datasets*, **Journal of Royal Statistical Society-C**, 2014.
    - A. Ba, M. Sinn, Y. Goude and P. Pompey*, Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting* **Advances in Neural Information Processing Systems** 25, 2012, 2519-2527.

GAM (GENERALIZED ADDITIVE MODELS)

- **Spline basis expansion:**

$$f_j(x) = \sum_{q=1}^{k_j} a_{j,q}(x)\beta_{j,q} \qquad y_i = X_i\beta + \sum_{q=1}^{k_1} a_{1,q}(x)\beta_{1,q} + \sum_{q=1}^{k_2} a_{2,q}(x)\beta_{2,q} + ... + \varepsilon_i$$

- **L2 Penalized regression, GCV score optimisation**

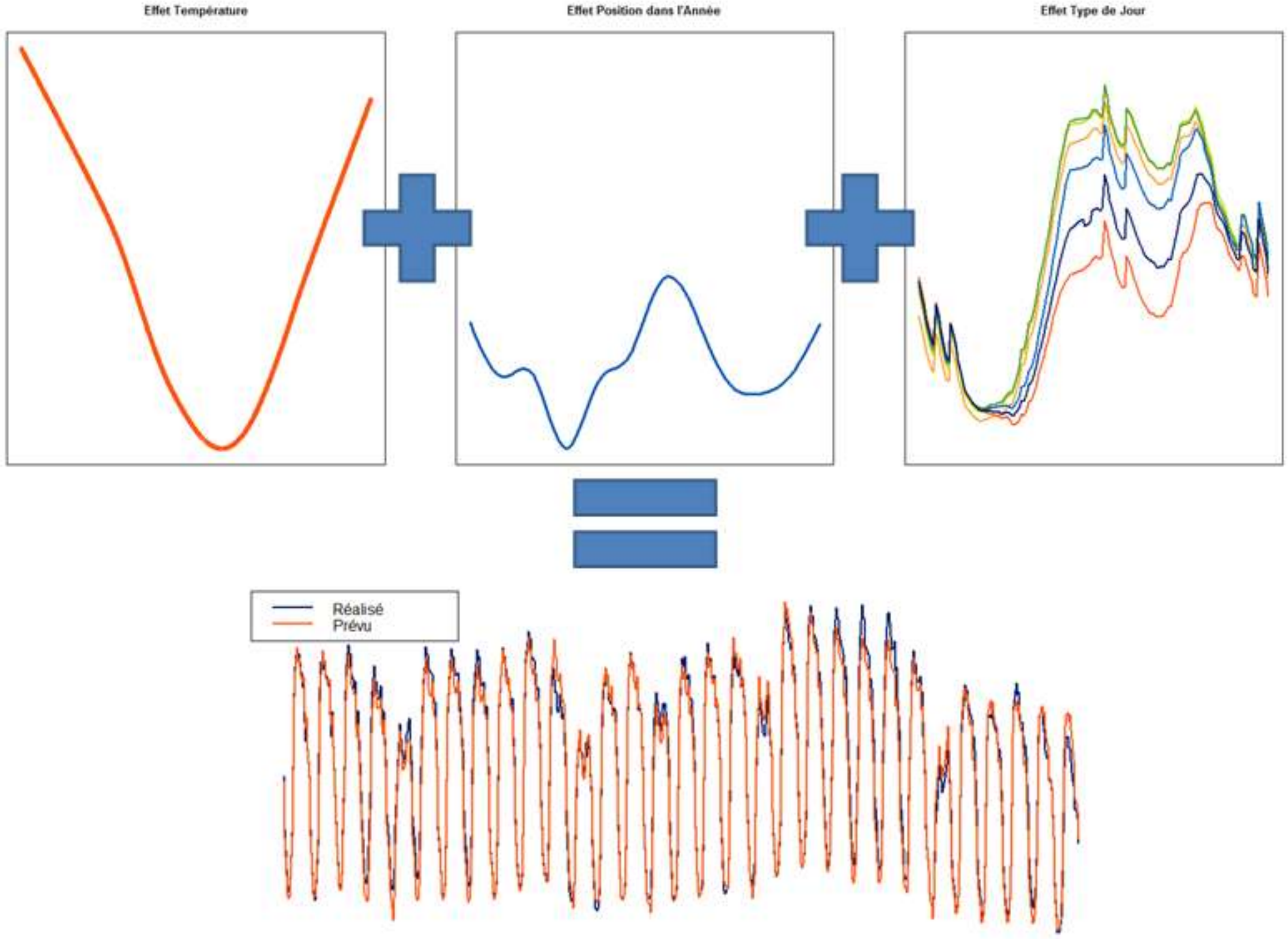
$$min_\beta ||y - X\beta||^2 + \sum \lambda_j \beta^T S_j \beta$$

$$V_g(\lambda) = n\|y - X\widehat{\beta}_\lambda\|^2/(n - tr(F_\lambda))^2$$

$$F_\lambda = (X^T X + \sum \lambda_j S_j)^{-1} X^T X$$

GAM

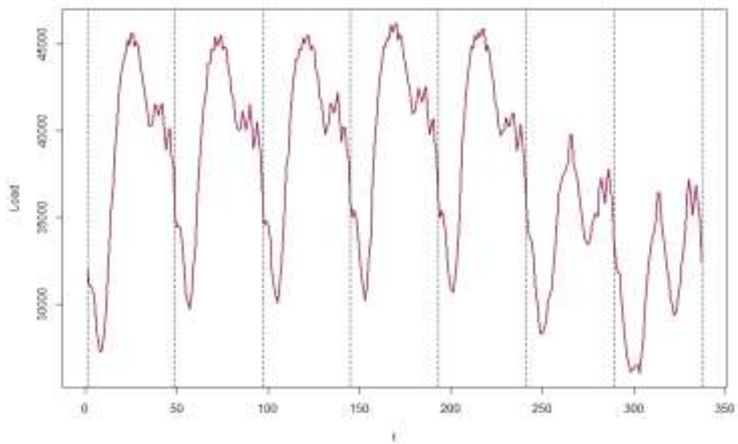
$$y_t = f_1(T_t) + f_2(I_t) + f_3(H_t) + \varepsilon_t$$



CURVE LINEAR REGRESSION

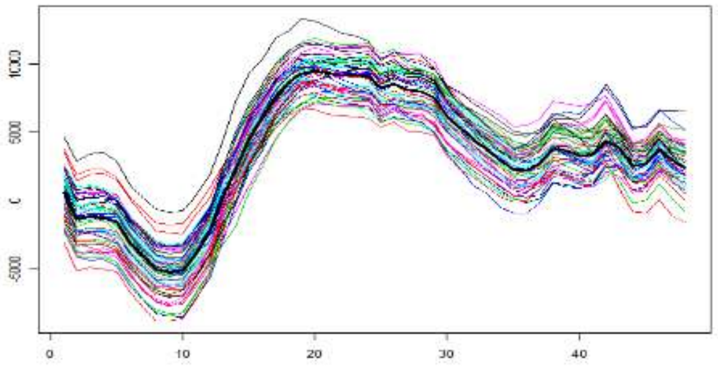
- **Regressing curves on curves**
  - Dimension reduction, SVD of cov(Y,X) , selection with penalised model selection
  - Scale to big data sets (SVD+linear regression)

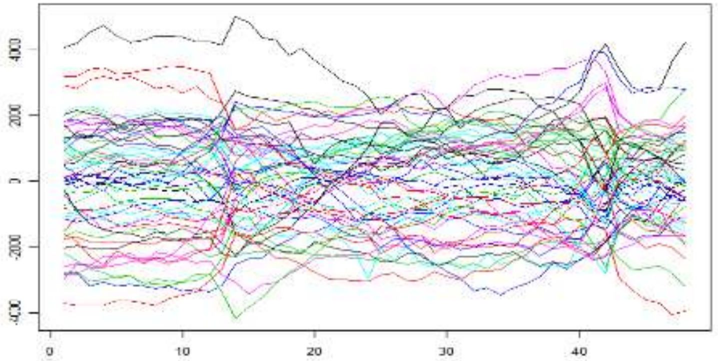
$$Y_i(u) = \mu_Y(u) + \int_{\mathcal{I}_2} \{X_i(v) - \mu_X(v)\} \beta(u,v) dv + \varepsilon_i(u)$$

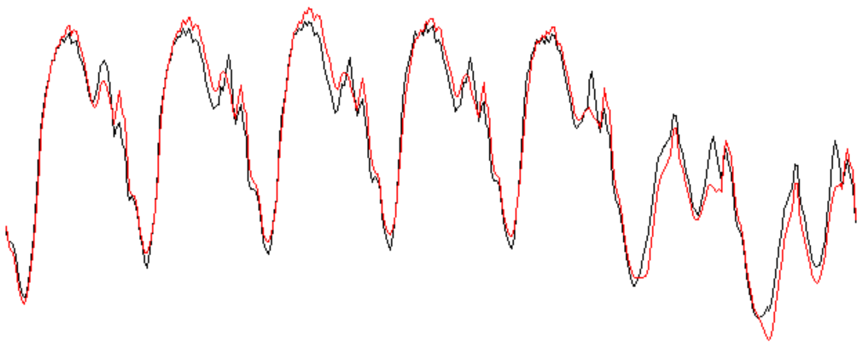


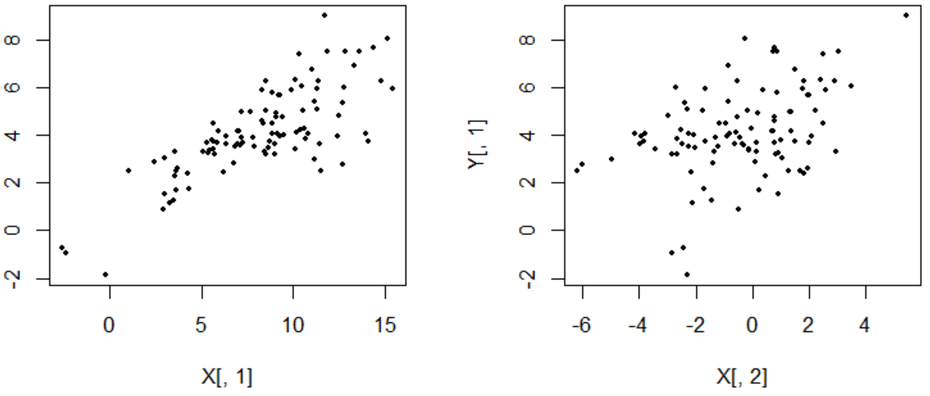
  - Application on electricity load forecasting:
    - H. Cho, Y. Goude, X. Brossat & Q. Yao, *Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach* **Journal of the American Statistical Association**, 2013, 108, 7-21.
    - Cho, H.; Goude, Y.; Brossat, X. & Yao, Q, *Modelling and forecasting daily electricity load using curve linear regression* to appear in **Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension**.

CURVE LINEAR REGRESSION







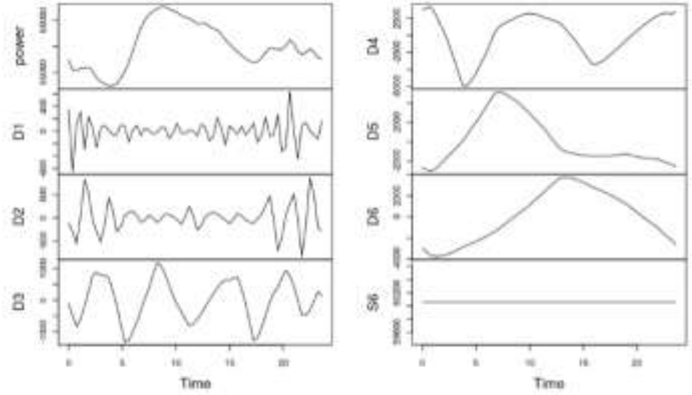


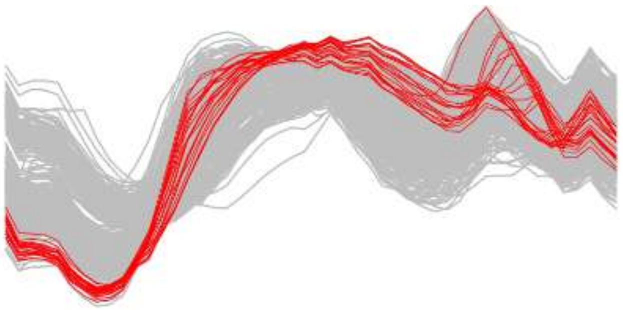
OTHER MODELS

- **Random forest: a popular machine learning method for classification/regression**
  - Breiman, L., . Random Forests, **Machine Learning,** 45 (1), 2001.

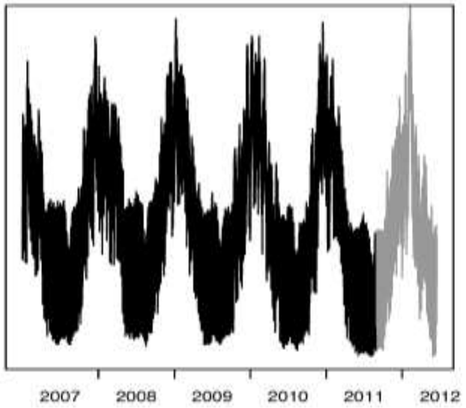
- **Generalized Boosted Regression Models**
  - Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009). *"10. Boosting and Additive Trees". The Elements of Statistical Learning (2nd ed.)*. **New York: Springer**. pp. 337–384.
  - Ridgeway, Greg (2007). *Generalized Boosted Models: A guide to the gbm package*.

- **KWF (Kernel Wavelet Functional): another approach for functional data forecasts**
  - See: Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J., *Clustering functional data using wavelets.* In: **Proceedings of the Nineteenth International Conference on Computational Statistics(COMPSTAT),** 2010.
  - Antoniadis, A., Paparoditis, E., Sapatinas, T., *A functional wavelet–kernel approach for time series prediction.* **Journal of the Royal Statistical Society: Series B** 68(5), 837–857, 2006.
  - *Prévision non paramétrique de processus à valeurs fonctionnelles. Application à la consommation d'électricité*, Jairo Cugliari, **PhD Université Paris-Sud,** 2011.





APPLICATION ON LOAD FORECASTING



- **initial « heterogenous » experts:**
  - GAM
  - Kernel Wavelet Functional
  - Curve Linear Regression
  - Random Forest

- **Designing a set of experts from the original ones: 4 « home made » tricks**
  - Bagging: 60 experts
  - Boosting: trained on $y'_t = (y_t - \gamma x_t)/(1-\gamma)$ such that $\gamma x_t + (1-\gamma)x'_t$ performs well
    45 experts
  - Specializing: focus on cold/warm days, some periods of the year… 24 experts
  - Time scaling: MD with GAM, ST with the 3 initial experts

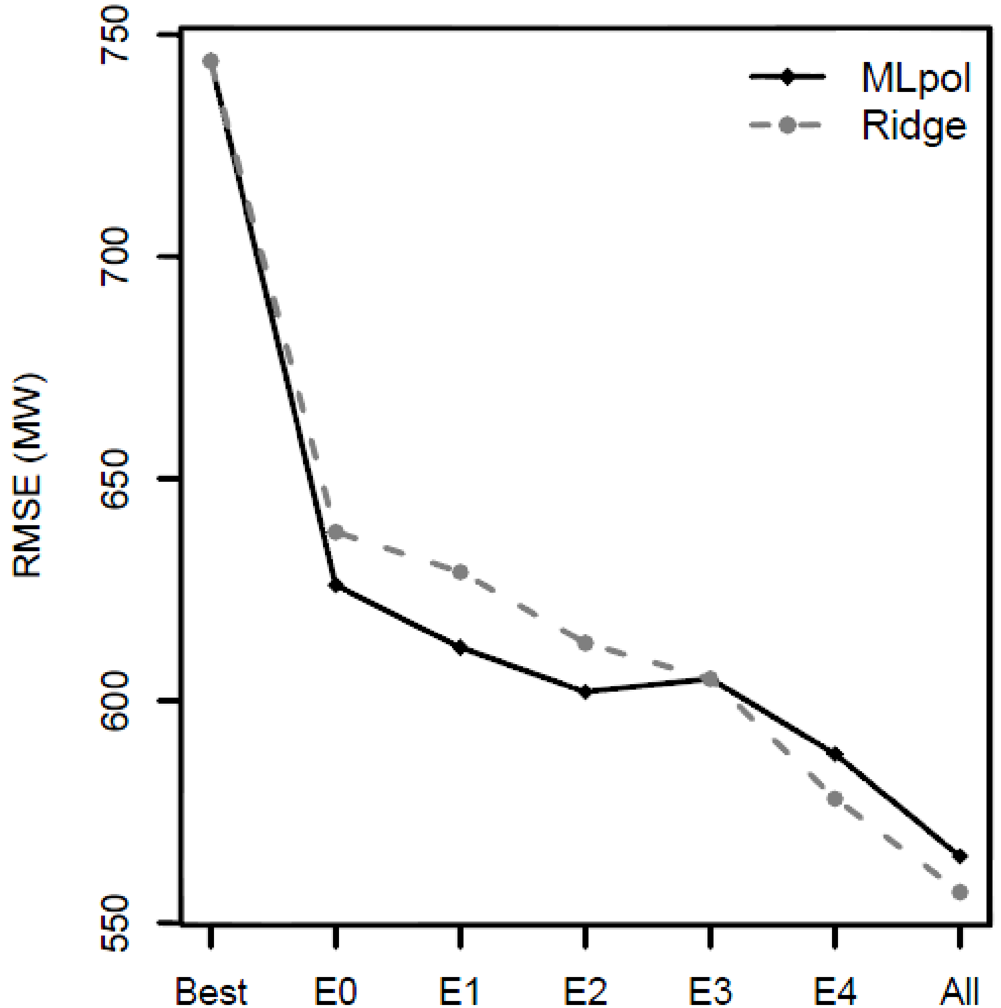
- M. Devaine, P. Gaillard, Y. Goude & G. Stoltz, *Forecasting electricity consumption by aggregating specialized experts - A review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions* **Machine Learning**, 2013, 90, 231-260.

- Gaillard, P. & Goude, Y., *Forecasting electricity consumption by aggregating experts; how to design a good set of experts* to appear in **Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension**, 2013.

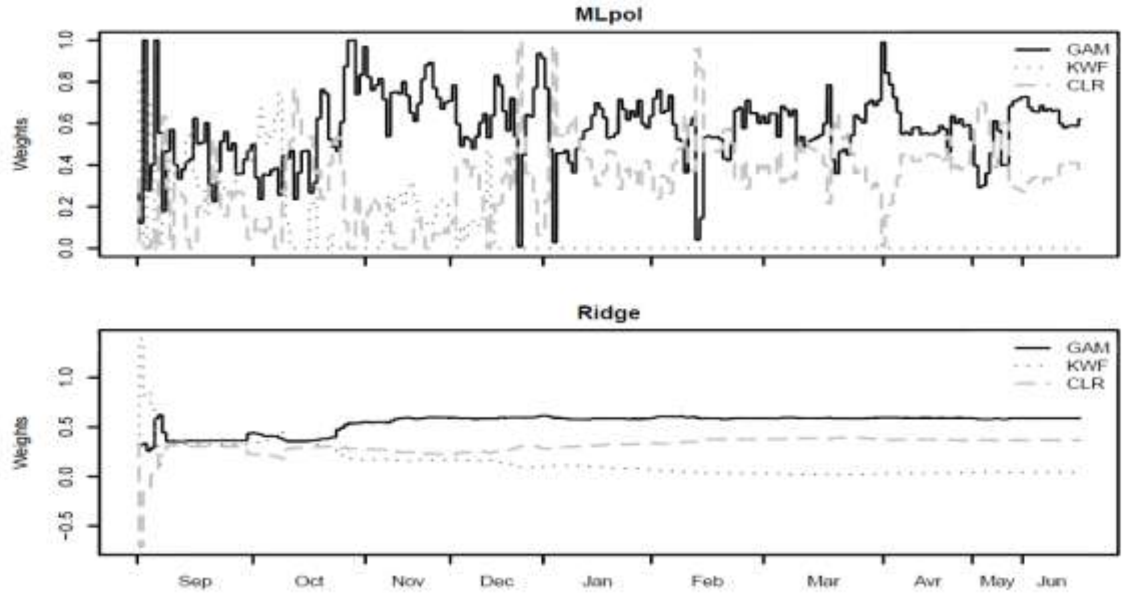
AGGREGATION

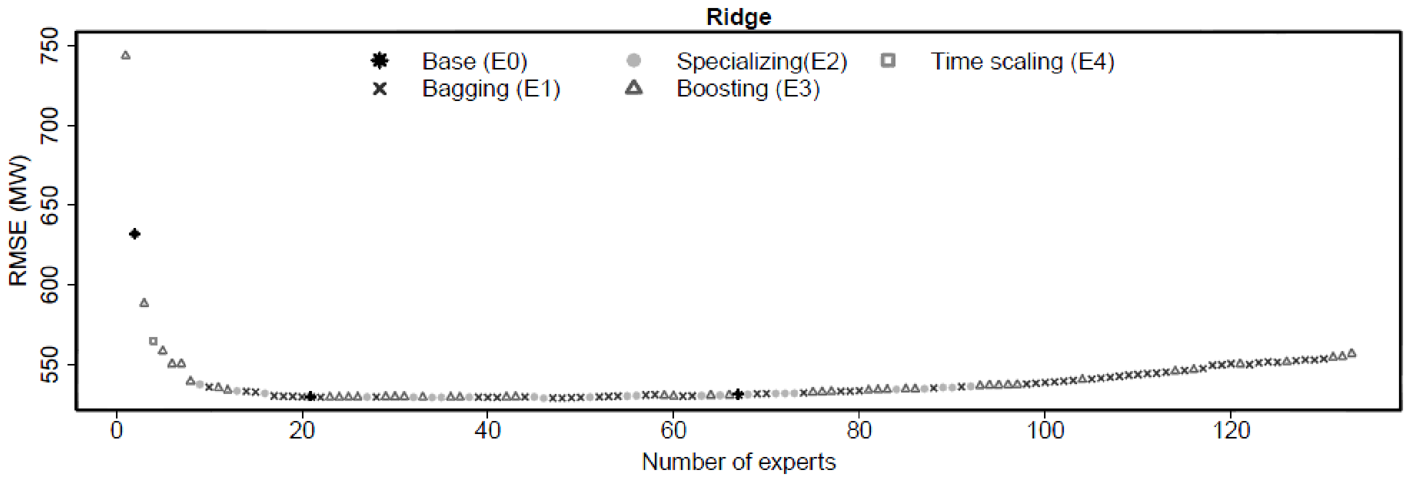
Combining
gain de 110MW

Designing experts
gain de 90MW



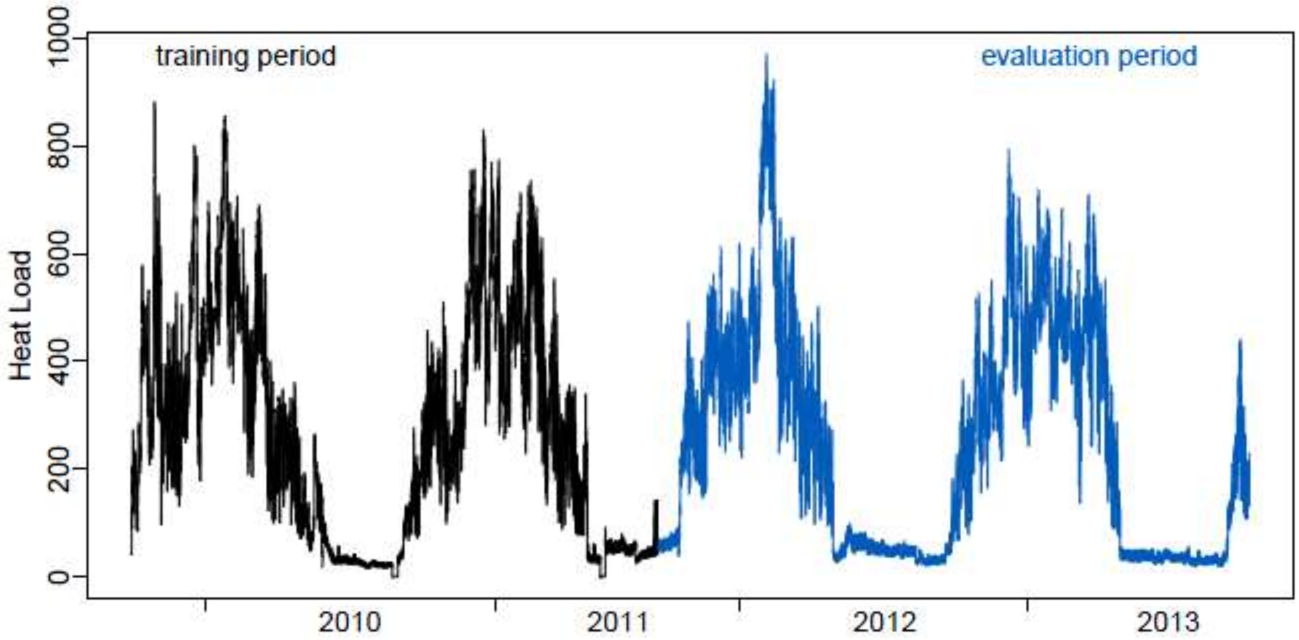
AGGREGATION





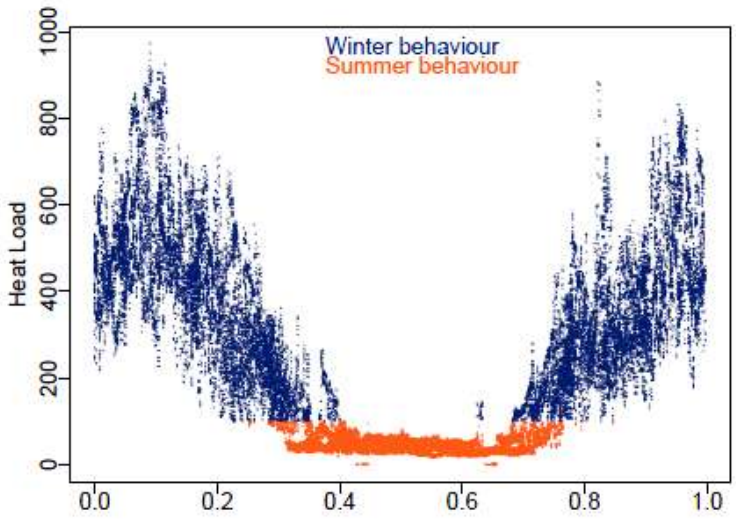
eDF

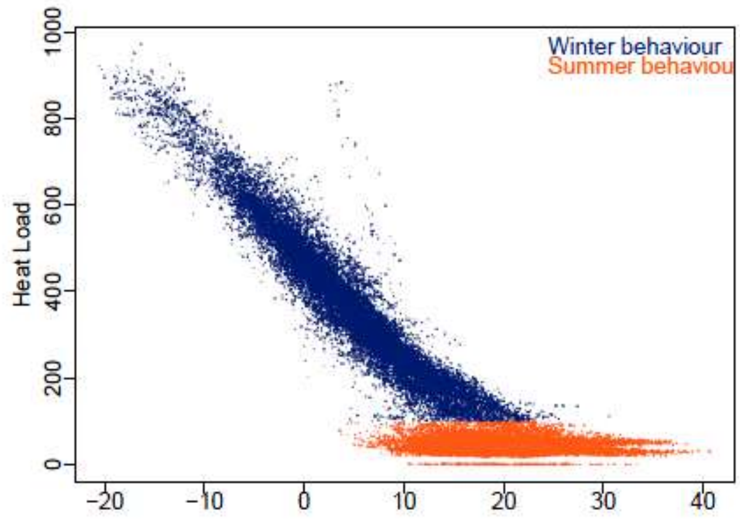
Heat Demand forecasting



- Temperatures: T_t^{ECK}, the outside temperature; T_t^{KR}, the temperature at the airport nearby; T_t^{w}, the temperature of the water leaving the plant; T_t^{zad}, the setpoint temperature of the water asked by the network operator a few hours in advance (not known at the time of prediction);
- Z_t^{KR}, the cloud cover ;
- Calendar variables: $Toy_t \in [0, 1]$, position in the year ; $DayType_t \in \{\text{Monday}, \ldots, \text{Sunday}\}$.

2 REGIMES





- winter : heavy correlation between the outside temperature and the load. The primary goal of the plant is to product heat and electricity comes as an extra.
- summer : low correlation between the outside temperature and the load. There is no need to produce heating, and the primary goal of the plant is to produce electricity.

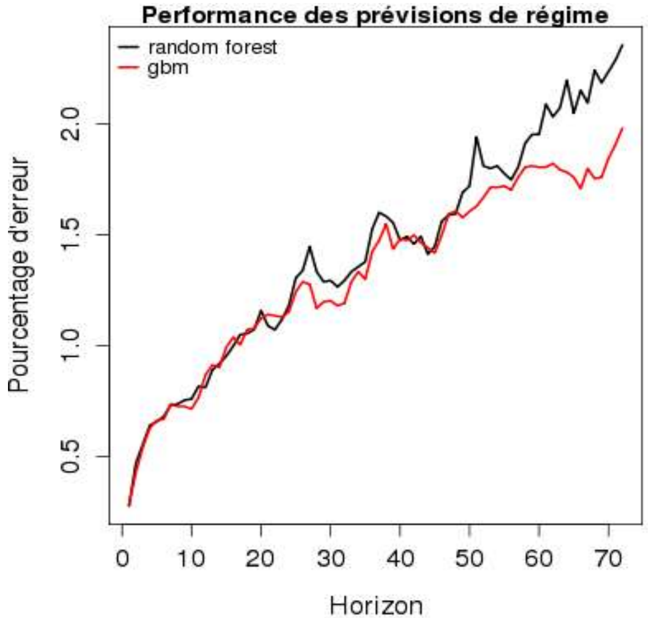
$$S_t = \mathbb{1}_{Y_t > 100}$$

FORECASTING MODELS

GAM: Generalized Addtitive Models
GAMMTCT: GAM middle Term+Short term correction
CLR: Curve Linear Regression
} *forecast Q/S*

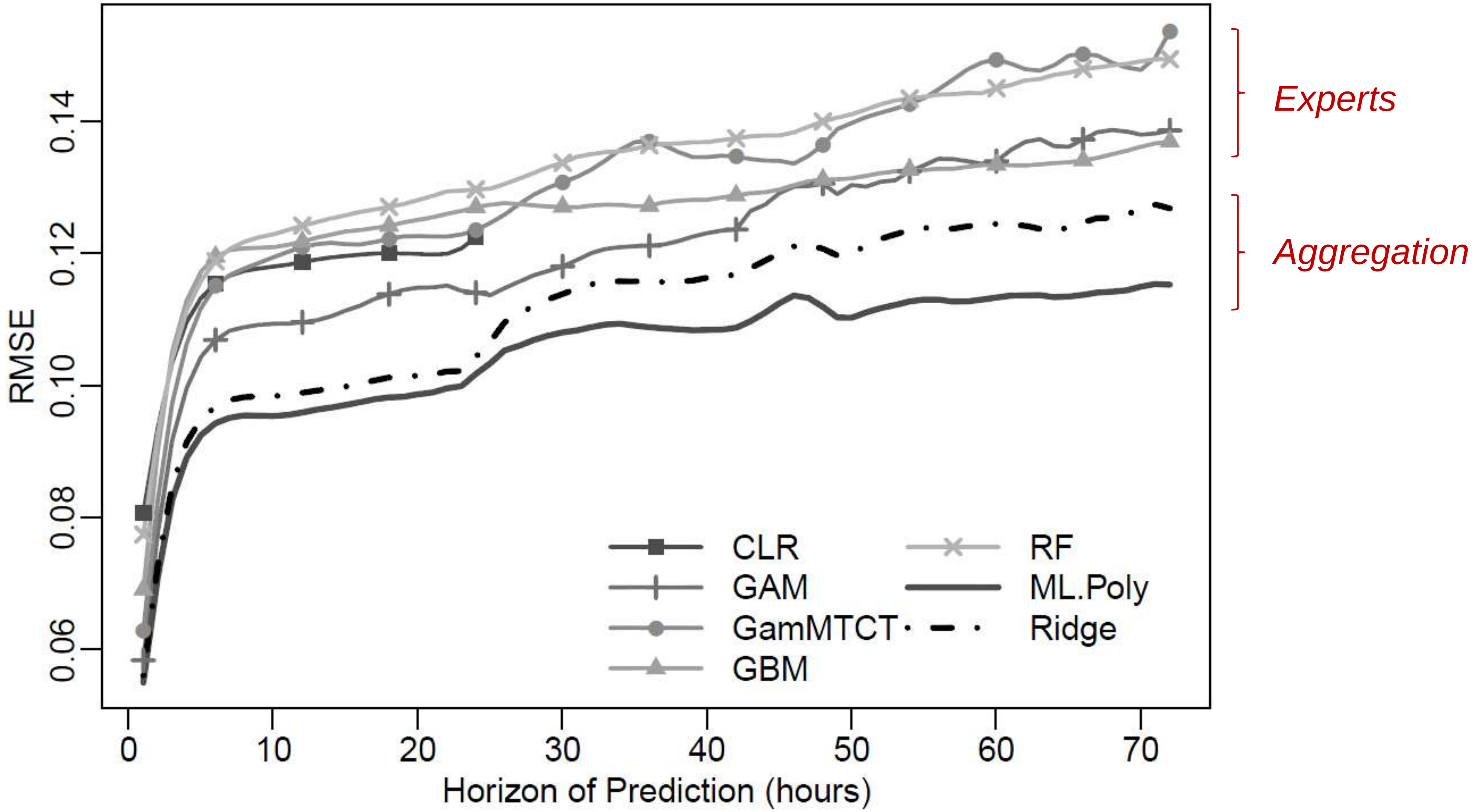
GAM: Generalized Addtitive Models } *forecast Q*

RF: Random Forest
GBM: Gradient Boosting Machine
} *forecast the regime S and Q/S*

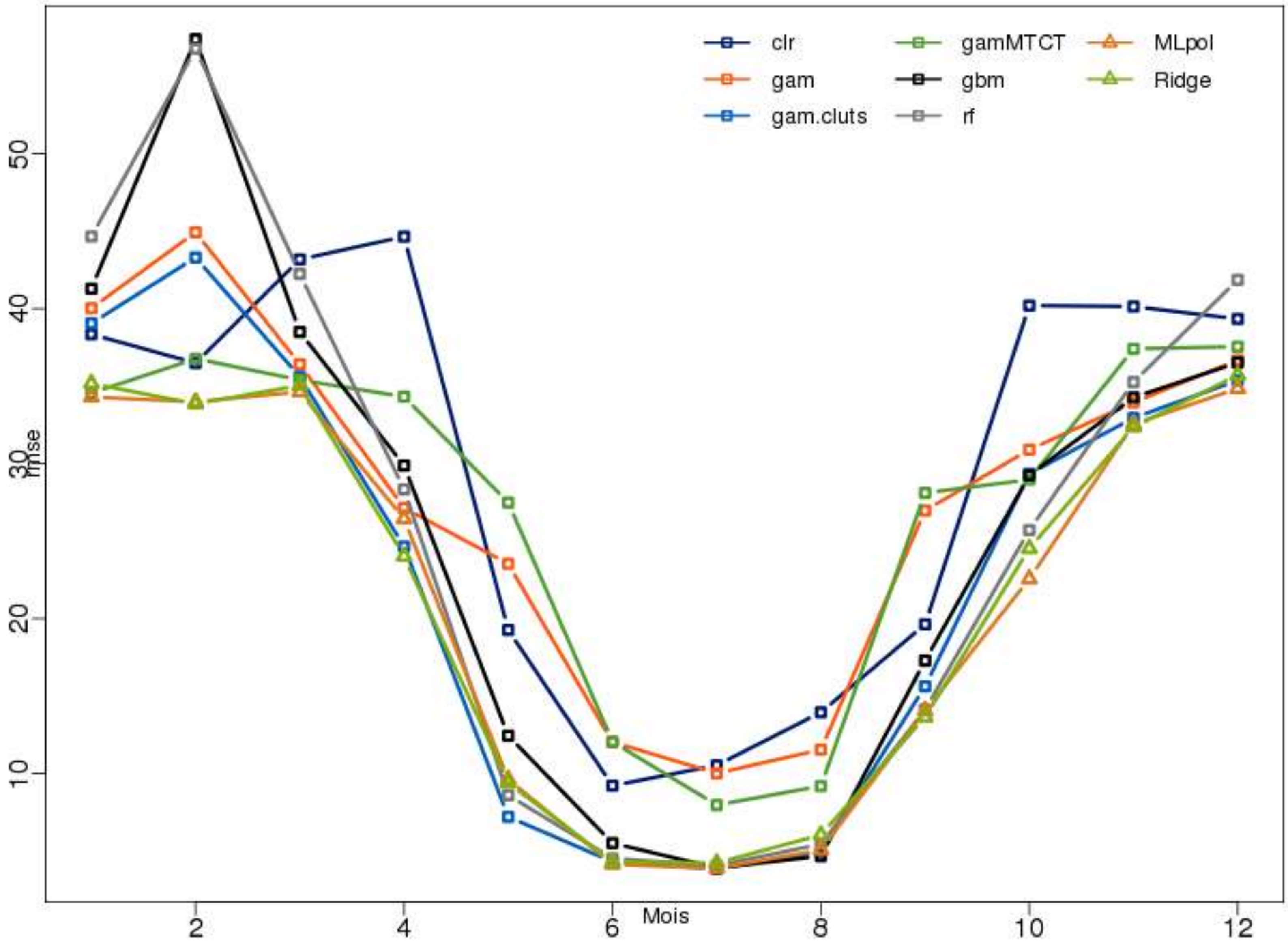


horizon of prediction going from 1 hour ahead to 72 hours ahead

RESULTS : rmse by horizon



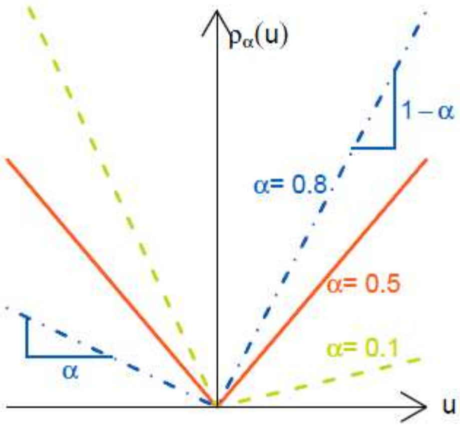
RESULTS : rmse by month *(horizon: 24 h)*



PROB. ELECTRICITY PRICE FORECASTING
GEFCOM14

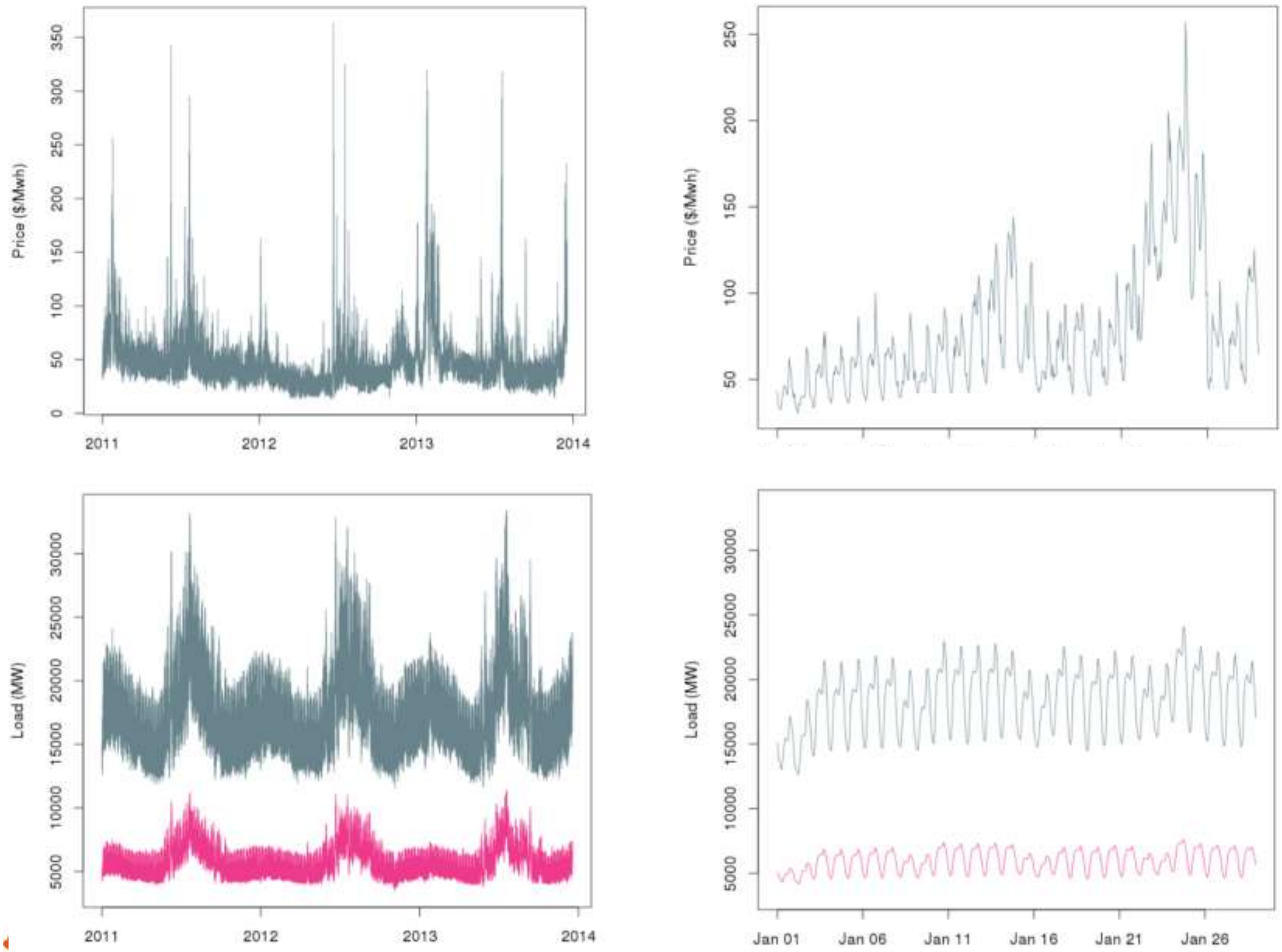
- **competition GEFCOM 2014, sponsored by IEEE Power and Energy Society**
  - Online forecasting of 15 days
  - Performance evaluation: pin-ball loss

$$L(q_a, y) = \begin{cases} (1 - a/100)(q_a - y), & \text{if } y < q_a; \\ a/100(y - q_a), & \text{if } y \geq q_a; \end{cases}$$



- **We proposed 3 methods:**
  - **Aggregation of 13 experts**
  - Non linear quantile regression: GAM quantile
  - Quantile Kernel lasso selection

ELECTRICITY PRICE DATA



PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **Aggregation of 13 experts:**
  - autoregressive model (AR)

$$\log(P_t) = \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-48}) + \alpha_3 \log(P_{t-168}) + \alpha_4 \log(P.\min_{t-24}) + h(\text{DayType}_t) + \varepsilon_t$$

  - An autoregressive model with forecasted electric loads as additional covariates (ARX).
  - A threshold autoregressive model TAR defined as an extension of AR to two regimes depending of the variation of the mean price between a day and eight days ago.
  - TARX the extension of ARX to the two regimes model.
  - Spike pre-processed autoregressive model PAR
  - PARX similar to PAR, but ARX is fitted with pre-processed prices.

inspire from Weron, R., Misiorek, A., 2008. *Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models*. **International Journal of Forecasting** 24 (4), 744 – 763

  - 2 linear regressions
  - 2 GAMS
  - 2 random forests
  - GBM

$$\log(P_t) = \alpha_1 \log(P_{t-24}) + \alpha_2 \log(P_{t-48}) + \alpha_3 \log(P_{t-168}) + \alpha_4 \log(P.\max_t) + \alpha_5 \text{FZL}_t^{(0.95)} + \alpha_6 \text{FTL}_t^{(0.95)} + \alpha_7 \text{FZL}_t^{(0.8)} + \alpha_8 \text{FTL}_t^{(0.8)} + h(\text{DayType}_t) + \varepsilon_t$$

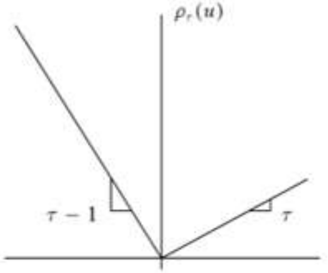
PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **(Convex) Aggregation with pin-ball loss:**

$$\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \qquad \widehat{p}_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{k,s})}}{\sum_{i=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{i,s})}}$$

$$l_t(x_{k,t}) = \rho_\tau(y_t - x_{k,t}) \longrightarrow$$

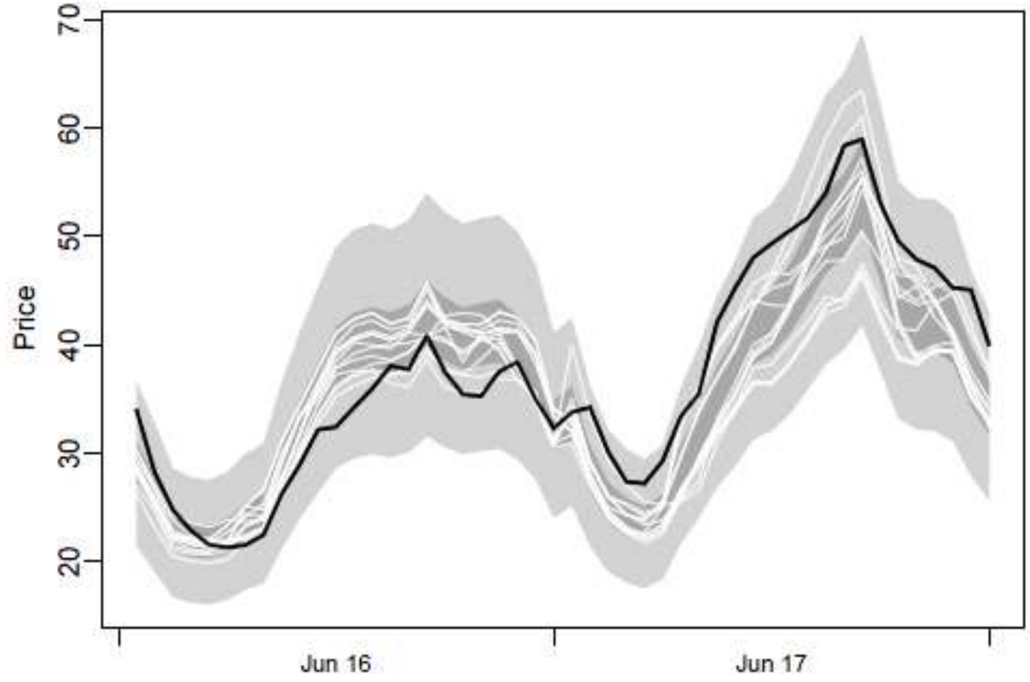


- **Extension to linear aggregation:**
  - substitute to original experts $\beta x_{1,t}, \ldots, \beta x_{K,t}, -\beta x_{1,t}, \ldots, -\beta x_{K,t}$

PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **Results**



PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **Results: 1st rank of the competition**

Task	TOLOLO	quantGAM	quantMixt	quantGLM
Jun. 06	XX	0.72	0.85	1.87
Jun. 17	1.06	1.15	1.37	0.71
Jun. 24	1.91	1.31	1.58	3.05
Jul. 04	1.71	2.06	1.27	1.59
Jul. 09	1.45	0.99	3.31	1.57
Jul. 13	1.10	2.23	1.20	1.18
Jul. 16	2.01	2.63	2.28	5.02
Jul. 18	9.15	5.13	7.90	11.72
Jul. 19	4.68	4.80	6.45	13.27
Jul. 20	1.59	1.90	2.35	2.80
Jul. 24	0.75	0.75	1.78	1.42
Jul. 25	2.46	2.30	0.84	2.12
Dec. 06	2.96	0.82	1.03	0.86
Dec. 07	1.35	3.63	3.23	3.22
Dec. 17	3.56	3.83	4.26	2.87

	Load		Price
Ranking	Team	Rating	Team
1	**Tololo**	**50,0%**	**Tololo**
2	**Adada**	**49,0%**	Team Poland
3	Jingrui (Rain) Xie	48,0%	GMD
4	OxMath	47,6%	C3 Green Team
5	E.S. Mangalova	45,4%	pat1

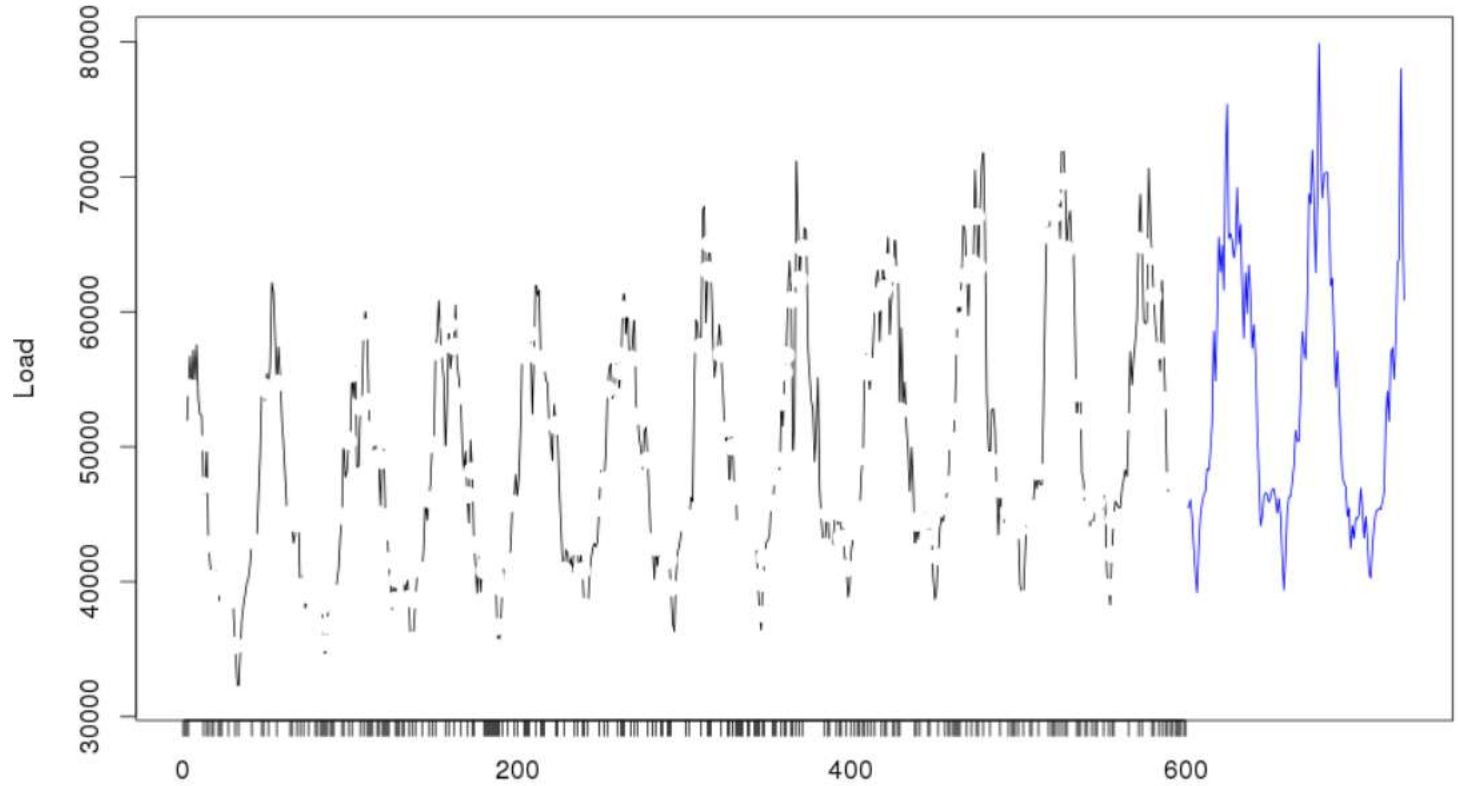
RANDOM MODEL GENERATION (WORK IN PROGRESS)

- **Random GAM:**

$$y_i = X_i\beta + \sum_{q=1}^{k_1} a_{1,q}(x)\beta_{1,q} + \sum_{q=1}^{k_2} a_{2,q}(x)\beta_{2,q} + \ldots + \varepsilon_i$$

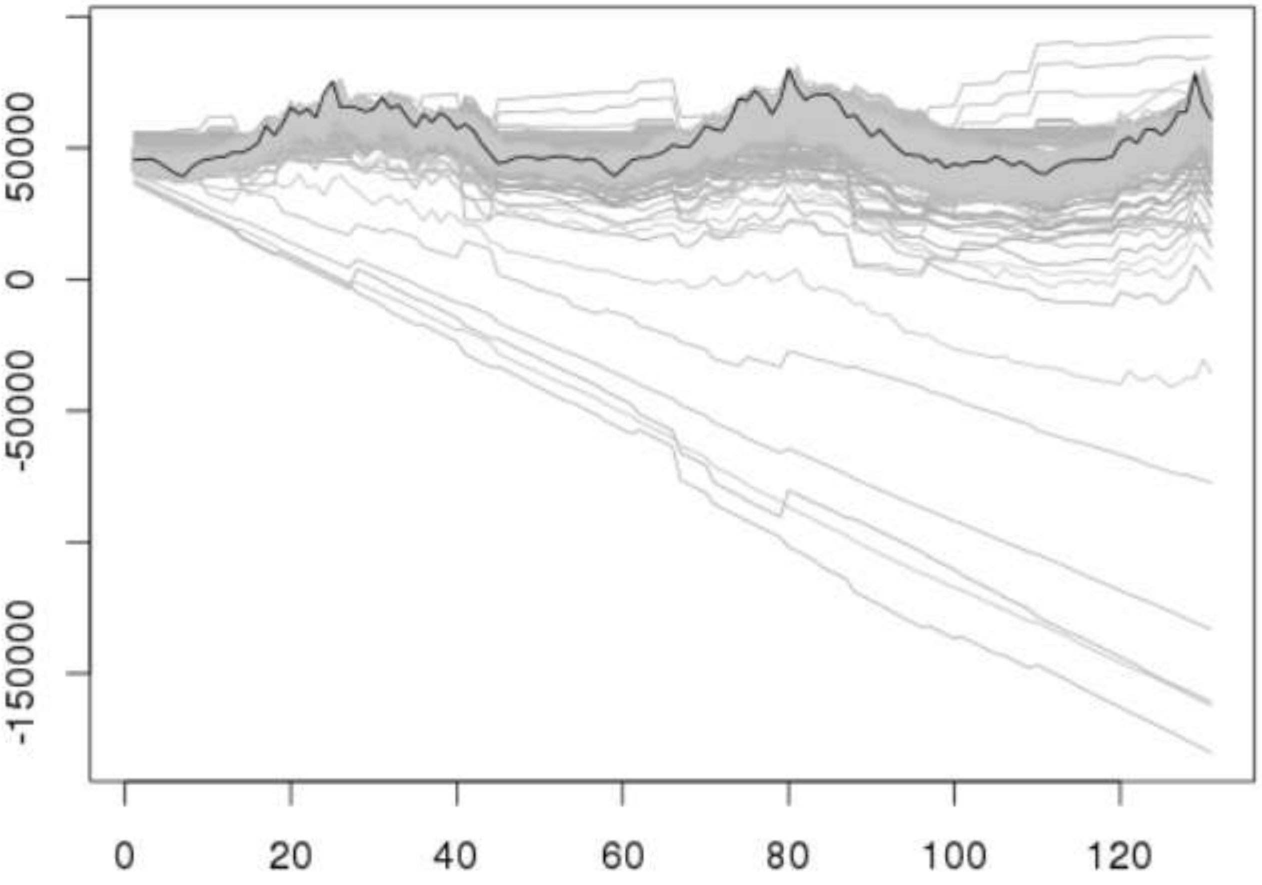
  - X are choosen at random among an initial subset of covariates
  - k are randomly sample in a realistic range (e.g. [3,50])
  - Subsampling the estimation set
    - weights can be initialized in an *out of bag* fashion (*OUT-OF-BAG ESTIMATION*, Leo Breiman https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf)
    - « scale » to potentially a high number of experts
  - Shrinkage algorithm
    - Calculate weights on the out of bag sample
    - Select experts adding a significant contribution (e.g.: significant average weights) to the aggregated forecast
    - Running the aggregation algorithm with these selected experts on the forecasting set

RANDOM MODEL ESTIMATION

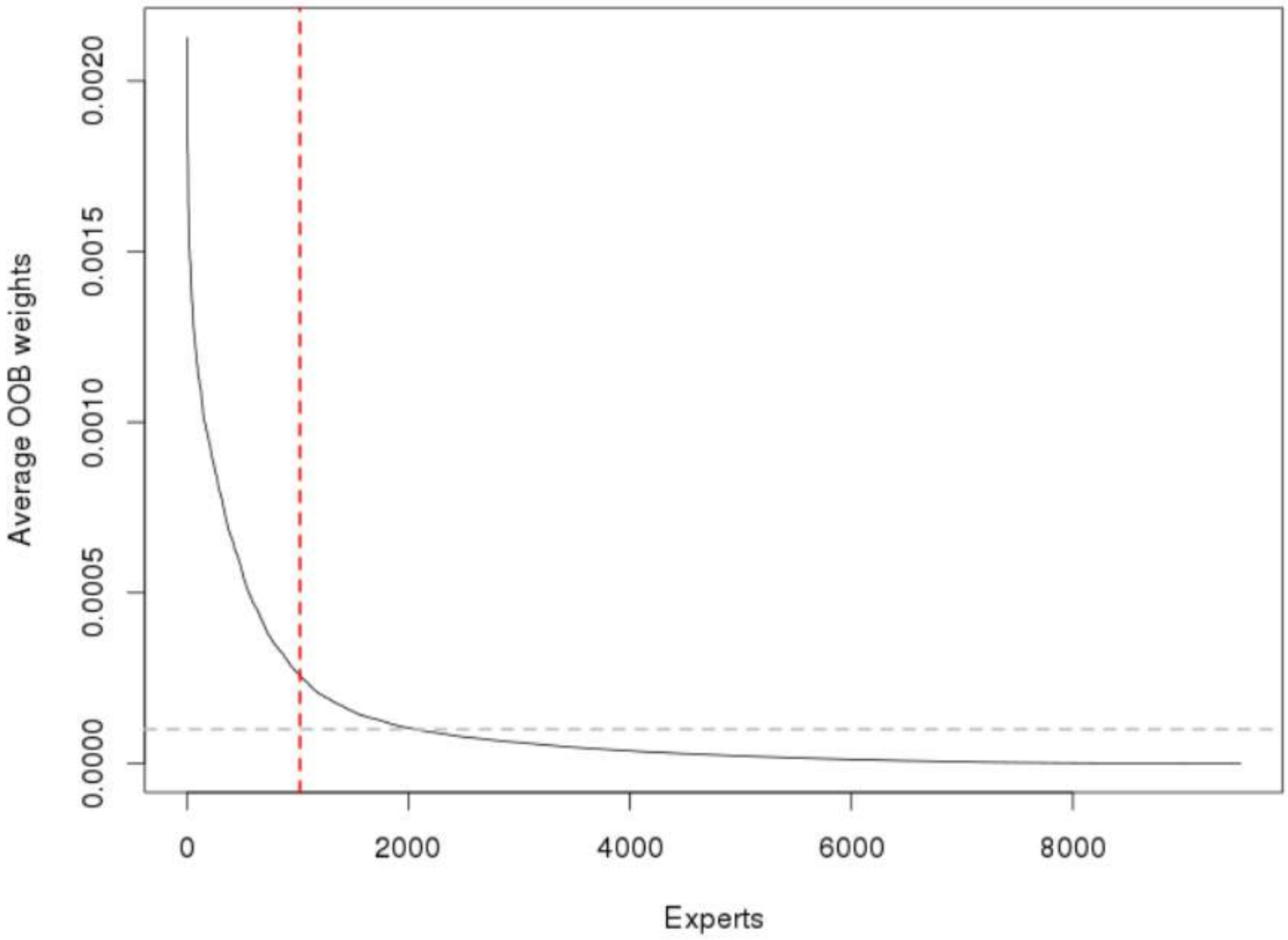



```
[1] "Load ~ Temp1 + s( NumWeek , k = 16 , bs = 'cr' )"

[2] "Load ~ Time + NumWeek + Temp + IPI_CVS"

[3] "Load ~ NumWeek + Temp + Temp1 + s( Temp , k = 12 , bs = 'cr' )"

[4] "Load ~ s( NumWeek , k = 16 , bs = 'cr' ) + s( Temp , k = 13 , bs = 'cr' ) + s( Temp1 , k = 17 , bs = 'cr' ) + s( IPI_CVS , k = 5 , bs = 'cr' )"

...

[9500] "Load ~ Time + NumWeek + Temp + Temp1 + IPI_CVS + s( NumWeek , k = 3 , bs = 'cr' ) + s( Load1 , k = 15 , bs = 'cr' )"
```

RANDOM MODEL GENERATION
10000 EXPERTS

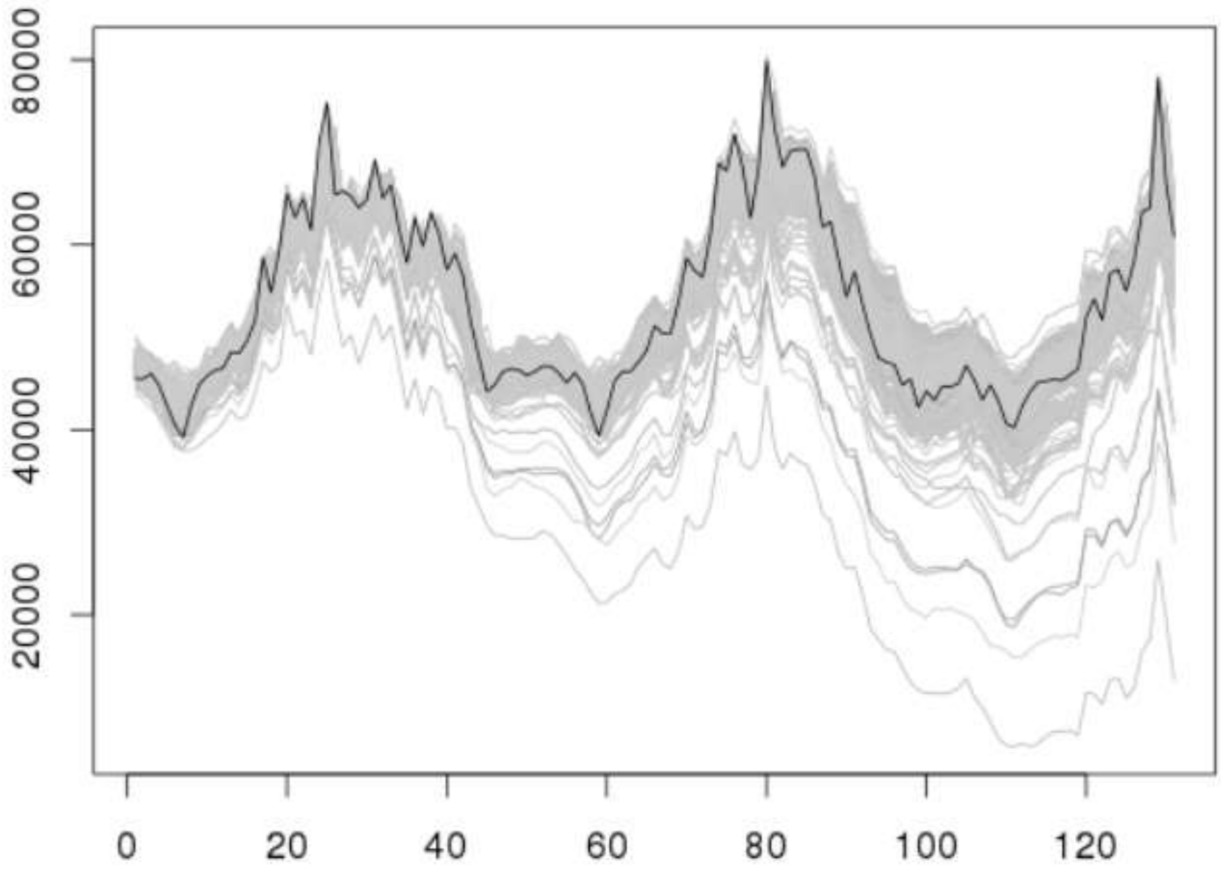


RANDOM MODEL
SELECTION



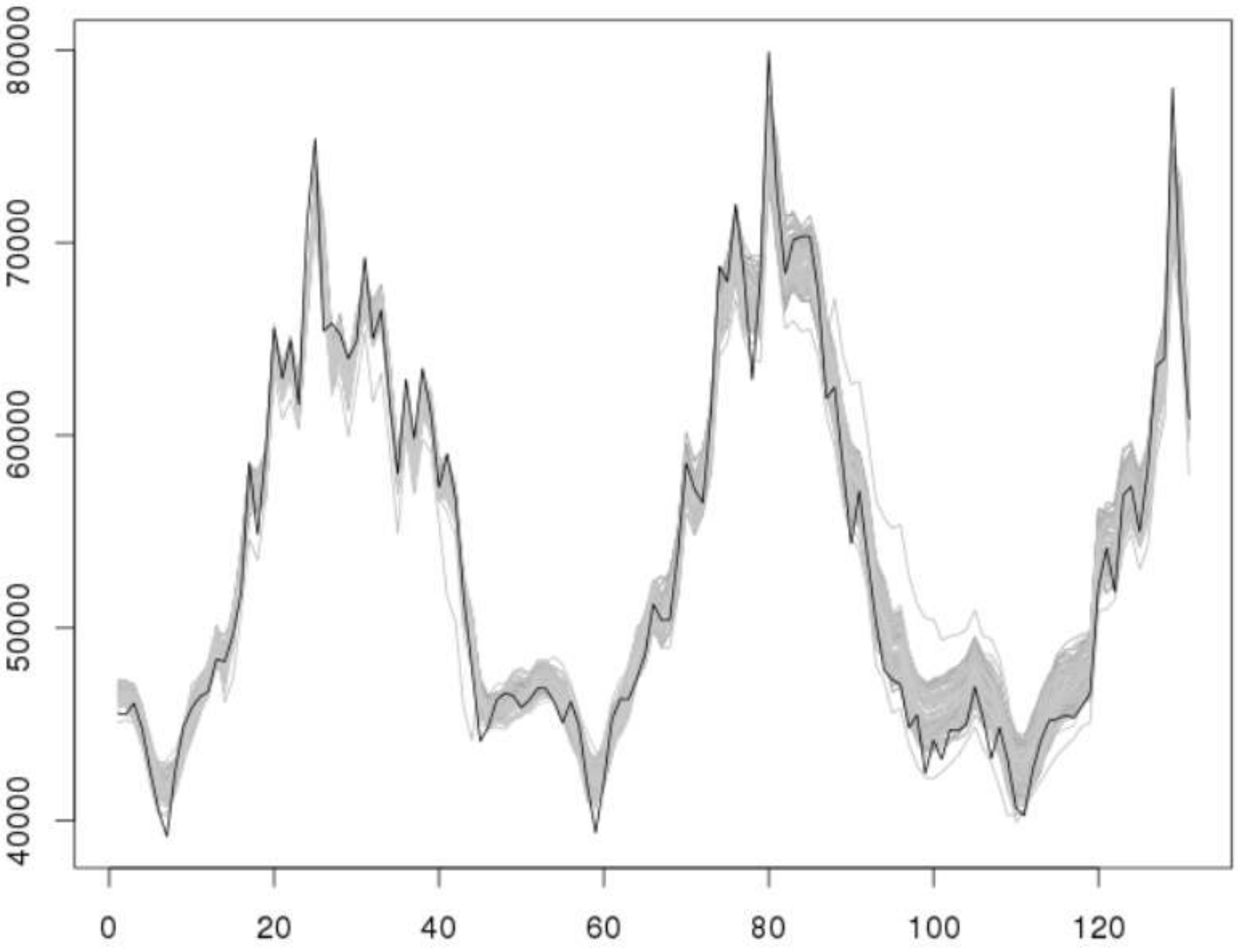
RANDOM MODEL GENERATION

1000 « BEST »

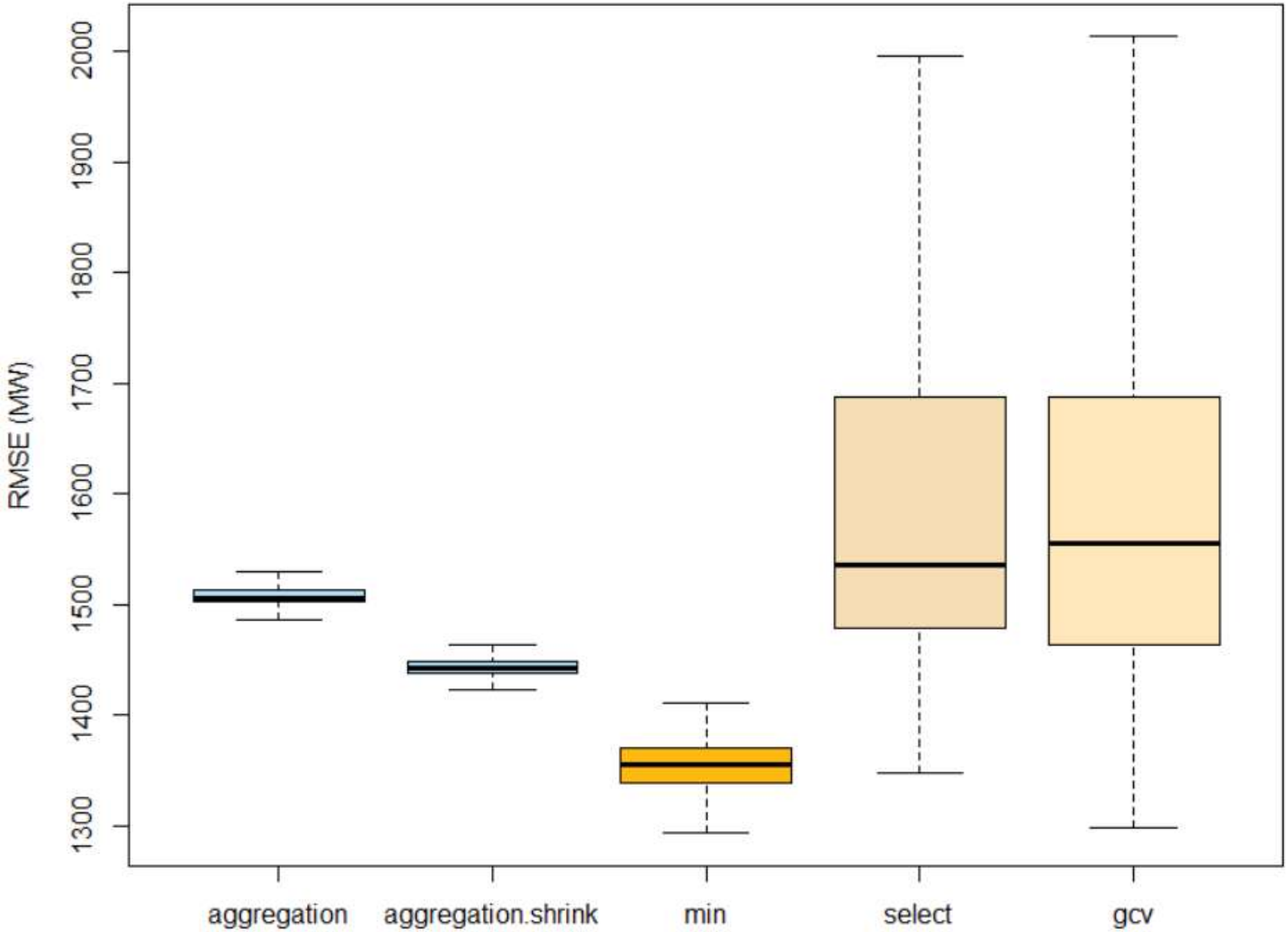


RANDOM MODEL GENERATION

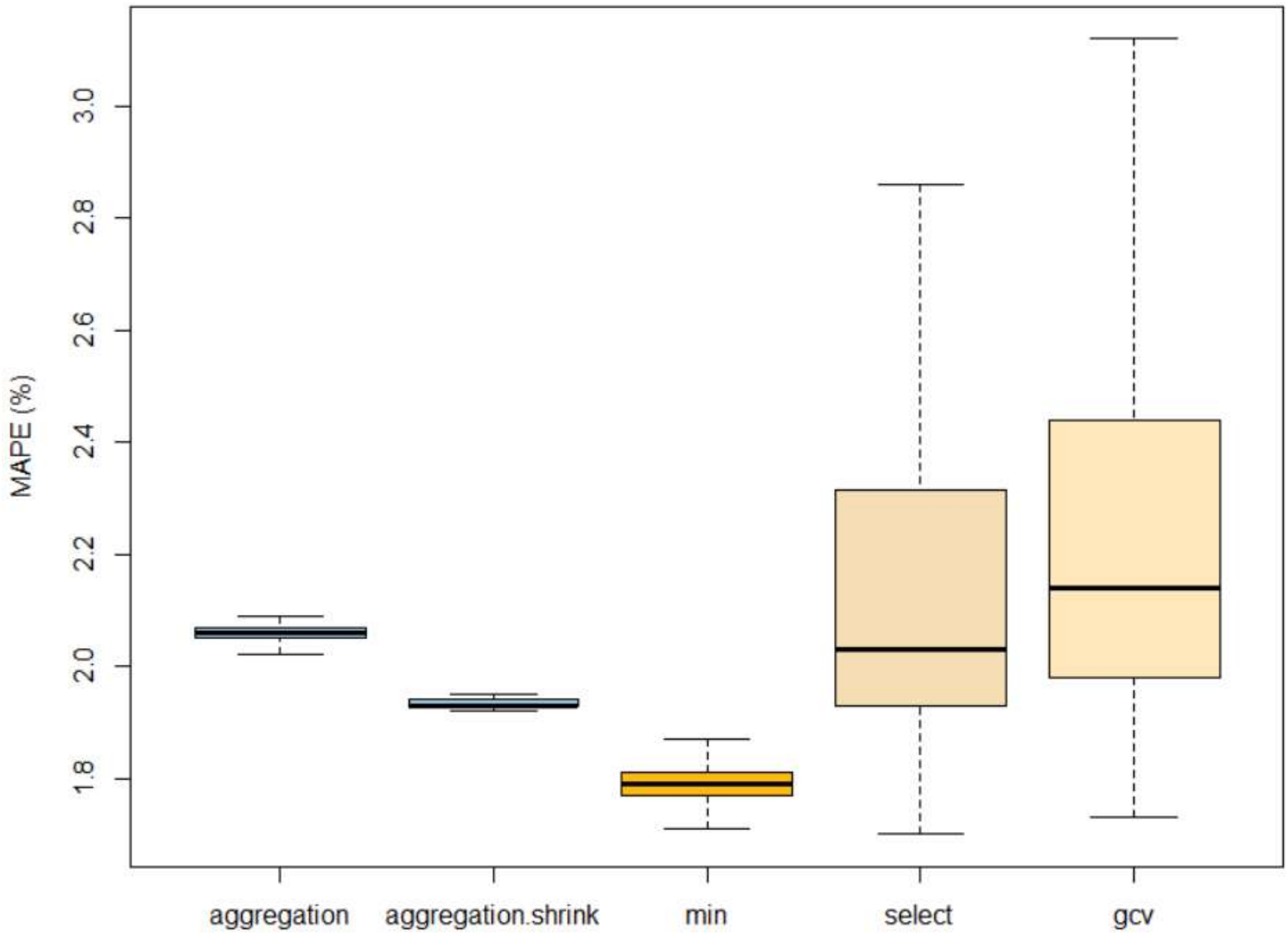
100 « BEST »



AGGREGATION OF 10 000 RANDOM EXPERTS



AGGREGATION OF 10 000 RANDOM EXPERTS



CONCLUSION/PERSPECTIVES

- **Forecasting methods:**
  - Industrial implementation on the way (national, substations, cogeneration central in poland: 30% better with GAM than with previous solution)
  - CLR: improve automatic clusturing, forecasting the clusters (HMM), derive probablistic forecasts

- **Aggregation of experts:**
  - R package OPERA (**O**nline **P**rediction through **E**xpe**R**ts **A**ggregation) – maintainer Pierre Gaillard

```
mixture(y, experts, aggregationRule = "MLpol", w0 = NULL, awake = NULL,
  href = 1, period = 1, delay = 0, y.ETR = NULL)
```

  - Automatic generation of experts
  - Combining a large number of experts: scale to very large number? (probably yes)

THANK YOU!