Ensemble Methods for Energy Forecasting

Yannig Goude, EDF R&D, LMO University of Paris-Sud Orsay

Journées MAS 2016

Phénomènes complexes et hétérogènes

Grenoble, du 29 au 31 août 2016. Campus universitaire de Saint-Martin-d'Hères

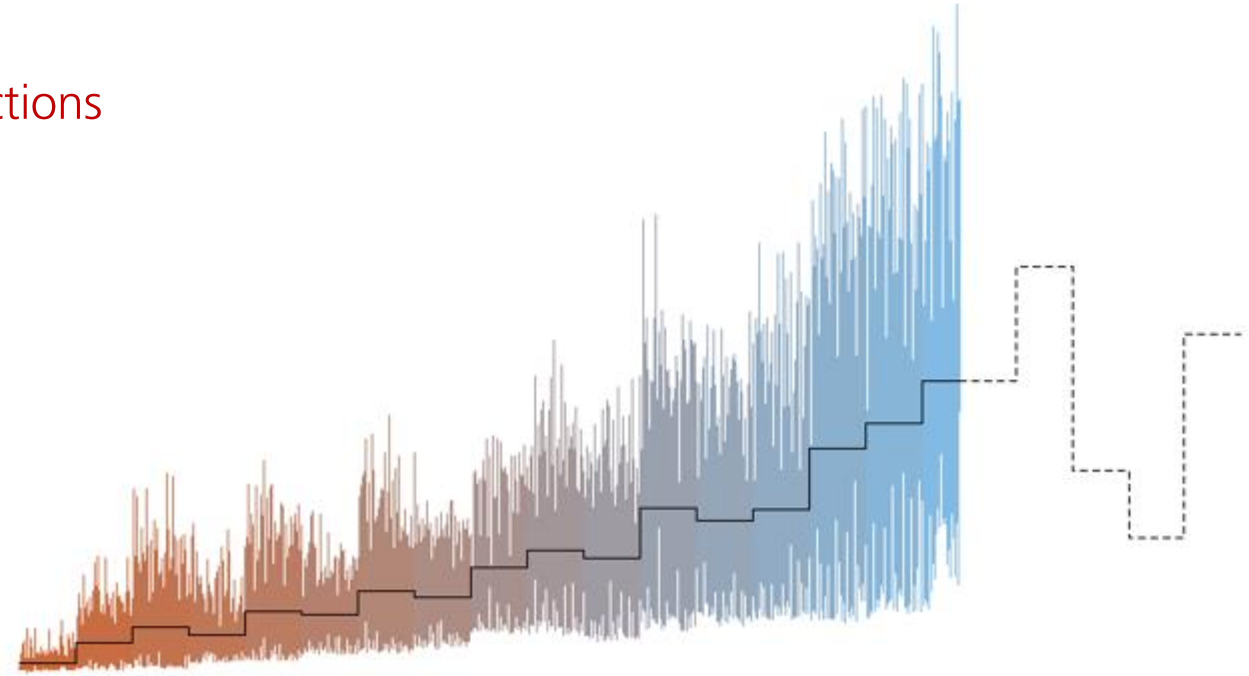
Online Forecasting

- We want to forecast a sequence of observations $y_1, y_2, \ldots, y_T$

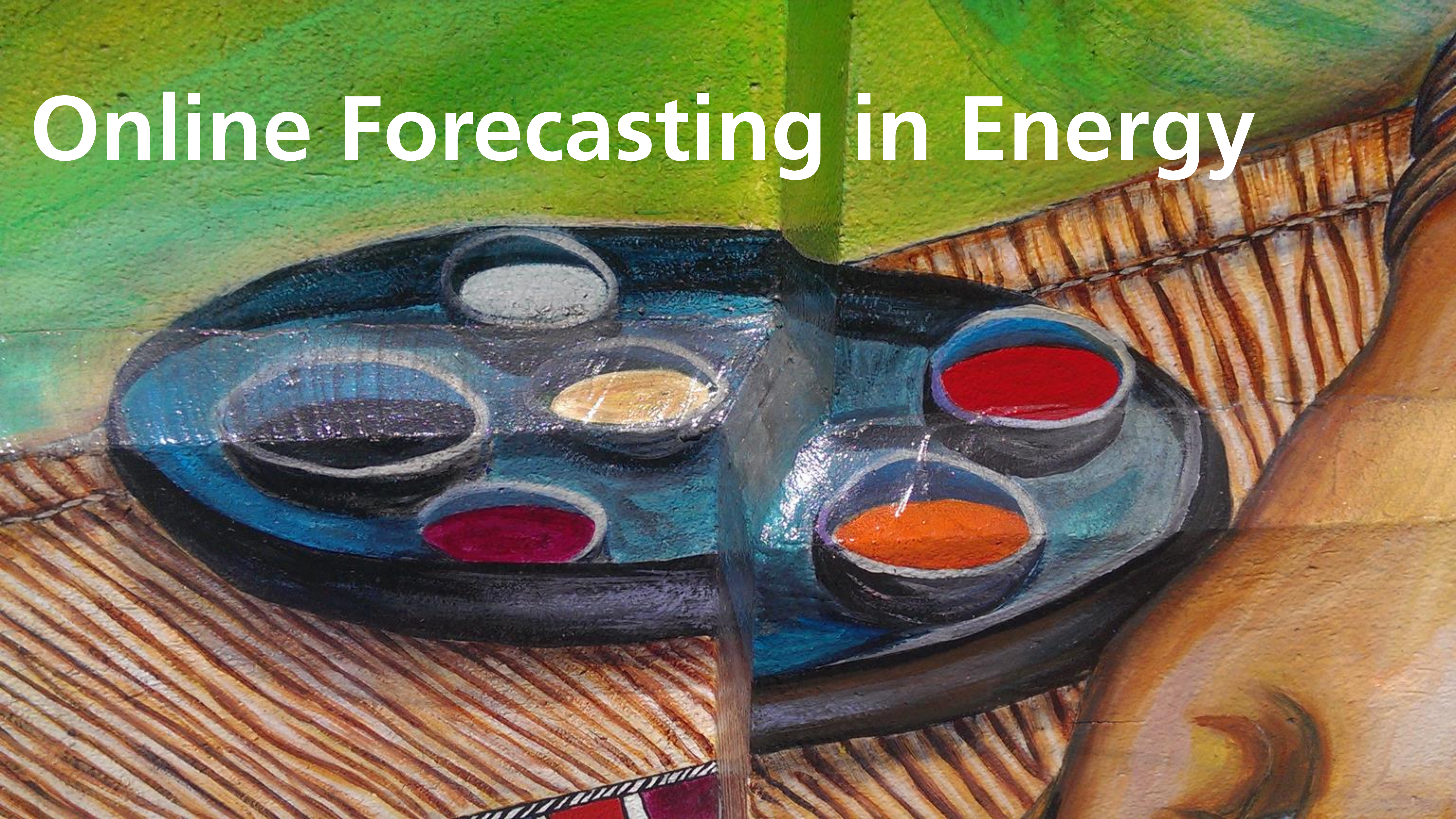
- Observations and predictions are made in a sequential fashion $y_1, y_2, \ldots, y_{t-1}$

  - predictions of y_t …

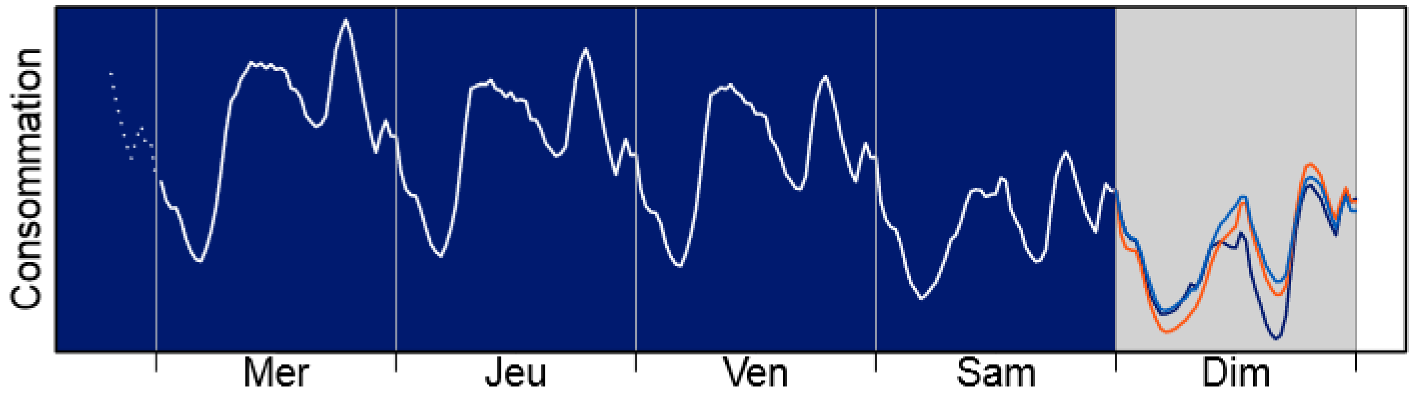
  … are based on past observations/predictions



Online Forecasting in Energy

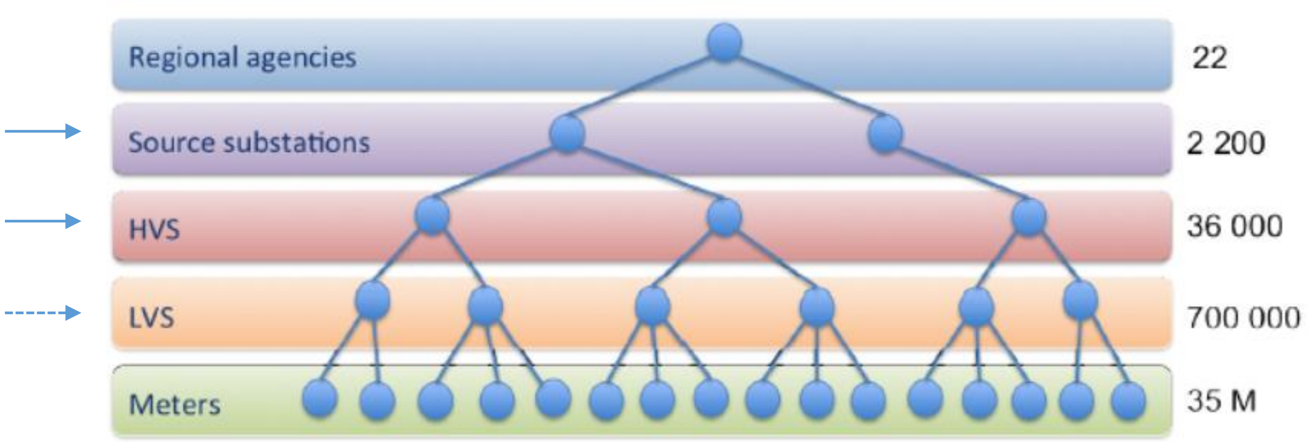


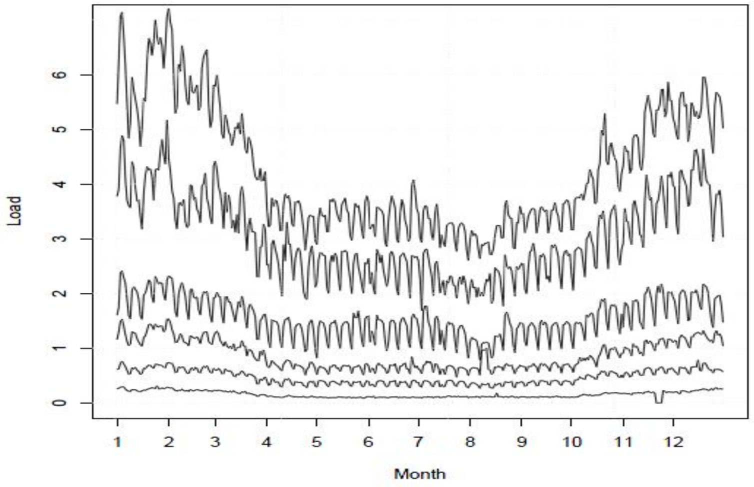
National electricity load forecasting



- A major concern for EDF
- Main entry for production planning

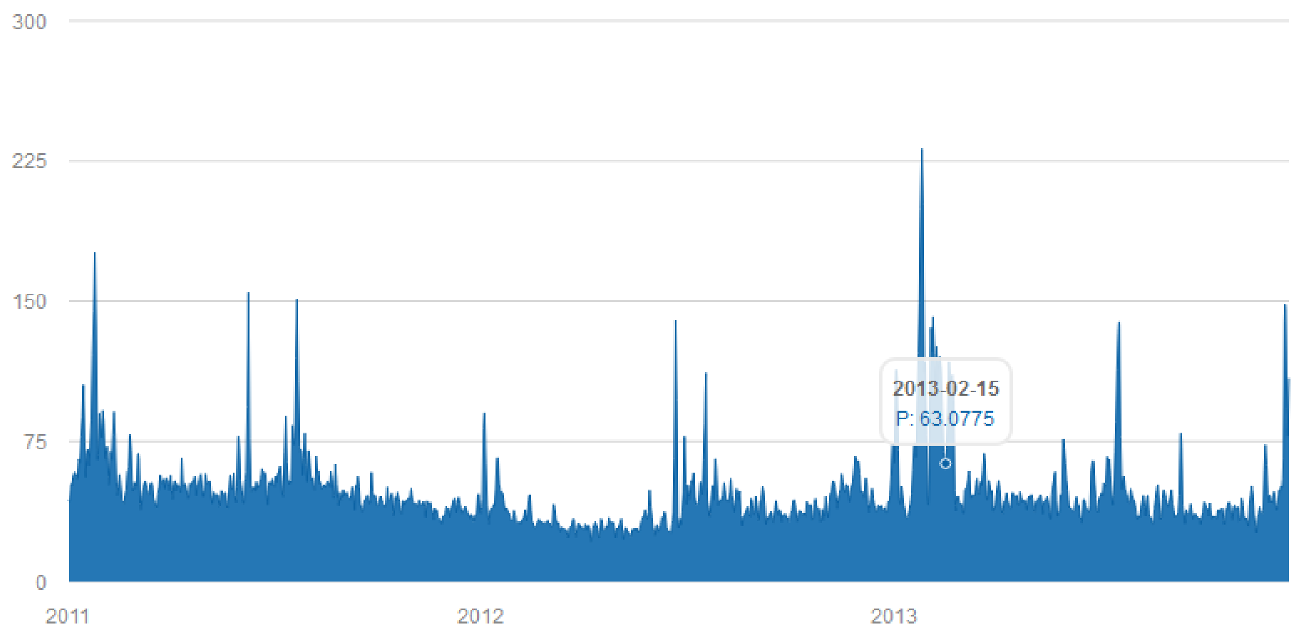
Local electricity load forecasting





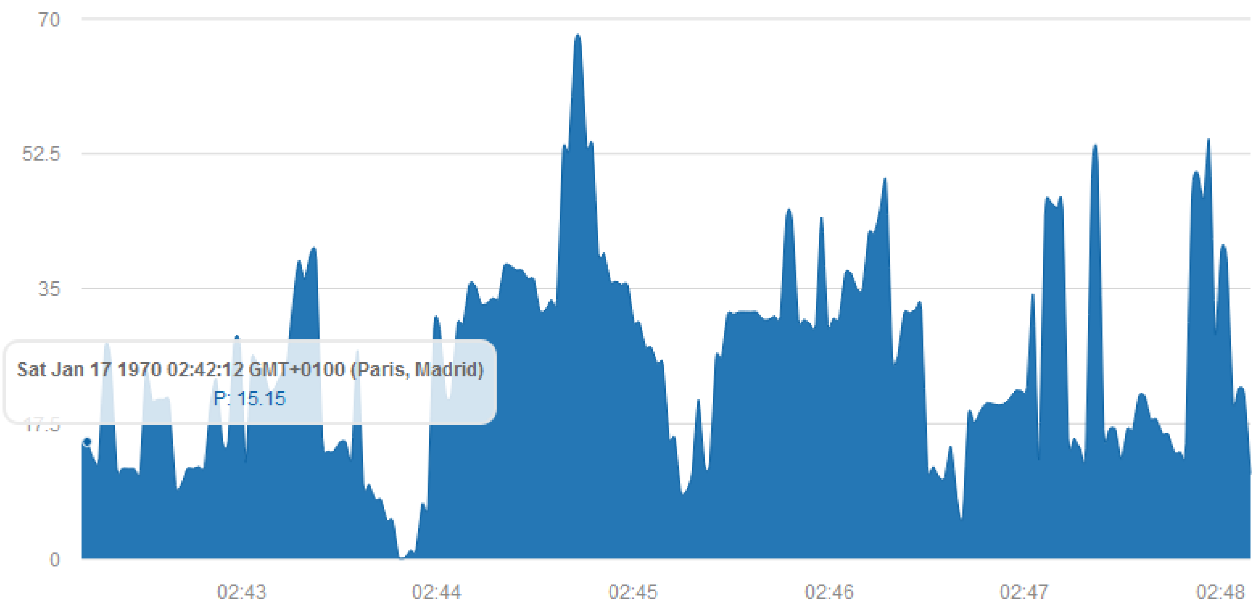
- Optimisation of the distribution grid
- Production planning at a local level

Price Forecasting



- trading
- Risk policy
- Churn forecast

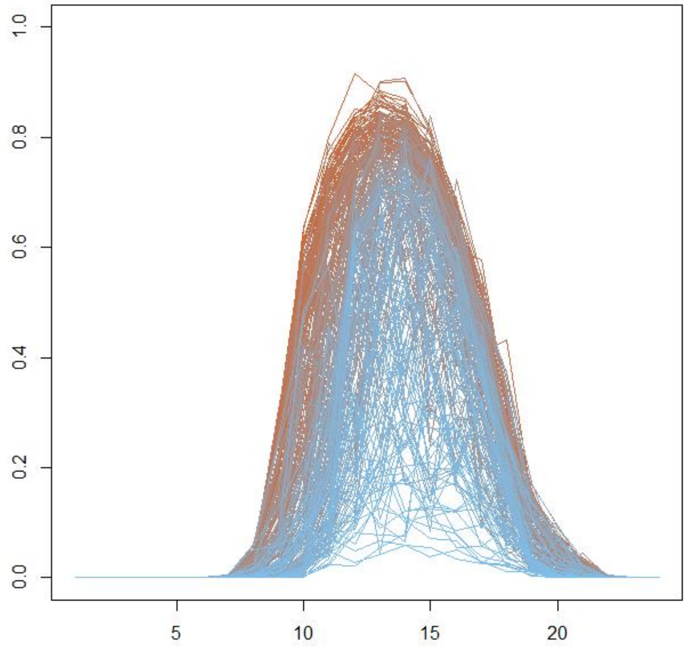
Real time bidding

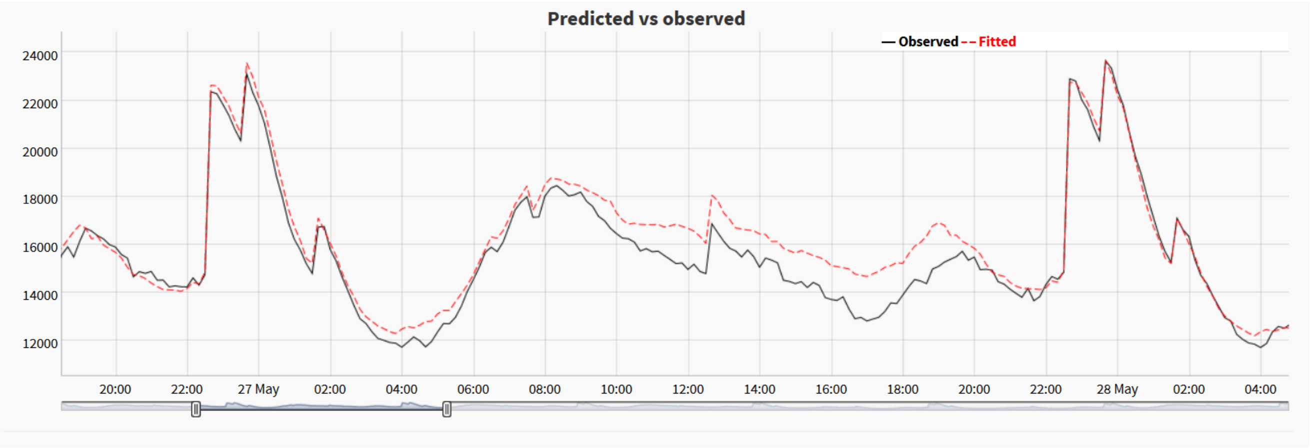


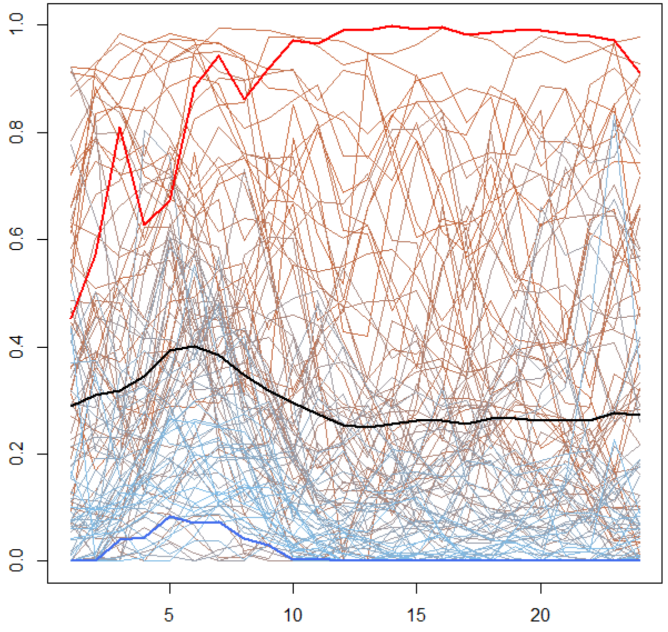
- trading
- Risk policy

Renewables

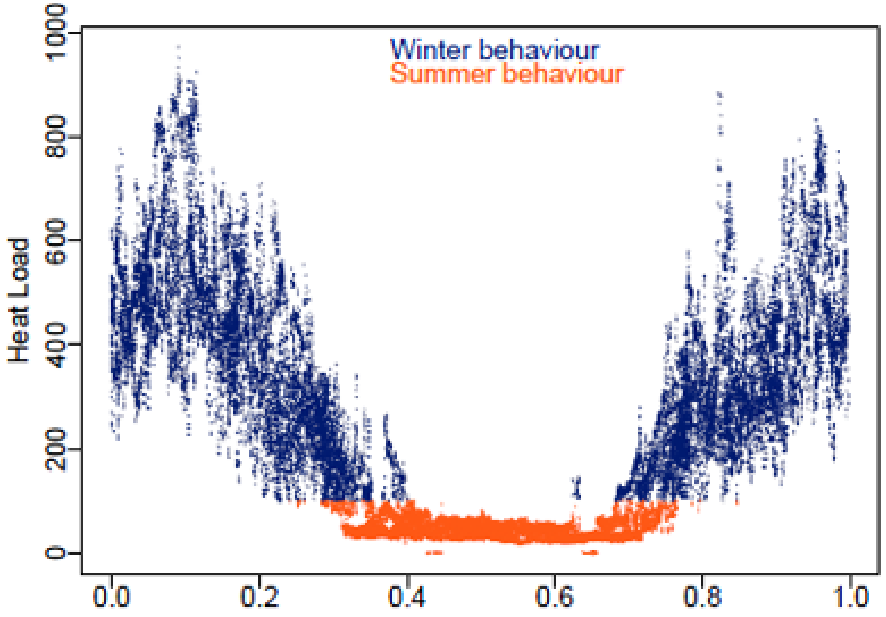
- *Random* production
- Net consumption modeling
- Local forecasts







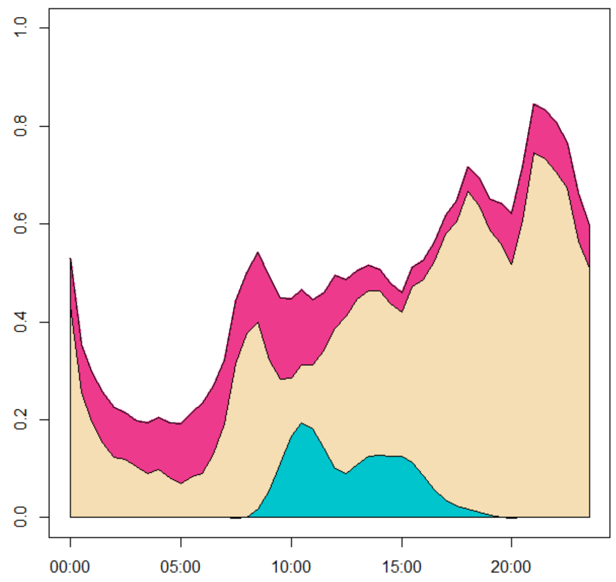
Heat demand forecasting

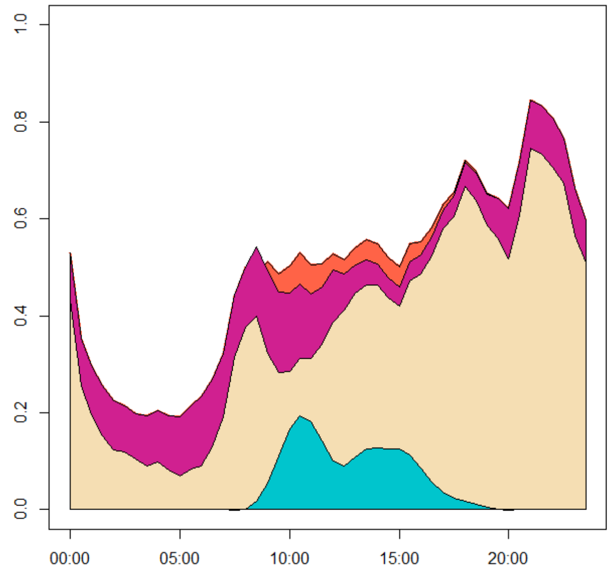


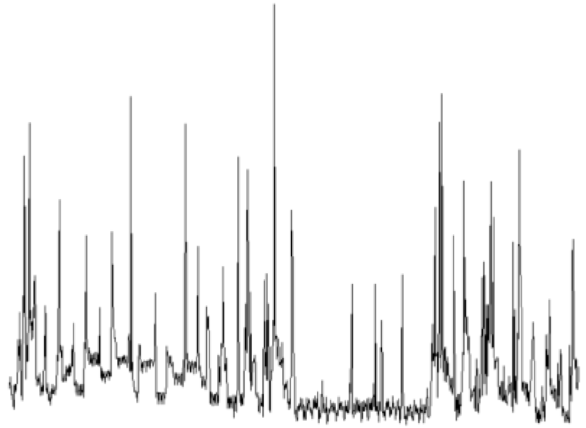
- Management of single production units

Other perspectives…

- Demand response
- Sensors data
- Smart meters
- Dynamic grids
- Local optimisation







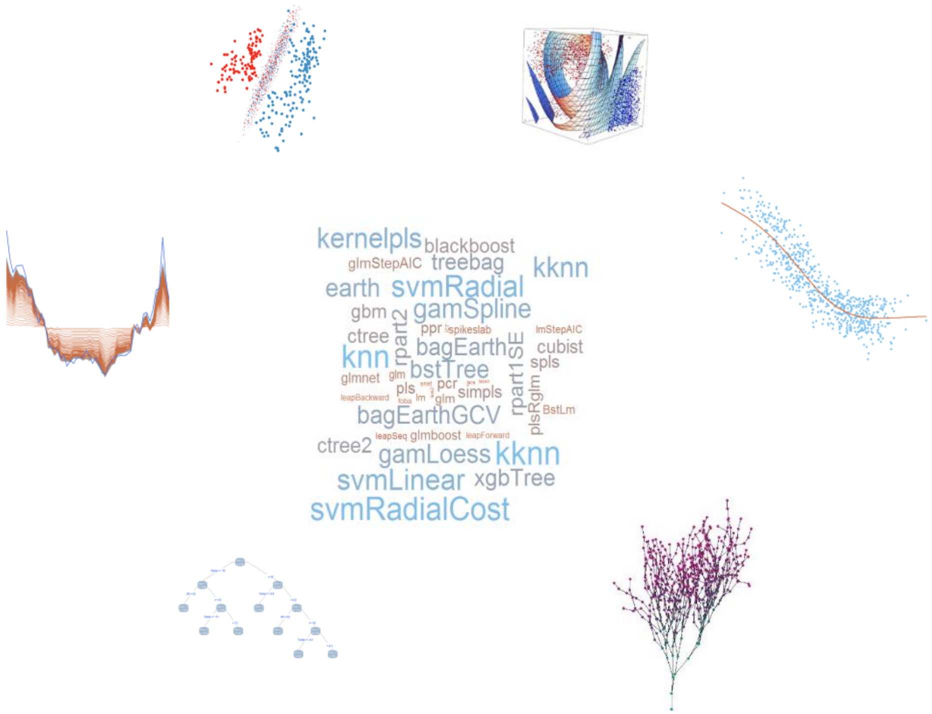
Ensemble Learning

Ensemble Learning

What is it?

In statistics and machine learning, **ensemble methods** use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms.[1][2][3] Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble refers only to a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives.

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$



Ensemble Learning
Does that work in practice?

During the nearly 3 years of the Netifx competition, there were two main factors which improved the overall accuracy: The quality of the individual algorithms and the ensemble idea.
BellKor's Pragmatic Chaos Team

1 million $

The XGBoost model got us to top 10. The meta-modelling then got us to the first position.
Marios Michailidis, Mathias Müller, HJ van Veen

10 000$

The idea was always to get models that are individually good on their own but have as little correlation as possible so that they can contribute meaningfully in the ensemble.
Andreas Merentitis, Alexander Bauer, Nurlanbek Duishoev

40 000$

In our final model, we had XGBoost as an ensemble model, which included 20 XGBoost models, 5 random forests, 6 randomized decision tree models, 3 regularized greedy forests, 3 logistic regression models, 5 ANN models, 3 elastic net models and 1 SVM model.
Darius, Davut, Song

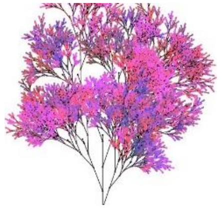
30 000$

I've tried ensembling XGBoost and RNN models but nothing could beat the simpler linear model.
Alejandro Mosquera

80 000$

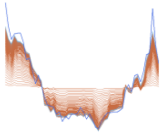
Ensemble Learning

Popular methods

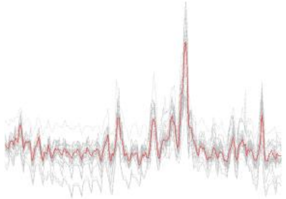


Random Forest
Breiman 2001

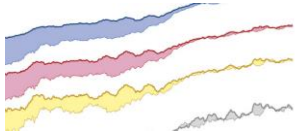
Bootstrap aggregating
Breiman 1994



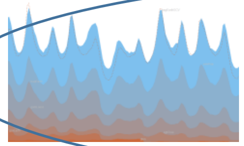
Gradient Boosting Machines
Friedman 2001



Stacking
Wolpert 1992



Bayesian Model Averaging
Roberts 1965; Leamer 1978



Robust aggregation of experts
Vovk 1990; Littlestone and Warmuth 1994

Ensemble Learning

Setting

dataset for batch learning

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}$$

N experts
(Machine learning regression methods including ensemble methods)

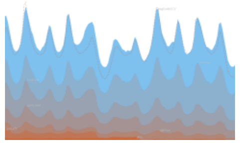
$$[\, f_1 \quad f_2 \quad \dots \quad f_N \,]$$

Online data

$$[\, y_{n+1} \quad x_{n+1,1} \quad x_{n+1,2} \quad \dots \quad x_{n+1,p} \,]$$

$$\vdots$$

$$[\, y_{n+m} \quad x_{n+m,1} \quad x_{n+m,2} \quad \dots \quad x_{n+m,p} \,]$$



Online Robust Expert Aggregation
Algorithms described and implemented in the OPERA R package

Ensemble Learning

Online robust aggregation algorithms

- At each instant t, expert j outputs a forecast $f_{j,t} = f_{j,t}(y_1^{t-1})$
- The aggregation algorithm determines his prediction based on
  - Past observations

    $$y_1^{t-1} = (y_1, \ldots, y_{t-1})$$

  - Current and past expert forecasts

    $$f_{j,s} \qquad s \in \{1, \ldots, t\} \qquad j \in \{1, \ldots, N\}$$

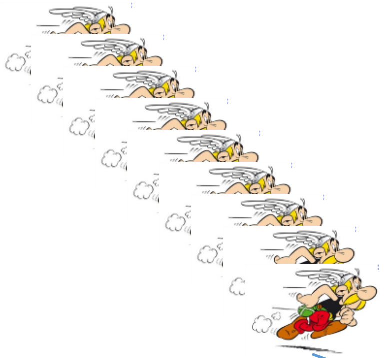
- At each instant, the aggregation algorithm computes a weight vector and forms

$$\widehat{y}_t = \sum_{j=1}^{N} p_{j,t} f_{j,t}$$

Online robust aggregation algorithms

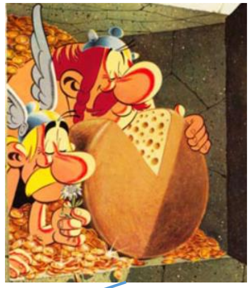
recipe

data stream

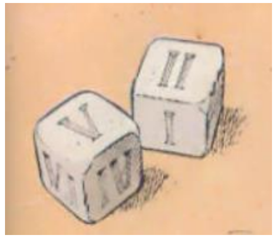


(sleeping) experts: human, machine learning methods…

loss function

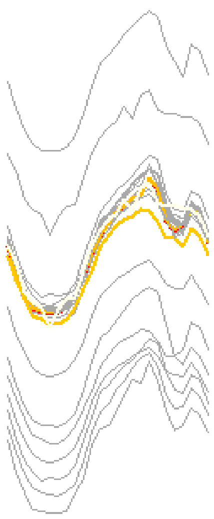


aggregation
algorithms

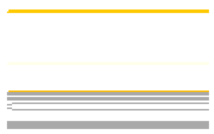


Online forecasting problem: example

Forecasted and Real Load Curves



Weights



Robust Sequential Expert Aggregation

Exponentially Weighted Aggregation

Loss of the expert j at time i

- Parameters $\eta > 0$ $p_0 = (\frac{1}{N}, \ldots, \frac{1}{N})$

- Weights update $$p_{j,t} = \frac{\exp(-\eta \sum_{i=1}^{t-1} l_{i,j})}{C}$$

- Oracle bounds $$\frac{1}{T} \sum_{t=1}^{T} \hat{l}_t - \min_k \sum_{t=1}^{T} \hat{l}_{t,k} \leq \sqrt{\frac{\log(N)}{T}}$$

Ensemble Learning

Setting: stacked experts

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ y_2 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_k & x_{k,1} & x_{k,2} & \dots & x_{k,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_j & x_{j,1} & x_{j,2} & \dots & x_{j,p} \\ y_n & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}$$

Random or block-random split

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_j & x_{j,1} & x_{j,2} & \dots & x_{j,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \xrightarrow{\text{Original experts calibration}} \begin{bmatrix} \widehat{f}_1 & \widehat{f}_2 & \dots & \widehat{f}_N \end{bmatrix}$$

$$\begin{bmatrix} y_2 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_k & x_{k,1} & x_{k,2} & \dots & x_{k,p} \end{bmatrix} \begin{bmatrix} \widehat{f}_1(x_2) & \widehat{f}_2(x_2) & \dots & \widehat{f}_N(x_2) & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{f}_1(x_k) & \widehat{f}_2(x_k) & \dots & \widehat{f}_N(x_k) & \end{bmatrix}$$

Augmented data set

stacking

Offline robust aggregation

$$\begin{bmatrix} \widehat{\widehat{f}}_1 & \widehat{\widehat{f}}_2 & \dots & \widehat{\widehat{f}}_N \end{bmatrix} \qquad \sum p_k \widehat{f}_k$$

Experts

50 experts from caret package

- Linear

lasso, lars2, lars, enet, foba, icr, leapBackward, leapForward, leapSeq, lm, lmStepAIC, spikeslab, glm, BstLm, glm, glmboost, glmnet, glmStepAIC

- Generalised Additive Models

bagEarth, bagEarthGCV, bstTree, earth, gamLoess, gamSpline, gcvEarth

- Projection based

pcr, ppr, pls, plsRglm, simpls

- Regression tree:

Gbm, blackboost, ctree, ctree2, rpart1SE, rpart2, treebag, xgbTree

- Kernel

Kernelpls, svmLinear, svmPoly, svmRadial, svmRadialSigma, svmRadialCost, knn, kknn

Calibrate default parameter of the caret::train function by cross validation 5-fold

The package OPERA

OPERA

Online Prediction by ExpeRt Aggregation

- Developed by Pierre Gaillard during his PhD at EDF R&D/Université Paris-Sud
- Online Robust Aggregation, joined work with:
  - Gilles Stoltz (CNRS-HEC Paris)
  - Marie Devaine (Ecole Normale Supérieure, Paris, France)
  - Yannig Goude (EDF R&D-Univ. Paris –Sud)

Pierre Gaillard

Gilles Stoltz

Marie Devaine

Yannig Goude

Opera

oracle calculation

```
oracle(Y, experts, model = "convex", loss.type = "square", awake = NULL,
  lambda = NULL, niter = NULL, ...)
```

- Y: the data stream to predict
- experts: the set of experts
- model: oracle, 'expert' best fixed (constant over time) expert oracle, 'convex' best fixed convex combination, 'linear' best fixed linear combination of expert, 'shifting' for all number m of switches the sequence of experts with at most m shifts that would have performed the best to predict the sequence of observations in Y.
- loss.type: the loss function, 'square', 'absolute', 'percentage', or 'pinball' (quantile reg.)
- awake: a matrix specifying the activation coefficients of the experts, lie in [0,1]
- lambda: a positive number used by the 'linear' oracle only. A possible L_2 regularization parameter for computing the linear oracle (if the design matrix is not identifiable)

See more @ ?oracle

Opera

the mixture fonction

```
mixture(Y = NULL, experts = NULL, model = "MLpol", loss.type = "square",
  loss.gradient = TRUE, coefficients = "Uniform", awake = NULL,
  parameters = list())
```

- Y: the data stream to predict
- experts: the set of experts
- model: aggregation algorithm, 'EWA' Exponential Weight Aggregation, 'FS' Fixed Share, 'Ridge' Ridge regression 'Mlpol', Polynomial Potential aggregation, 'OGD' Online Gradient Descent
- loss.type: the loss function, 'square', 'absolute', 'percentage', or 'pinball' (quantile reg.)
- loss.gradient: should it take the gradient of the loss or not
- coefficients: prior weights of the experts (not possible for 'MLpol')
- awake: a matrix specifying the activation coefficients of the experts, lie in [0,1]
- parameters: optional parameters for the aggregation rule

See more @ ?mixture

Opera

Online Prediction by ExpeRt Aggregation

Package + electricity load data to test is available @

https://cran.rstudio.com/web/packages/opera/index.html

opera: Online Prediction by Expert Aggregation

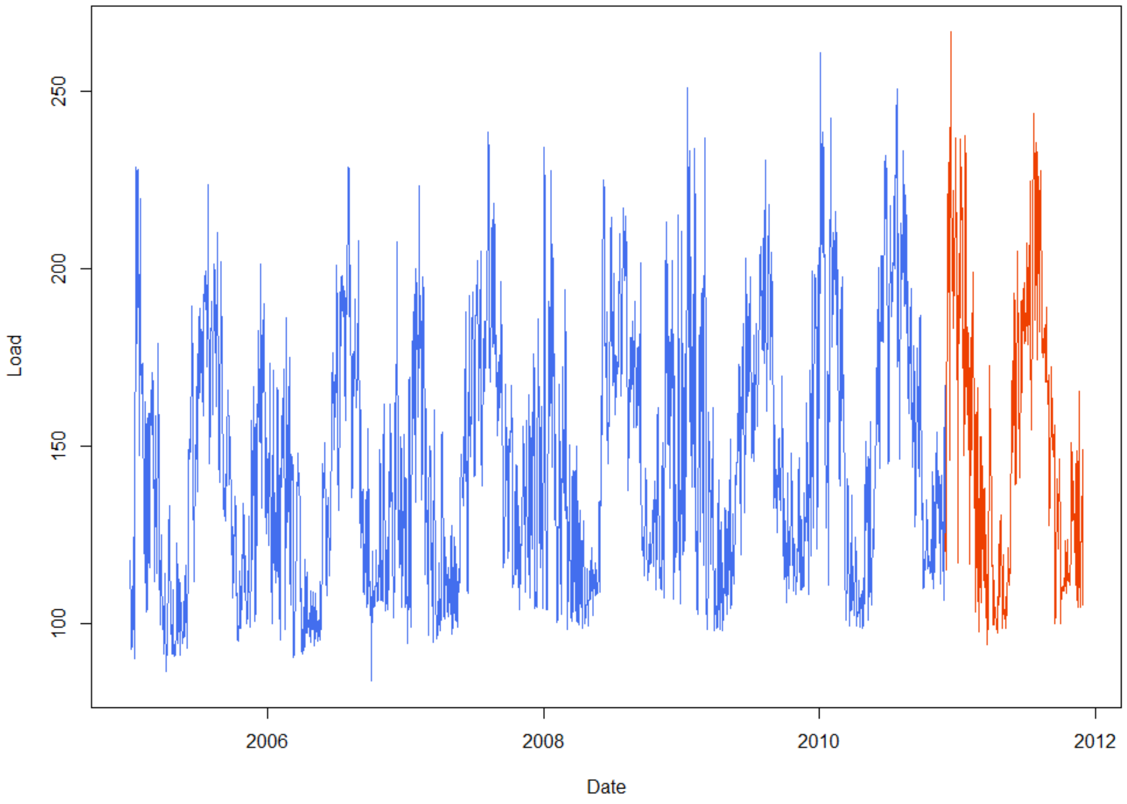
Misc methods to form online predictions, for regression-oriented time-series, by combining a finite set of forecasts provided by the user.

Version:	1.0
Depends:	R (≥ 3.1.0)
Imports:	quadprog, quantreg, RColorBrewer
Suggests:	testthat, splines, caret, mgcv, survival, knitr, gbm
Published:	2016-08-17
Author:	Pierre Gaillard [cre, aut], Yannig Goude [aut]
Maintainer:	Pierre Gaillard <pierre at gaillard.me>
BugReports:	https://github.com/dralliag/opera/issues
License:	LGPL-2 | LGPL-2.1 | LGPL-3 [expanded from: LGPL]
Copyright:	EDF R&D 2012-2015
URL:	http://pierre.gaillard.me/opera.html

Applications

GEFCOM14 Electricity Load Data set

- Y: electricity demand on the US grid, daily mean, 2005-01-01 / 2011-11-30
- X: lag load, temperature 25 meteo stations, calendar information



GEFCOM14 Electricity Load Data set

Results

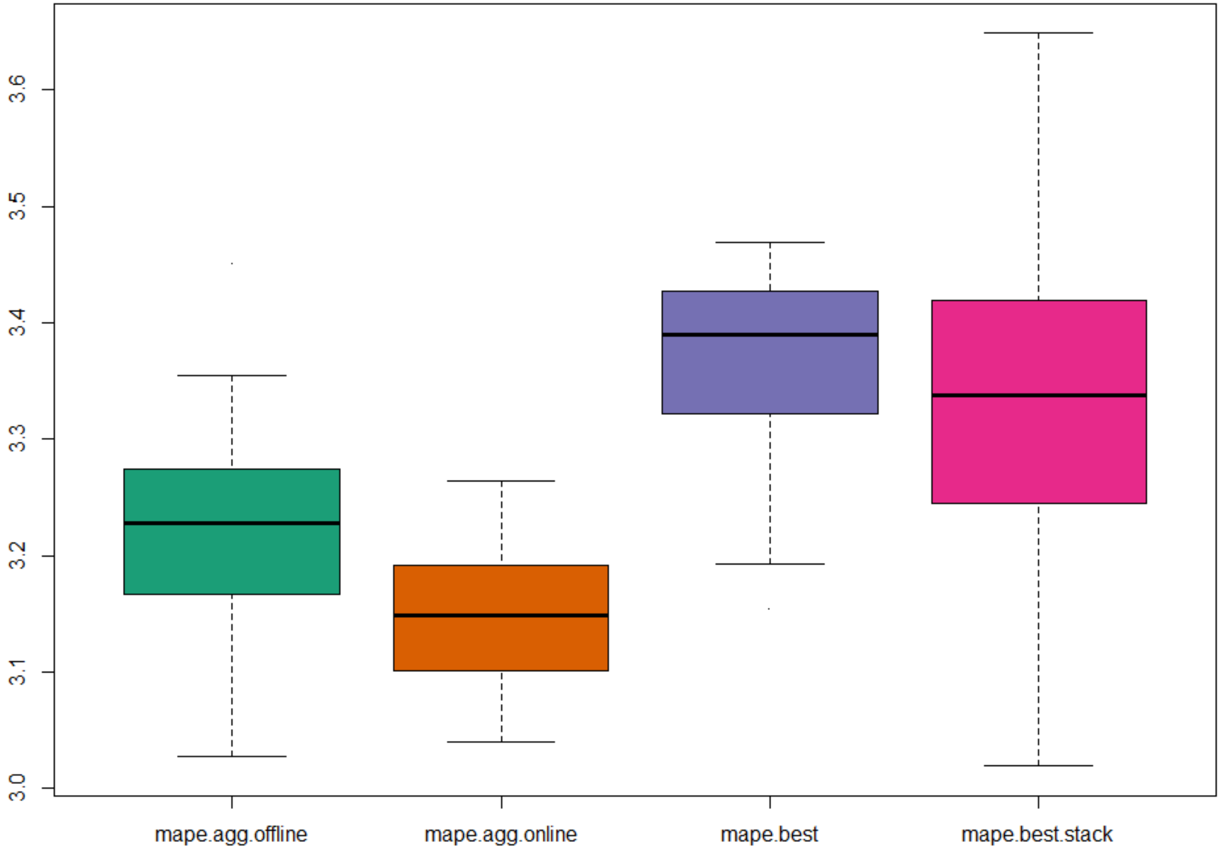
4.7% improvement over the best expert for the off-line MLpol

3.2% improvement over the best stacked expert for the off-line MLpol

6.9% improvement over the best expert for the on-line MLpol

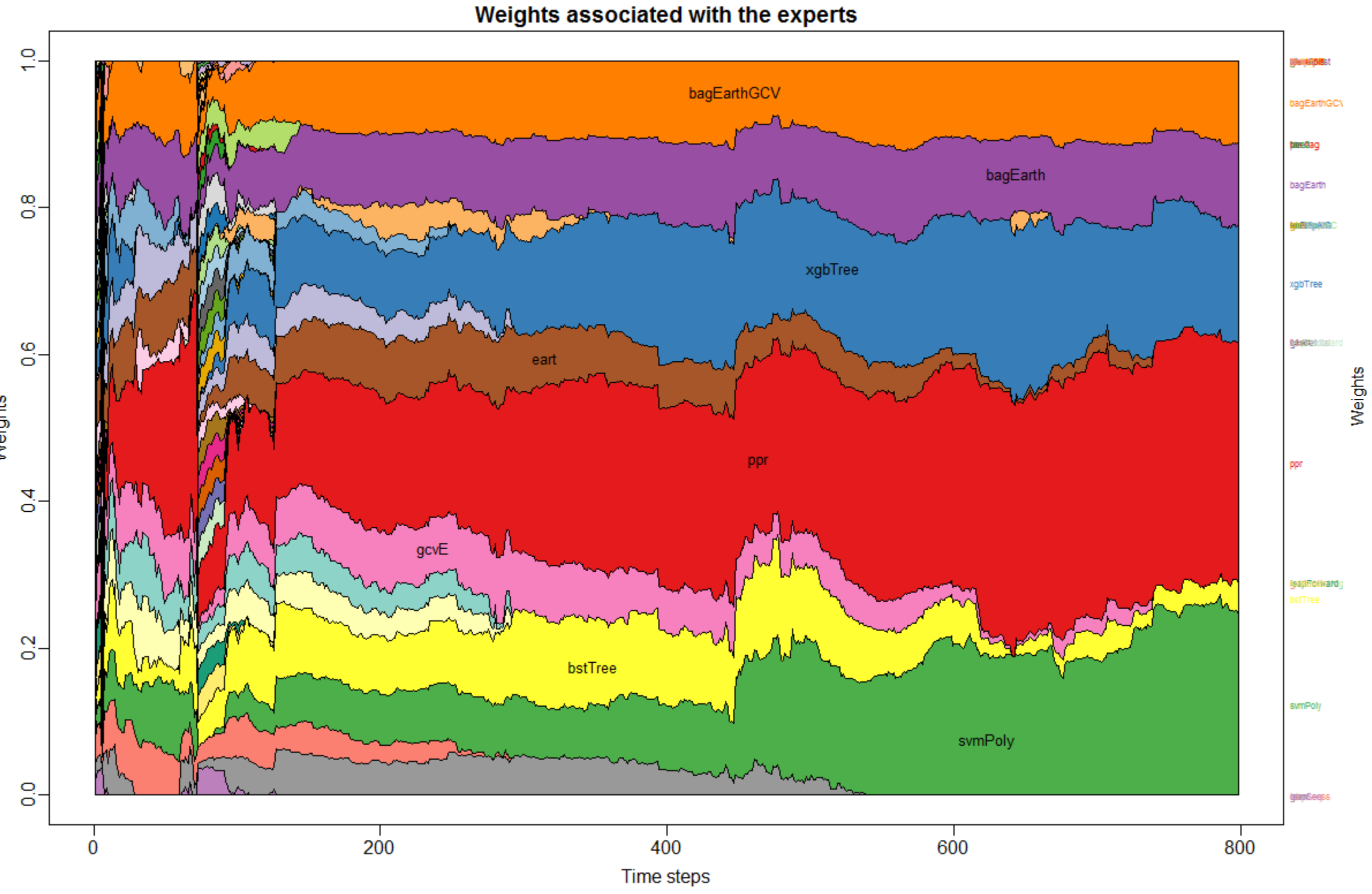
5.2% improvement over the best stacked expert for the on-line MLpol

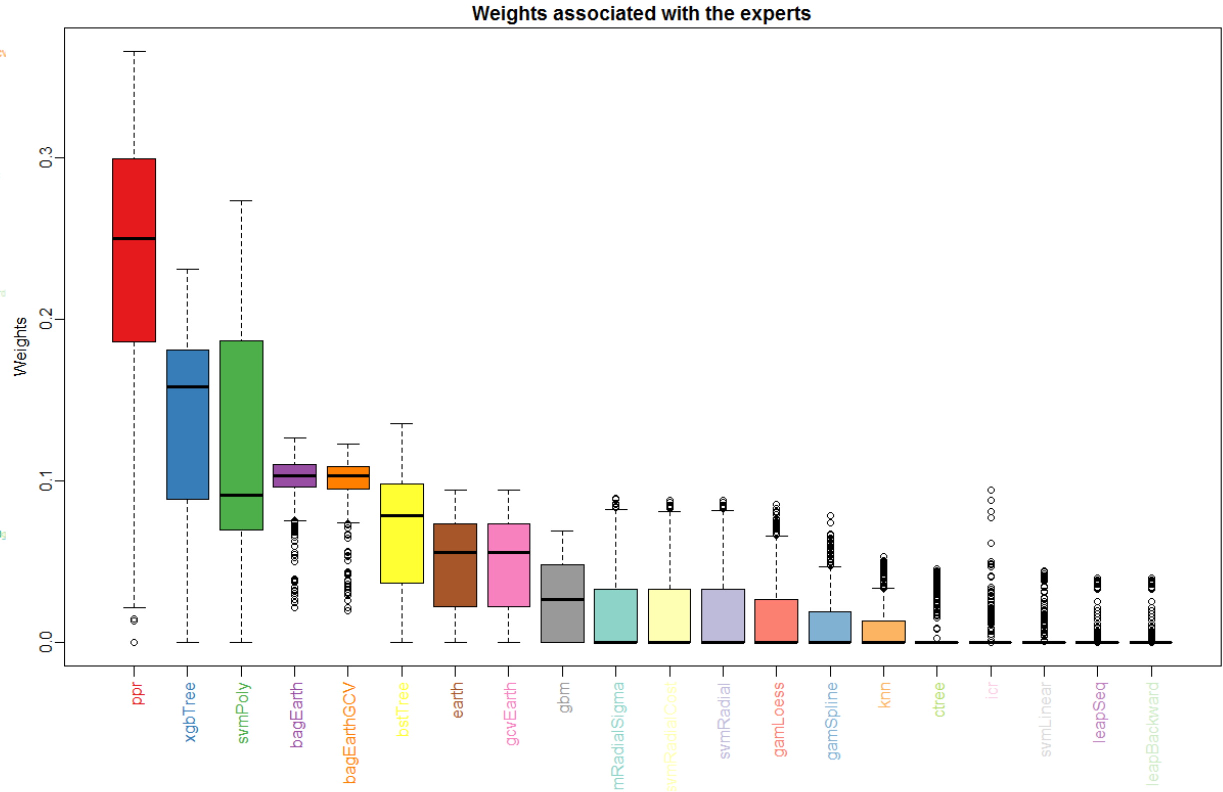
Mean Absolute Percentage Error
Obtained for100 random splits (stacking)



GEFCOM14 Electricity Load Data set

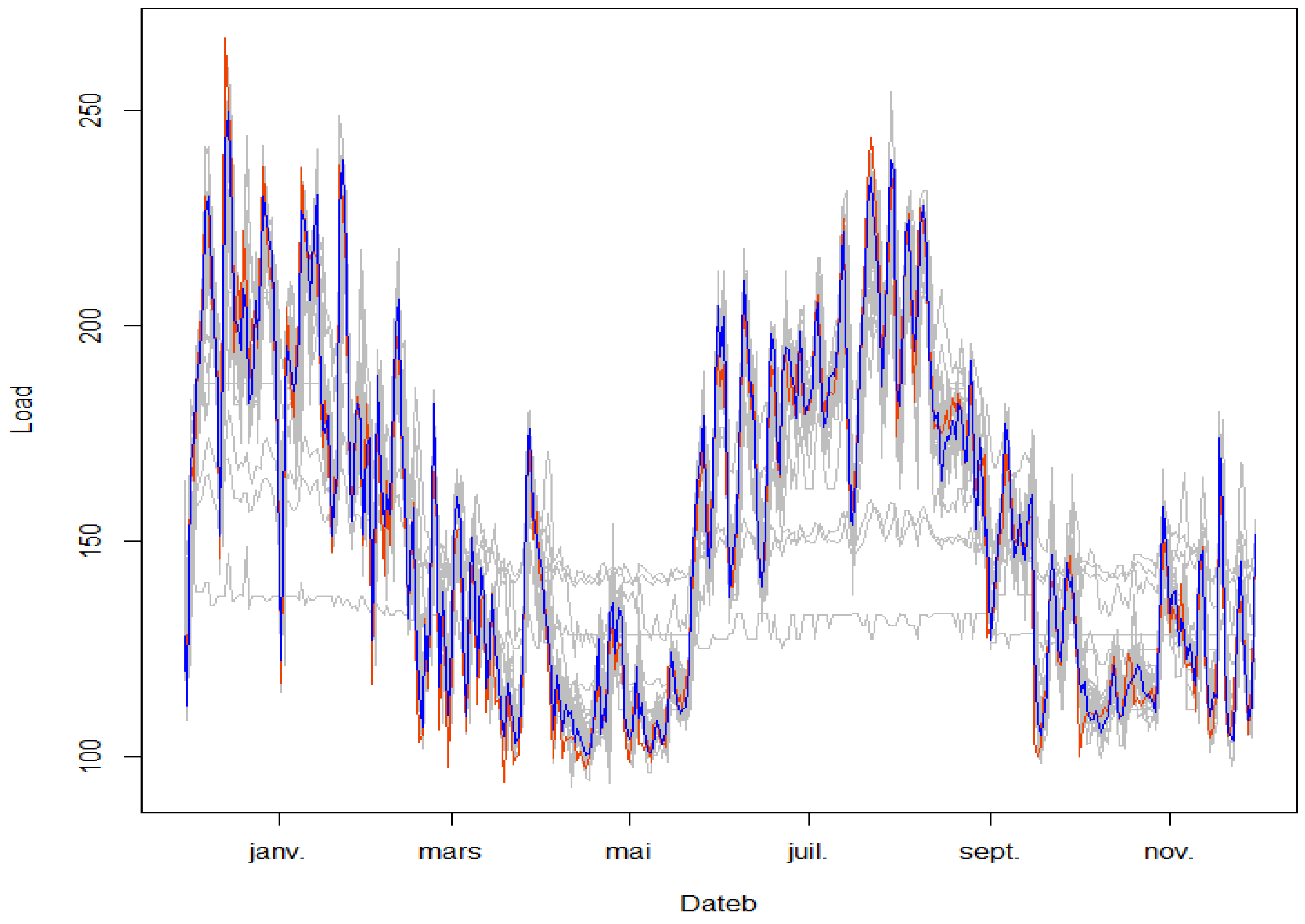
Results





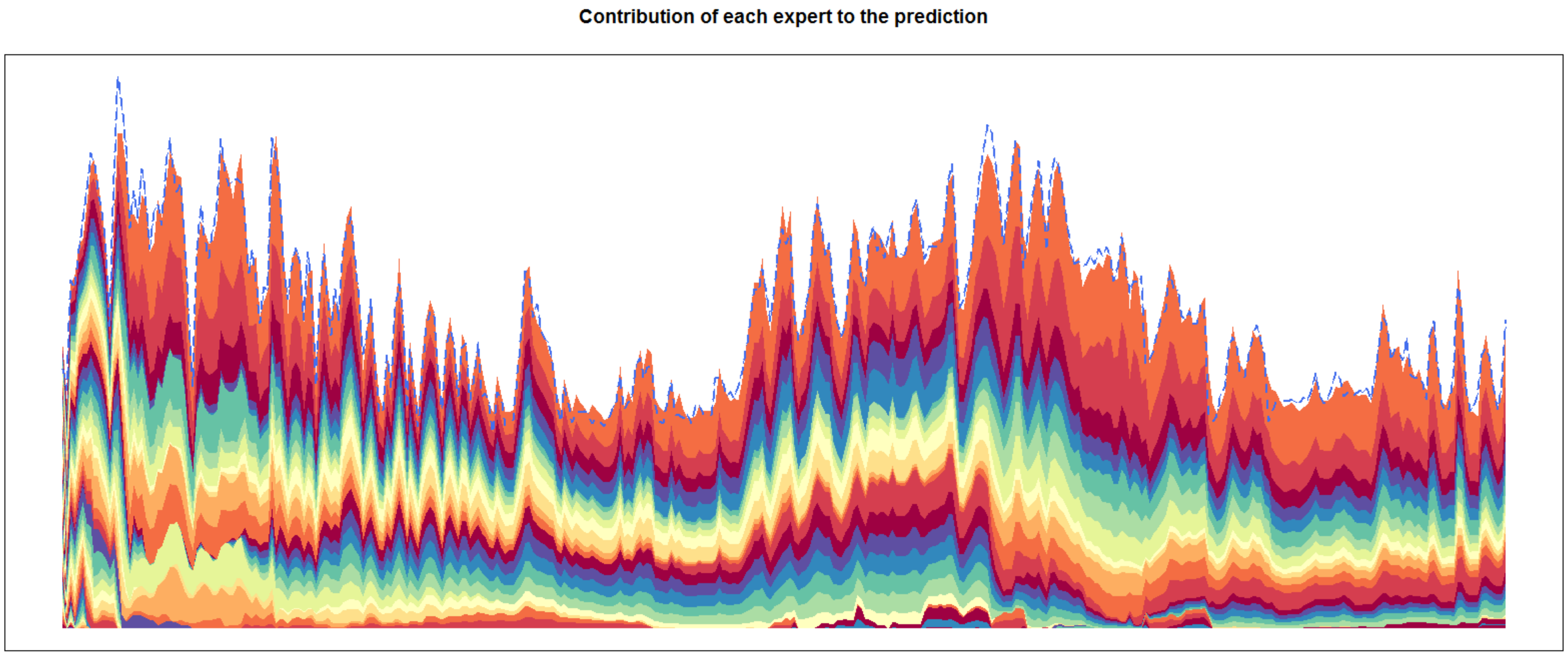
GEFCOM14 Electricity Load Data set

Results



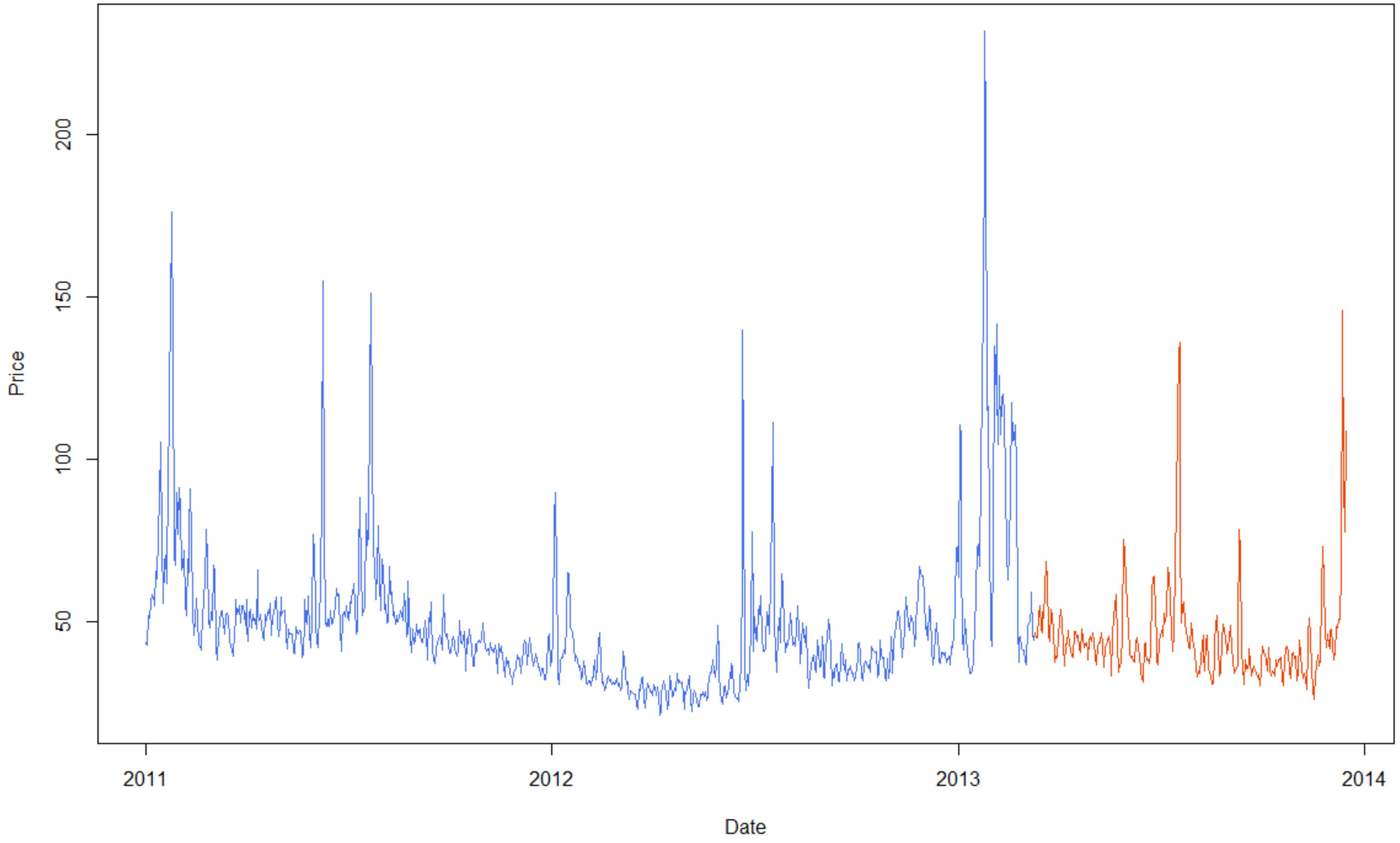
GEFCOM14 Electricity Load Data set

Results



GEFCOM14 Electricity Price Data set

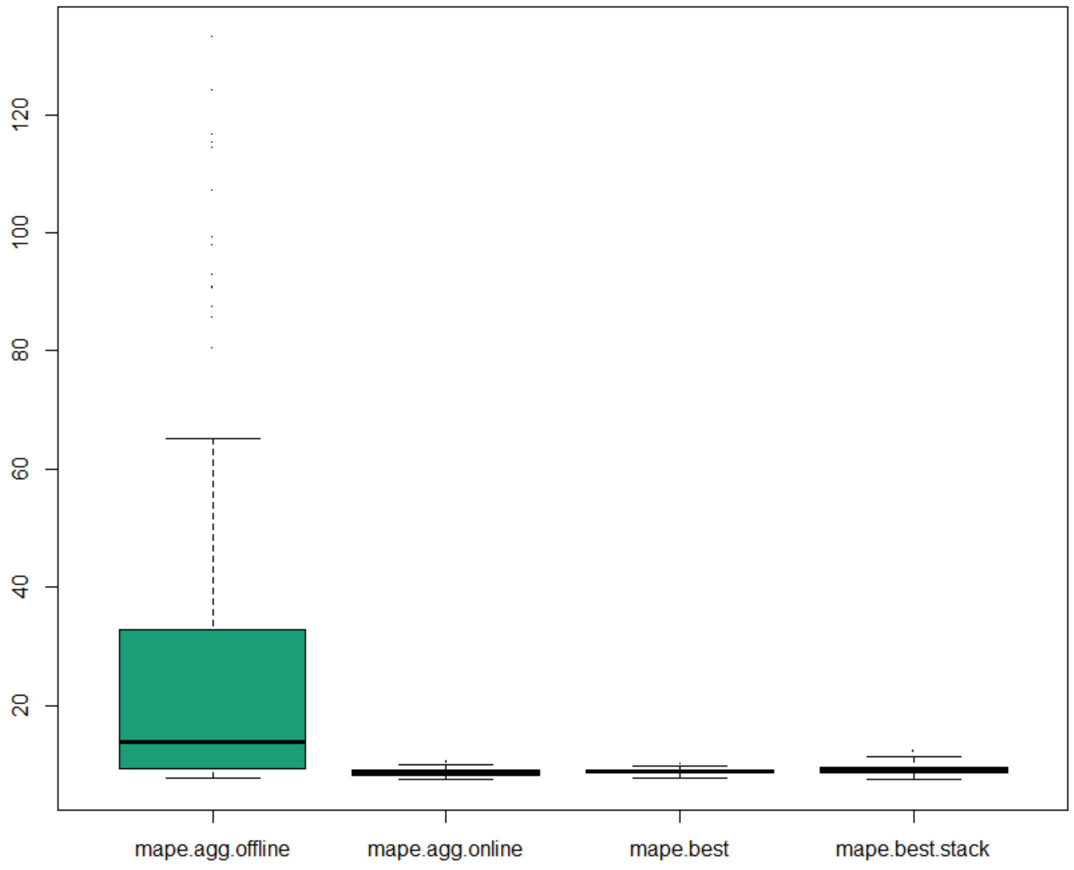
- Y: electricity price of a US zone, daily mean, 2011-01-01 / 2013-12-16
- X: lag price, zonal load and total load forecasts, calendar information



GEFCOM14 Electricity Price Data set

Results

Mean Absolute Percentage Error
Obtained for100 random splits (stacking)



Off Ml-pol fails due to non-stationarity

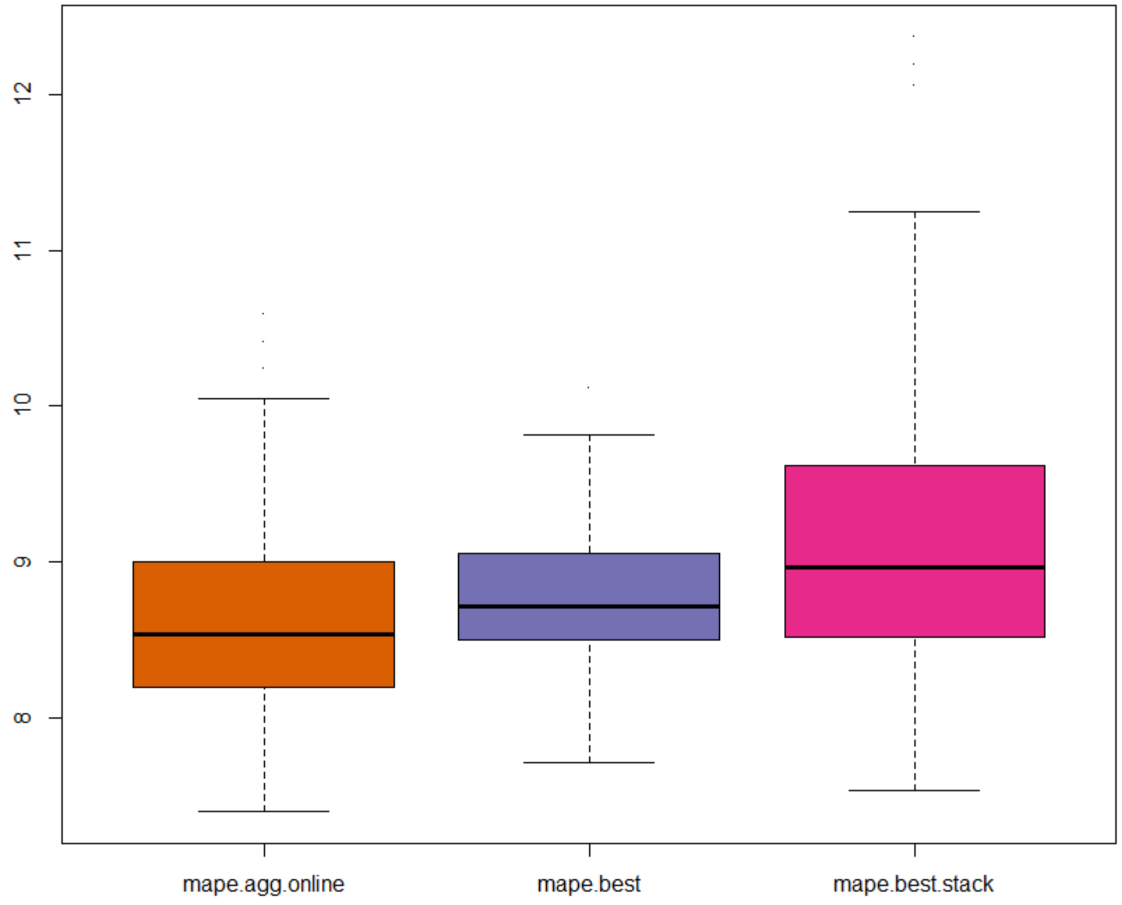
GEFCOM14 Electricity Price Data set

Results

1.7% improvement over the best expert for the on-line MLpol

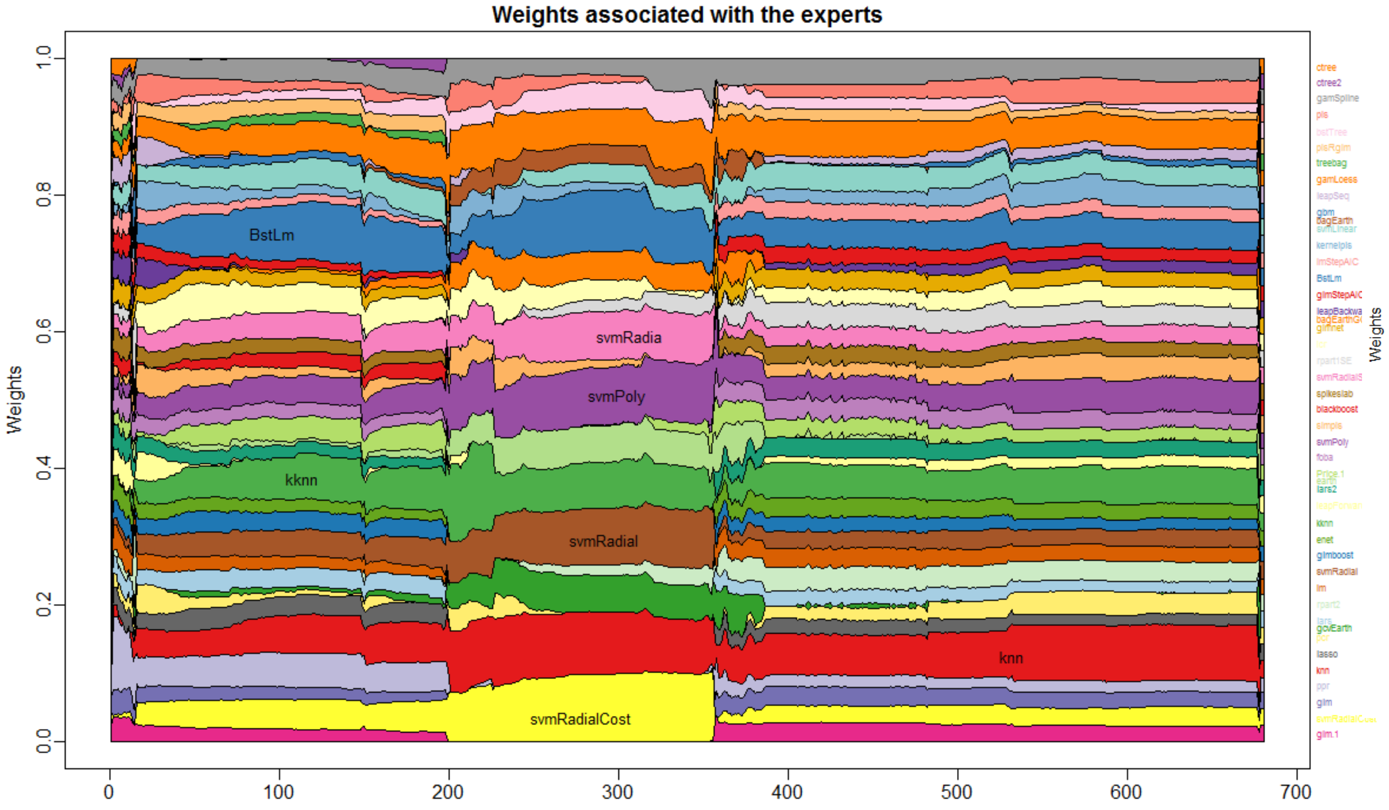
5.8% improvement over the best stacked expert for the on-line MLpol

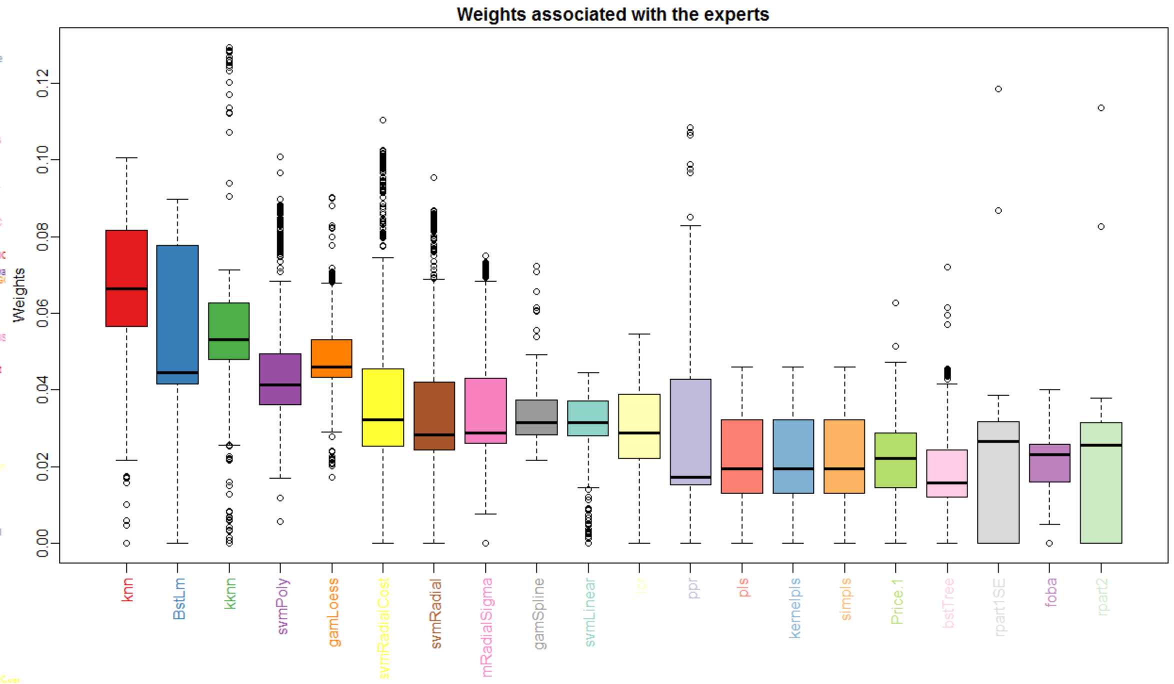
Mean Absolute Percentage Error

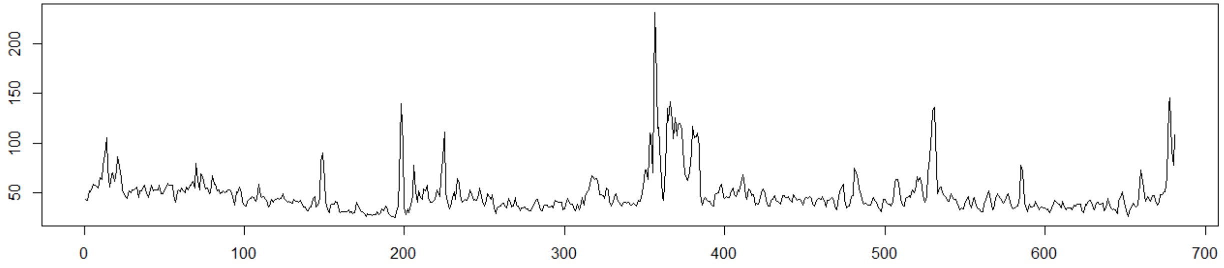


GEFCOM14 Electricity Price Data set

Results







Remarks/Perspectives

- Stacking for time series is not straightforward (block CV, stationnarity)
- Robust online aggregation works well in both cases (load and price)
- The OPERA package is on CRAN

  OPERA, Online Prediction by ExpeRt Aggregation
- Calibration of the experts, a need to improve default train parameters, scalability of some experts to p is not obvious
- Future work: PV, wind others benchmark datasets (ideas are welcome)

Bibliography

Breiman, L. Random Forests Mach. Learn., Kluwer Academic Publishers, 2001, 45, 5-32

Devaine, M.; Gaillard, P.; Goude, Y. & Stoltz, G. 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

Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine Annals of Statistics, 2000, 29, 1189-1232

Kuhn, M. The caret Package **2009**

Gaillard, P. & Goude, Y. Forecasting Electricity Consumption by Aggregating Experts; How to Design a Good Set of Experts
Modeling and Stochastic Learning for Forecasting in High Dimensions, Springer International Publishing, **2015**, *217*, 95-115

Gaillard, P. Contributions à l'agrégation séquentielle robuste d'experts~: travaux sur l'erreur d'approximation et la prévision en loi. Applications à la prévision pour les marchés de l'énergie
Université Paris-Sud 11, **2015**

Gaillard, P.; Goude, Y. & Nedellec, R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting *International Journal of Forecasting, Elsevier,* **2016**, *32*, 1038-1050

THANKS!