Forecasting electricity consumption by aggregating forecasts

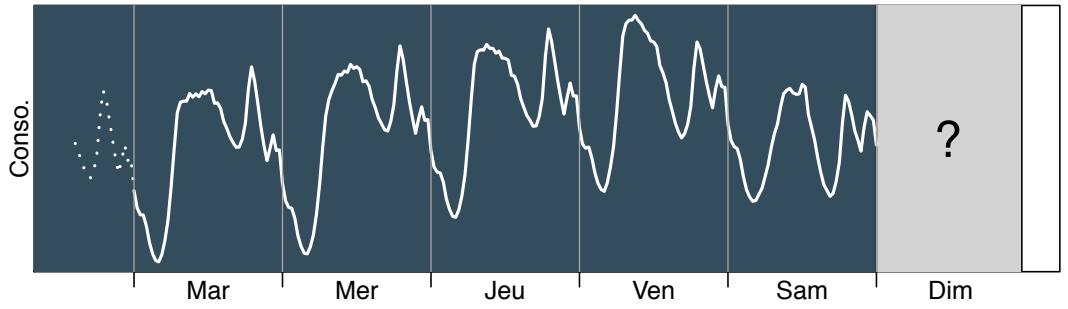
Pierre Gaillard

INRIA Paris

Joint work with: Yannig Goude (EDF R&D), Gilles Stoltz (Université Paris-Sud), Raphael Nedellec (EDF), and Marie Devaine

Industrial context

- Short term prediction (one day ahead) of the electricity consumption



- Important because electricity is hard to store

Production

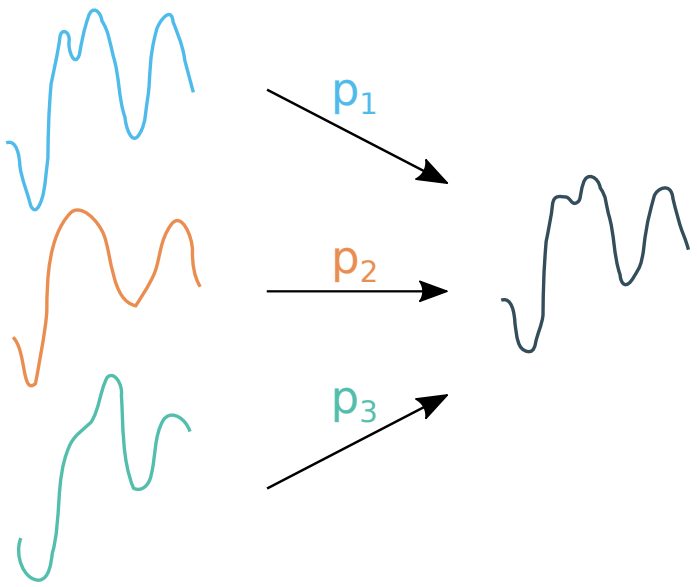
Demand

Many models are possible

- Regression models on splines (GAM)
- Autoregressive models
- Models on curves (CLR)
- Models based on similarities in the past (KWF)
- Machine learning models: random forests, boosting methods, deep learning,…
- Historical models of EDF
- …

Which model to choose?

Instead of picking one, we want to combine them

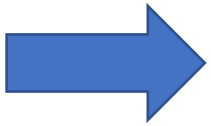


By
1) assigning a weight to each model
2) predicting the weighted average

How to choose the weights?

The electric environment is constantly evolving

- Changes in the usage: Electric cars, Energy saving light bulb
- Changes in the production: more renewable energy
- Political changes: opening to competition of electricity marketing and production



We want a model that can evolve over time and adapt itself automatically

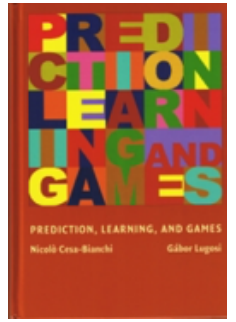
What assumption on the data?

- The stochastic process behind electricity consumption is complicated
- Forecasting models come from different communities and make highly different assumptions on the data

Hard to unify into a common framework!

We make no assumption on the data (for the combination part)

Setting of prediction of arbitrary sequences



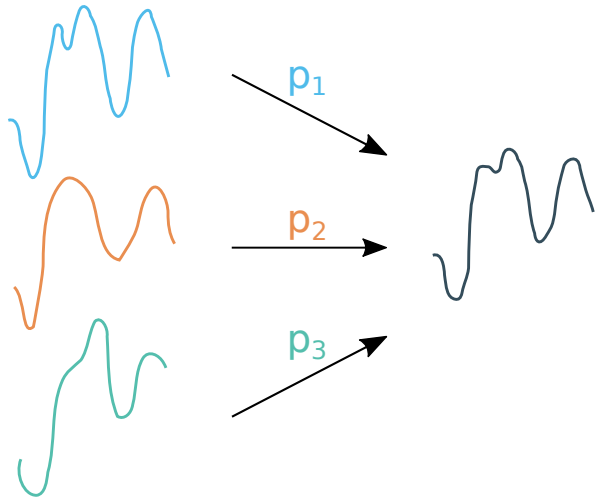
Prediction, learning, and Games, 2006
Cesa-Bianchi and Lugosi

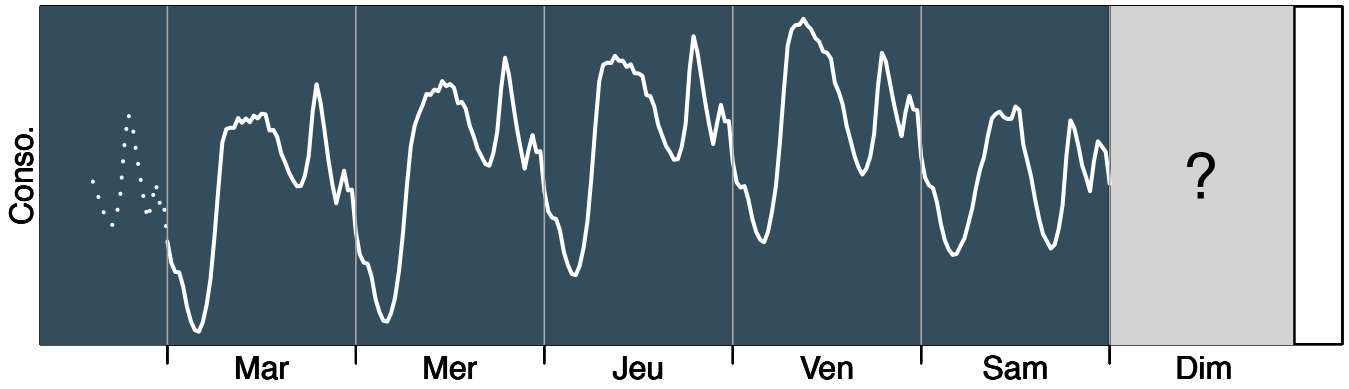
A sequential framework

- Each day t = 1, … ,T
  - Some experts (or models) suggest predictions $x_{k,t}$
  - We assign a weight p_{kt} to each expert k
  - We predict the weighted average

$$\hat{y}_t = \sum_k p_{k,t} x_{k,t}$$

  - We observe the true consumption y_t





Performance criterion

- Weights are updated every day according to past performances
- Garbage in garbage out: if no expert is good, there is no hope to provide good predictions

- Goal: perform almost as well as the best fixed predictor in hindsight

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}_t - y_t\right)^2 \underset{T\to\infty}{\longrightarrow} \min_k \frac{1}{T}\sum_{t=1}^{T}\left(x_{k,t} - y_t\right)^2$$

Our average error

Average error of the best model

This for all possible data y_t and x_t !

Application: electricity load forecasting

Goal: one day-ahead forecasting of the French electricity load

Data characteristics:

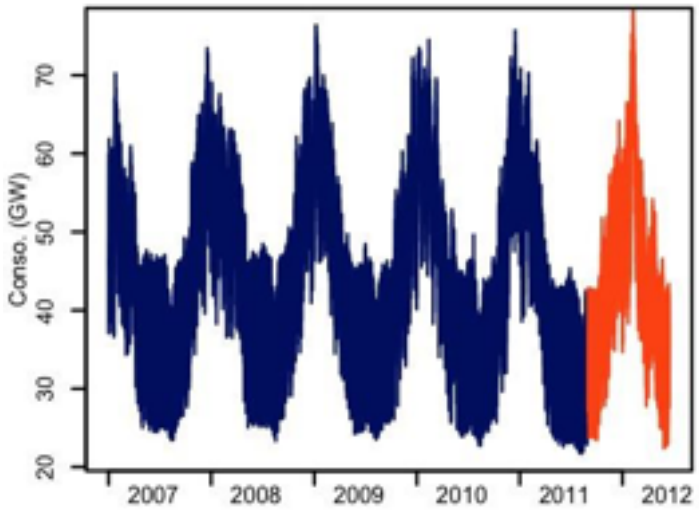
- Electricity demand for EDF clients, at half-hour steps
- Side information: weather (temperature, wind, nebulosity), data, loss of clients
- 2008 – 2011: training data set (for training the models)
- 2011 – 2012: test set (for combining them online)
- Typical values:

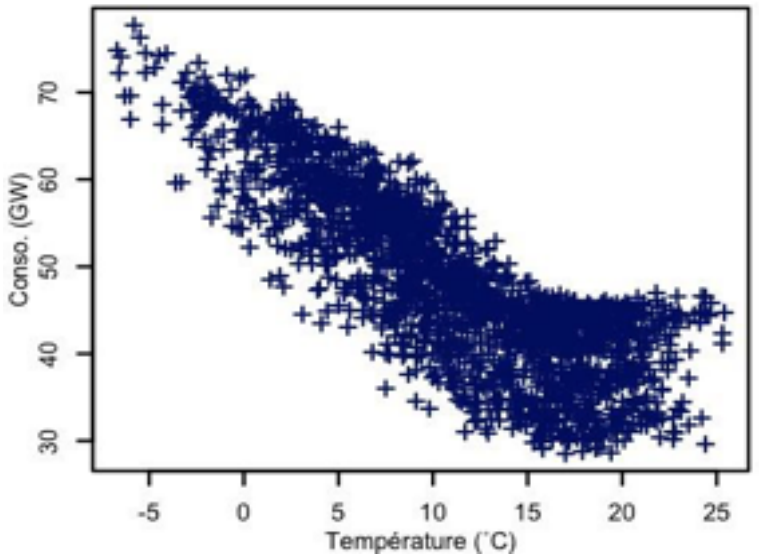
  Median: 43 GW, Max: 79 GW

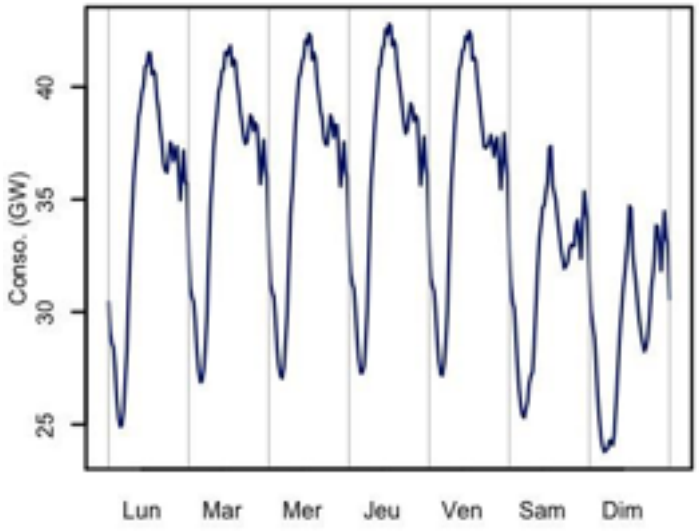
In total there are 1696 days.

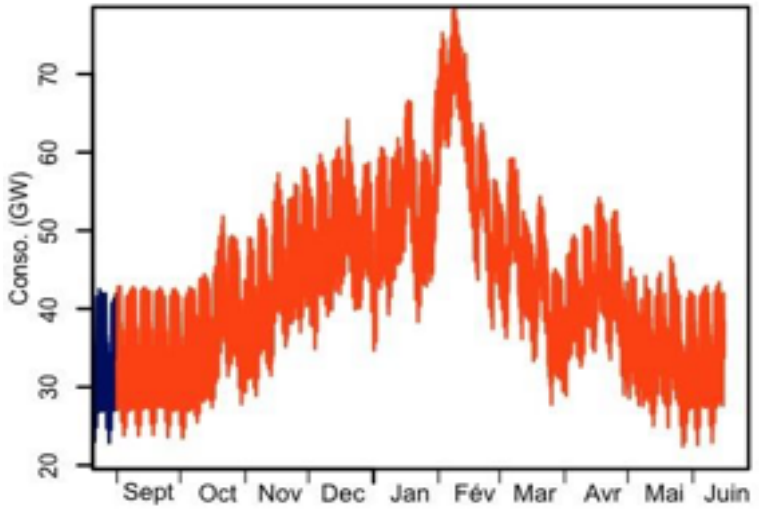
We remove uncommon days (public holidays +-2)

Data looks like…









Expert forecasters

- GAM: generalized additive models

(see Wood 2006, Wood, Goude, Shaw 2014)

- CLR: curve linear regression

(see Cho, Goude, Brossat, Yao 2013, 2014)

- KWF: functional wavelet-kernel approach

(see Antoniadis, Paparoditis, Sapatinas 2006, Antoniadis, Brossat, Cugliari, Poggi 2012, 2013)

How good are the experts?

Loss: RMSE and MAPE on the testing sets (with no warm-up period)

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \qquad \sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{y_t}$$

The smaller the better!

We look at the performance of the oracles

	Uniform average	Best expert	Best combination p
RMSE (MW)	725	744	629
MAPE (MW)	1.18	1.29	1.06

The exponentially weighted average forecaster

Vovk ’90
Littlestone and Warmuth ’94

Parameter: η

$$p_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (y_s - x_{k,s})\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} (y_s - x_{j,s})\right)}$$

Theoretical guarantees: for any bounded data

$$\frac{1}{T}\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - \min_k \frac{1}{T}\sum_{t=1}^{T} (x_{k,t} - y_t)^2 \leq \frac{\log K}{\eta T} + \frac{\eta B^2}{8} \leq B\sqrt{\frac{\log K}{T}} \longrightarrow 0$$

for $\eta = B^{-1}\sqrt{\frac{8 \log K}{T}}$

our average loss

average loss of the best expert in hindsight

Proof

Lemma (Hoeffding)

Let X be a random variable taking value in $[0, B]$. Then for any $s \in \mathbb{R}$

$$\log \mathbb{E}\left[e^{sX}\right] \leq s\mathbb{E}[X] + \frac{s^2 B^2}{8}$$

1. Upper bound the instantaneous loss $\widehat{\ell}_t$

$$\begin{aligned}
\widehat{\ell}_t = \ell(\widehat{p}_t \cdot x_t, y_t) &\overset{\text{by convexity}}{\leq} \widehat{p}_t \cdot \ell(x_t, y_t) \\
&\overset{\text{by Hoeffding}}{\leq} -\frac{1}{\eta} \log\left(\sum_{k=1}^{K} \widehat{p}_{k,t} e^{-\eta \ell_{k,t}}\right) + \frac{\eta B^2}{8} \\
&\overset{\text{by definition of } \widehat{p}_{k,t+1}}{=} -\frac{1}{\eta} \log\left(\frac{\widehat{p}_{k,t}}{\widehat{p}_{k,t+1}} e^{-\eta \ell_{k,t}}\right) + \frac{\eta B^2}{8} \\
&= \ell_{k,t} + \frac{1}{\eta} \log \frac{\widehat{p}_{k,t+1}}{\widehat{p}_{k,t}} + \frac{\eta B^2}{8}
\end{aligned}$$

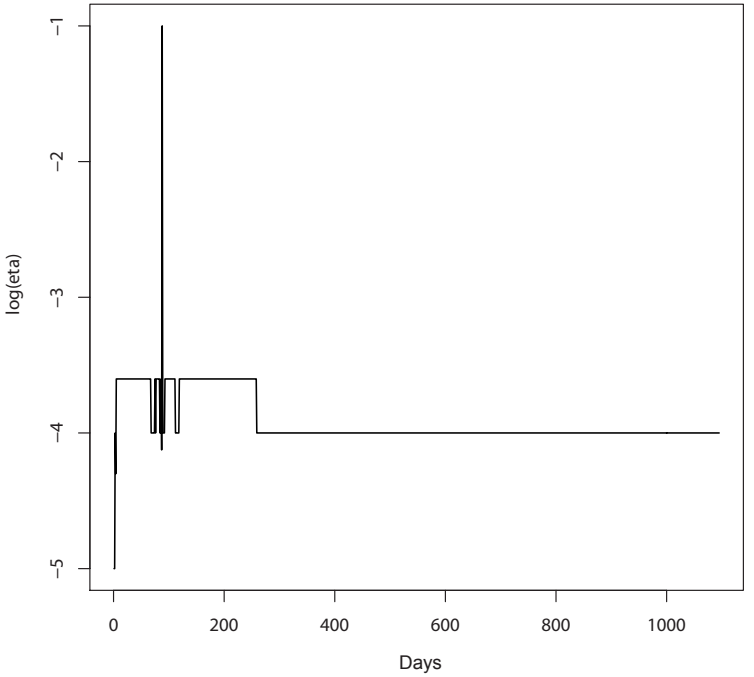
2. Sum over all t, the sum telescopes

$$\sum_{t=1}^{n} \widehat{\ell}_t - \ell_{k,t} \leq \frac{1}{\eta} \log \frac{\widehat{p}_{k,n+1}}{\widehat{p}_{k,1}} + \frac{\eta n B^2}{8} \leq \frac{\log K}{\eta n} + \frac{\eta B^2}{8}$$

Calibration of the learning rate

Best theoretical value:

$$\eta = B^{-1}\sqrt{\frac{8\log K}{T}}$$



Issue: T and B are not known in advance!

Solutions:

- Doubling trick
- Online adaptation by picking η_t according to the theoretical value η_t
- Optimization on a finite grid by choosing:

$$\eta_t \in \arg\min_{\eta} \left\{ \text{Loss of Exp. weights with } \eta \text{ until time } t-1 \right\}$$

Performance on the EDF data set

Benchmark and oracles (RMSE)

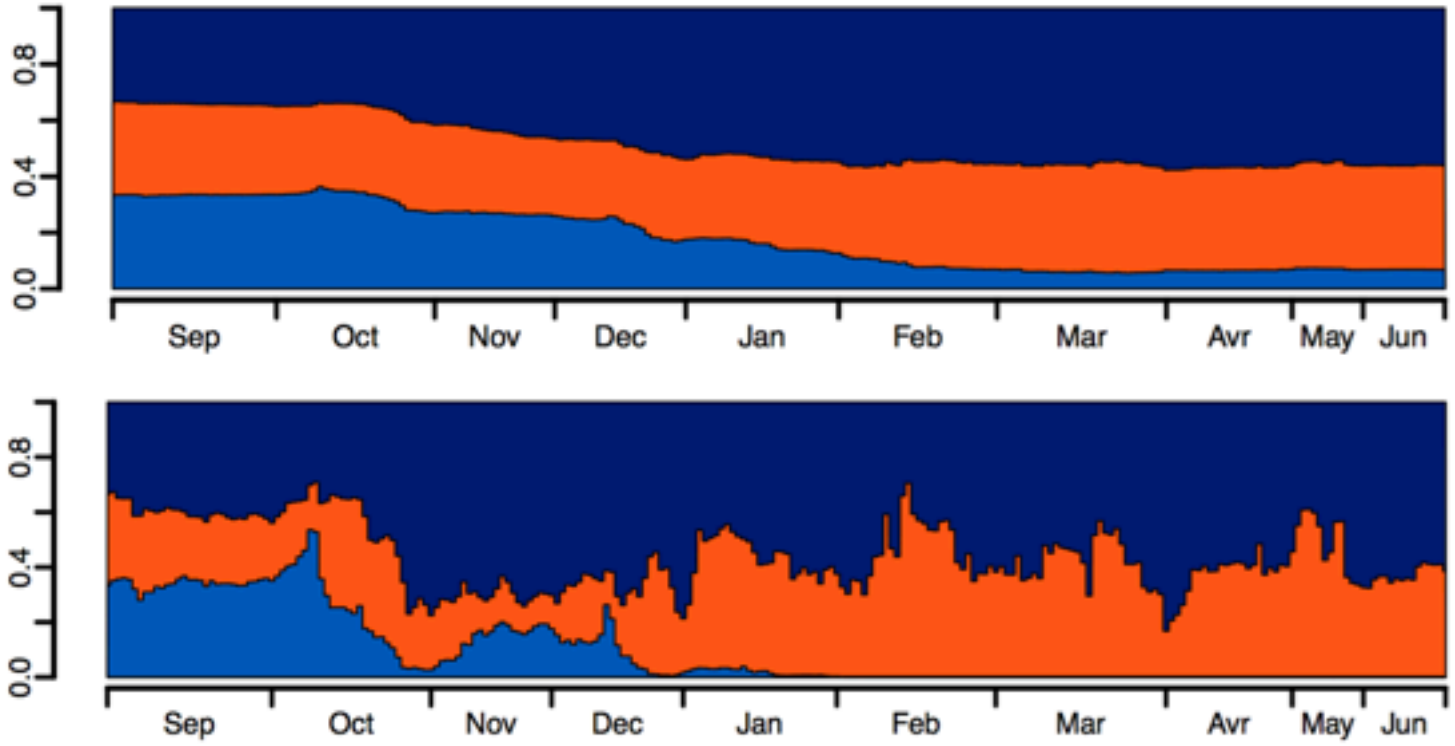
	Uniform	Best expert	Best weight p
RMSE (MW)	725	744	629

vs.

Aggregated forecasts

Exp. weights (best η for theory)	644
Exp. weights (best η on data)	644
Exp. weights (best η tuned on data)	625

Evolution of the weights



← Exp. weights (theory)

← Exp. weights (best η)

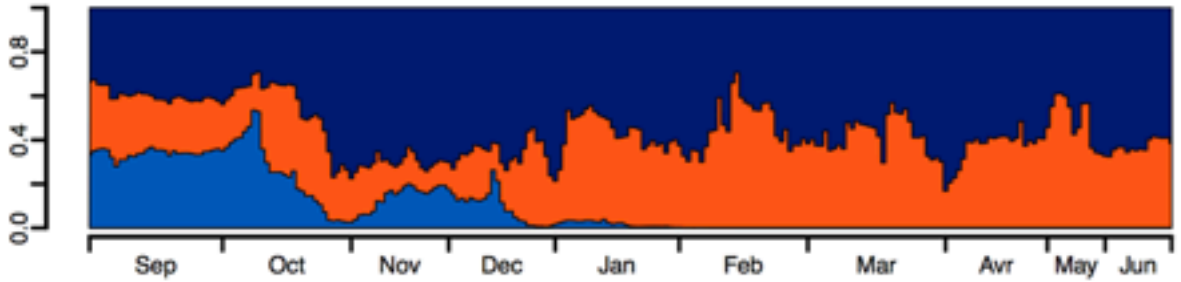
Weights change significantly over time and do not converge

Are all forecasters useful?

Yes!

Performance of Exp. Weights with 3 forecaster: 625

Vs only best two: 644



← Exp. weights (best η)

Forecasters not considered anymore can come back if needed

Time adaptive method

If the performance of experts vary over time, we need to make sure no weight vanishes completely to zero to never completely forget any expert.

Fixed share algorithm (Herbster and Warmuth '98): add a mixing step to ensure this

$$p_{t+1} = (1-\alpha)\text{Exp. weight update} + \frac{\alpha}{K}$$

Theoretical guarantees for competing with respect to the best sequences of experts (with few changes).

Our average error < Average error of the best sequence of experts with at most m changes + $\sqrt{\frac{m\log(K)}{T}}$

Similar performance on our data set than normal Exp. weights. More reactive (but less stable) weights.

Time adaptive method

Version with memory (i.e under sparsity assumption: only a small number of experts are useful)

[Bousquet and Warmuth, 2001, Cesa-Bianchi et al. 2011]

Other solutions:

- Sliding windows
- add a forgetting discount factor γ^{T-t} to forget old instances

$$p_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \gamma^{T-t}(y_t - x_{k,t})^2\right)}{\sum_j \exp\left(-\eta \sum_{s=1}^{t-1} \gamma^{T-t}(y_t - x_{j,t})^2\right)}$$

Best convex combination

If an expert provides inaccurate forecasts which compensate other expert forecasts, we should increase its weight!

More ambitious goal than competing with the best single expert:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}_t - y_t\right)^2 \underset{T\to\infty}{\longrightarrow} \min_p \frac{1}{T}\sum_{t=1}^{T}\left(p\cdot x_t - y_t\right)^2$$

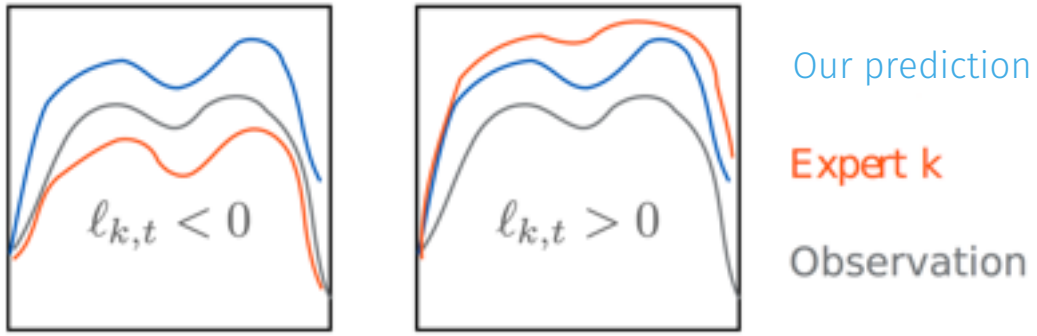
Our average error

Average error of the best fixed combination of experts

The gradient trick formalizes this idea.

For the square loss:

$$(x_{k,t} - y_t)^2 \quad \rightarrow \quad (\widehat{y}_t - y_t)(x_{k,t} - y_t)$$



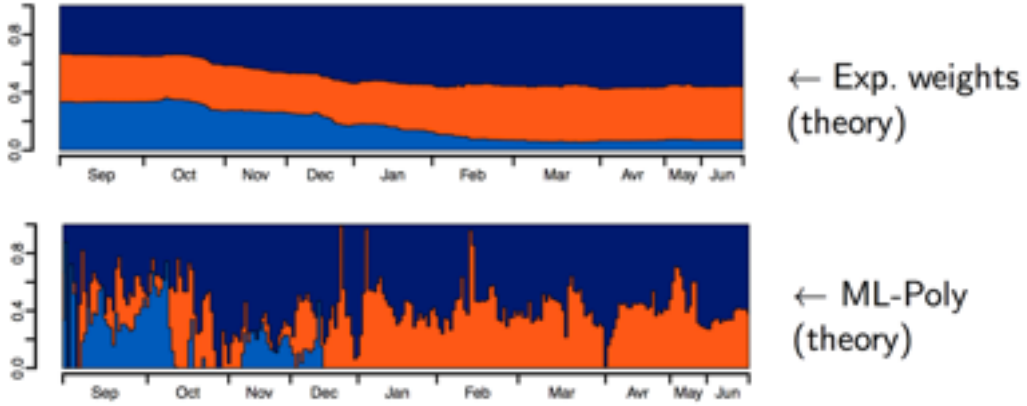
Real data are not arbitrary…

There is regularity, structure,…

Algorithms tuned to face arbitrary data are too pessimistic and careful

Can we adapt to data to exploit its structure and improved performance? Sometimes yes!

MLPoly, AdaHedge, Squint, MLProd, BOA



Better performance:

- If some expert is significantly better
- If the loss function is strongly convex
- Quantile bounds if many experts are good

$$\sqrt{\frac{\log K}{T}} \longrightarrow \frac{\log K}{T}$$

$$\sqrt{\frac{\log K}{T}} \longrightarrow \sqrt{\frac{\log(\text{proportion of good experts})}{T}}$$

Most of these methods are based on well-calibration of the learning parameter: Learn faster if the data allows it

Missing data (sleeping experts)

Some expert predictions may be missing at some times

Trick: replace their prediction with the one of the algorithm $\widehat{y}_t = p_t \cdot x_t$

$$p_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (y_t - x_{k,t})^2\right)}{\sum_j \exp\left(-\eta \sum_{s=1}^{t-1} (y_t - x_{j,t})^2\right)}$$

This works with any algorithm and has theoretical guarantees!

Useful, if one has specialized experts

Ref: Gaillard, Stoltz, van Erven, 2015

Specialized experts

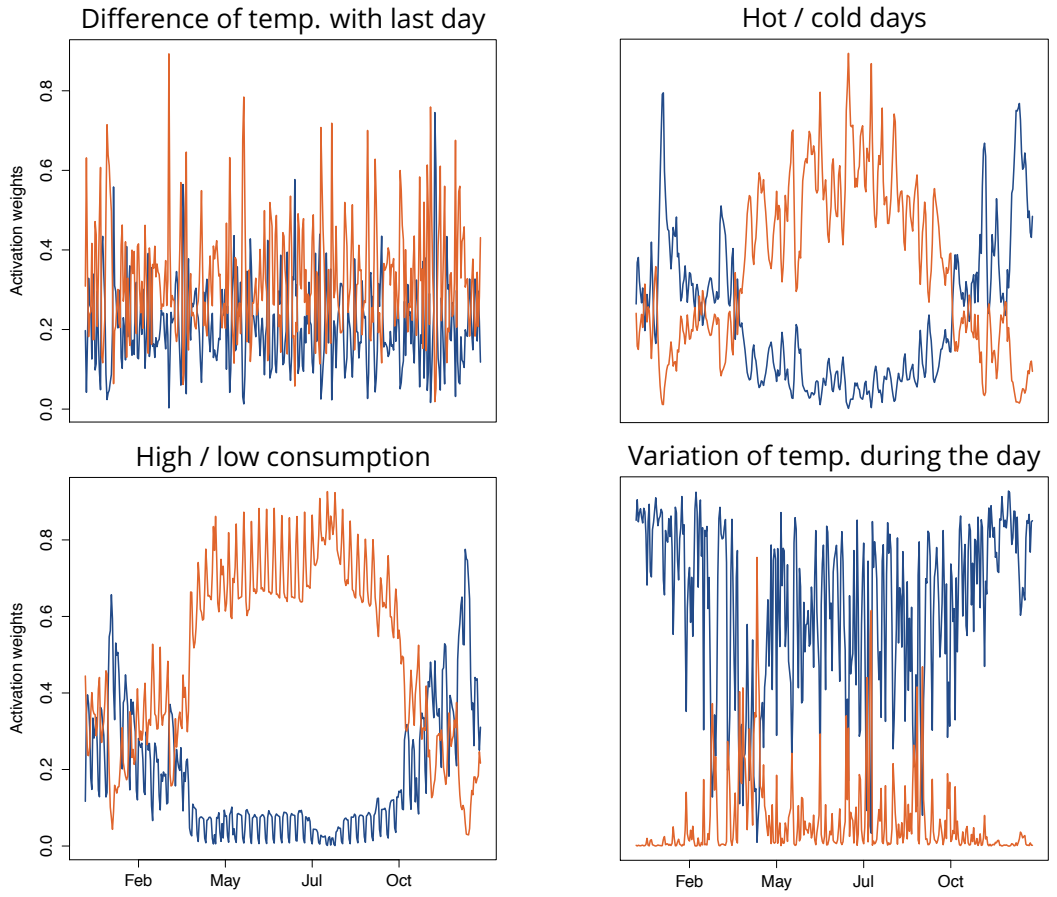
Idea: focus the training of some experts on specific situations

Examples:

- meteorological situations

(high/low temperature, change of temperature)

- Winter / summer
- High / low consumption



Such specialized experts suggest prediction only the days corresponding to their scenario

Enrich the set of experts?

Ref: Gaillard, Goude (2014)

The performance of our method is a trade-off between two errors

Our error = Error of the best expert + regret

Approximation error

Estimation error

The theoretical performance increases slowly with the number of experts as $\sqrt{\log(K)/T}$

Enriching the set of experts can be highly beneficial:

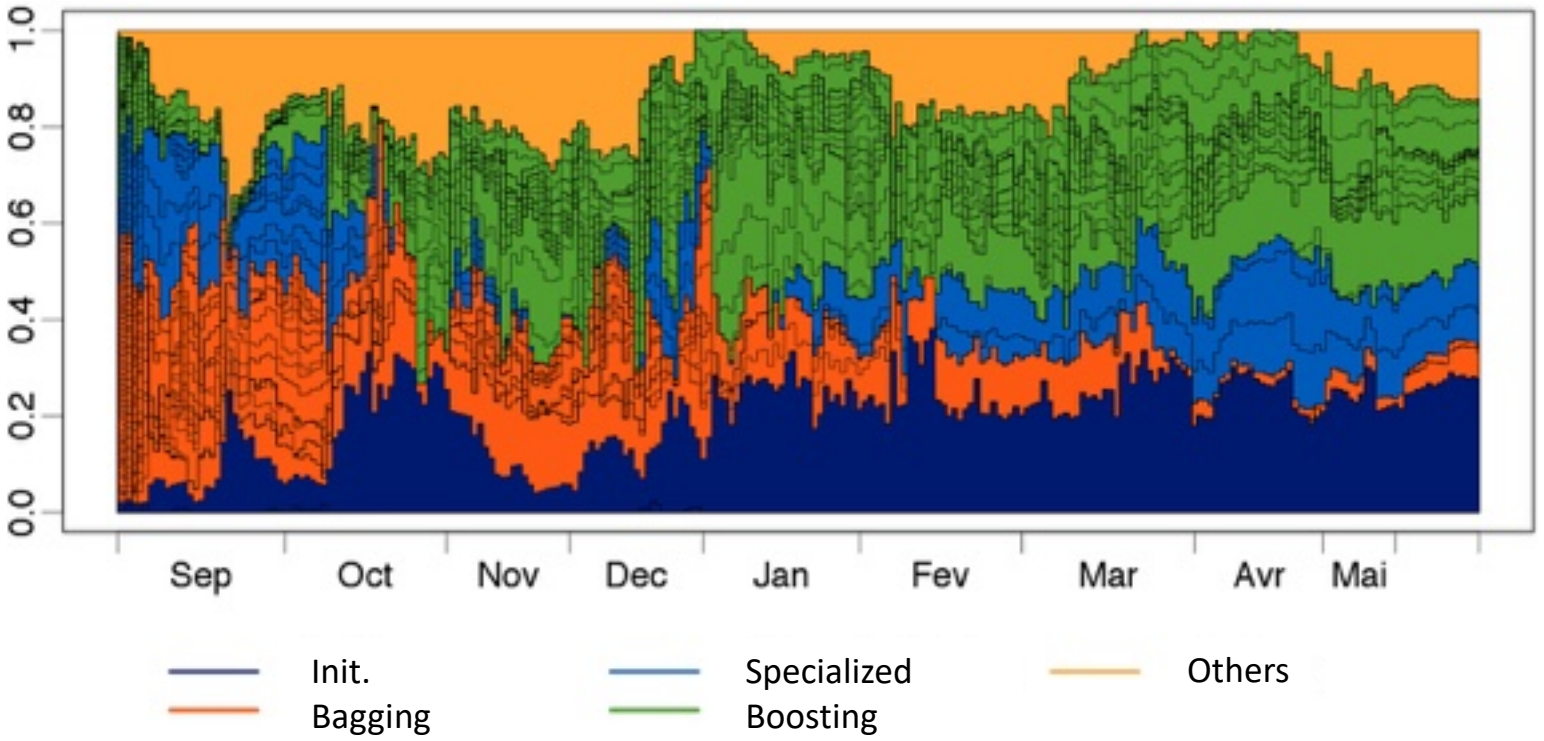
- Modifying initial experts using bagging methods
- Adding new experts that aim at correcting errors of current experts
- Specialized experts,…

3 experts -> 133 experts

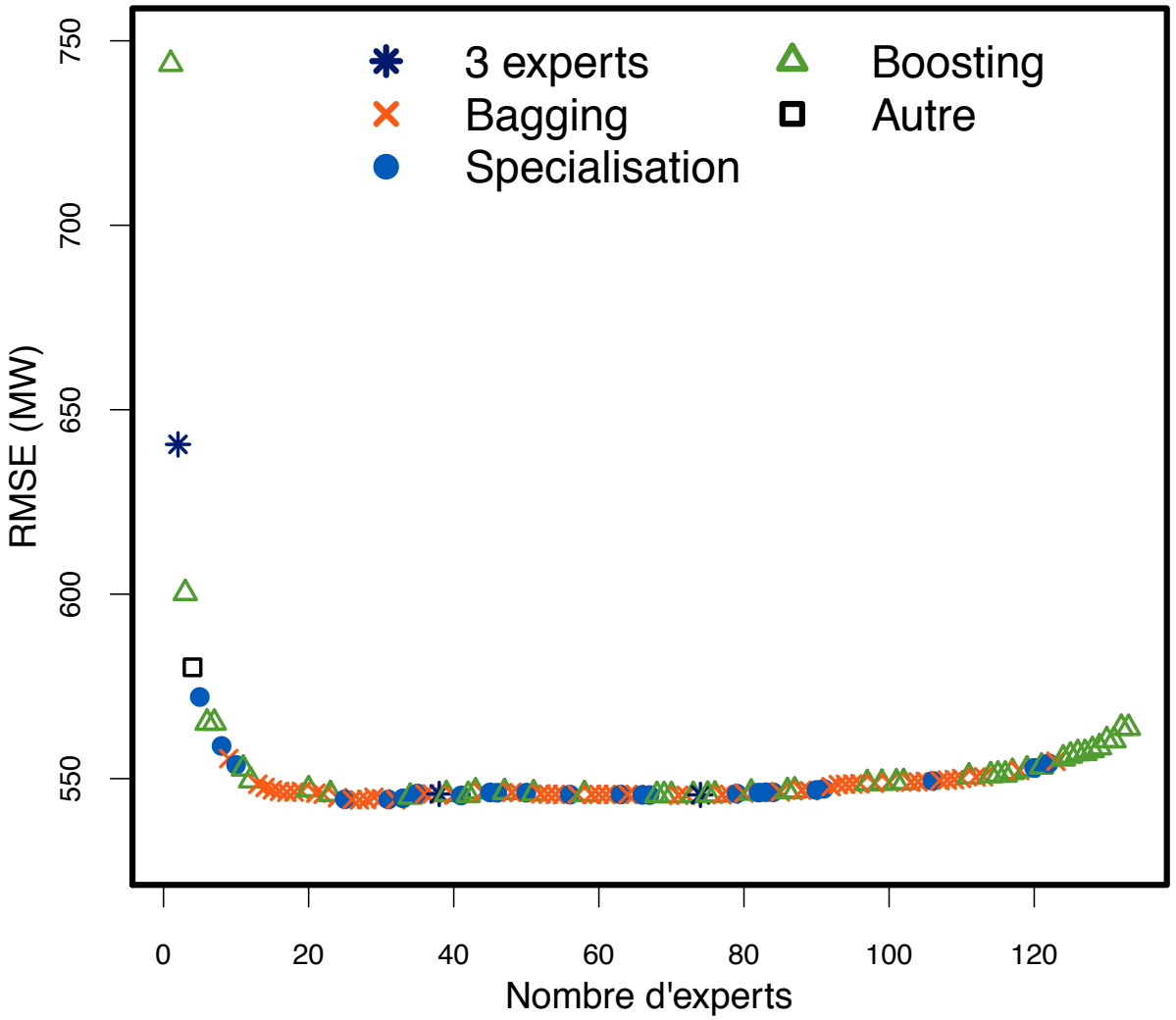
Performance of adding new experts

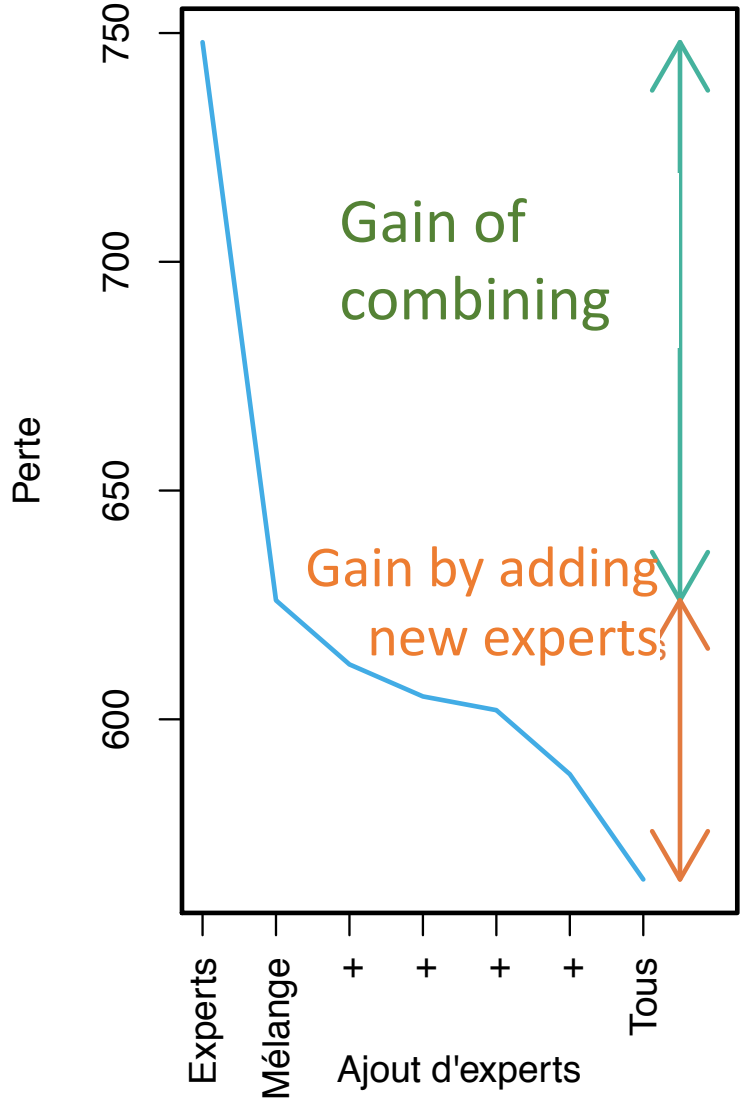
	Best expert	Best fixed combination	Exp. Weights (best η $for\ theory$)	Exp. Weights (best η for data)	ML-Poly
3 experts	744	629	644	625	626
133 experts	744	521	737	591	565

Evolution of the weights



Performance of adding new experts



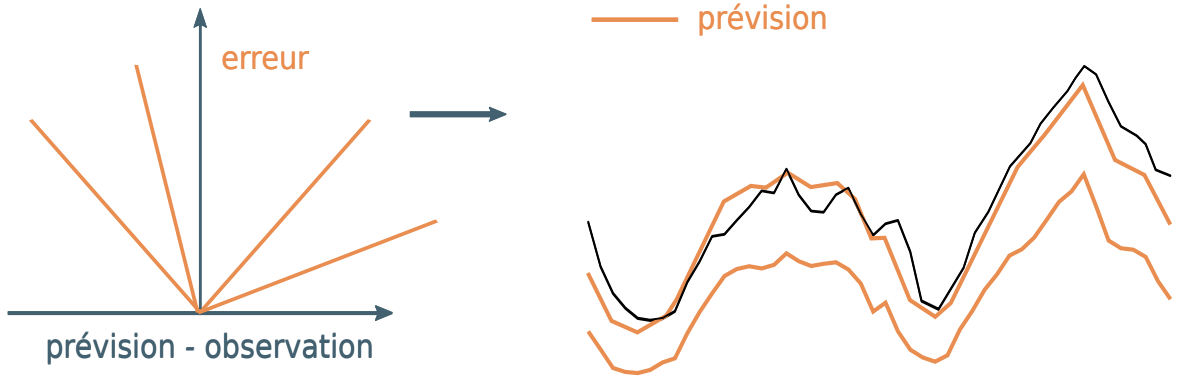


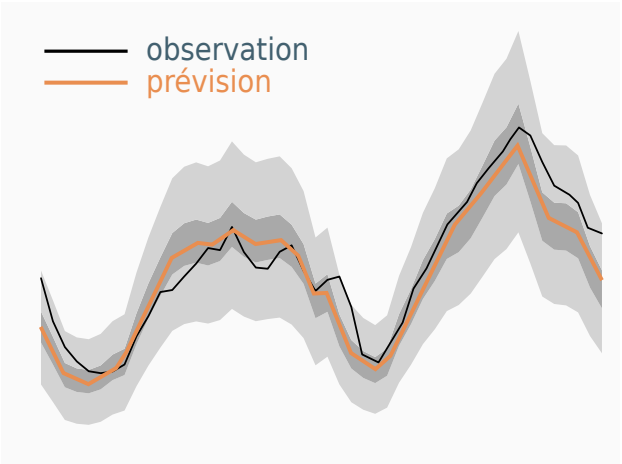
Probabilistic prediction

Can we trust our forecasts? Renewable energies add randomness in the production.

Need of probabilistic forecasts (instead of forecasting the mean only)

Easy trick here: just change the loss function: use the quantile loss

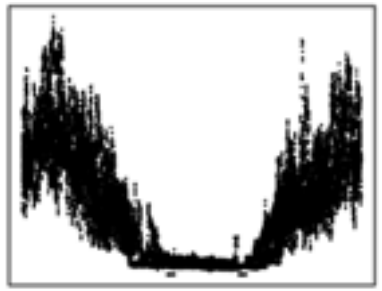




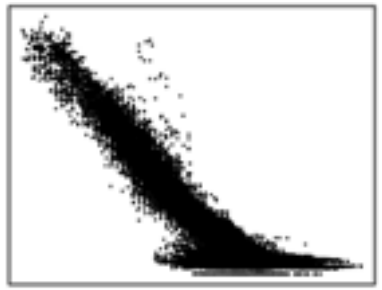
We used this method on GefCom 2014 competition with Y. Goude and R. Nedellec. Worked great!

Universality – Other data sets

Heat load forecasting

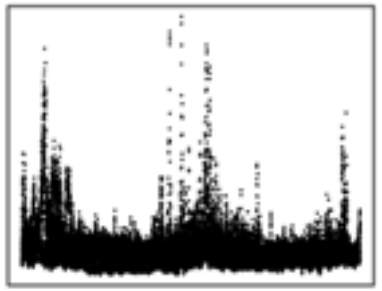


Annual position

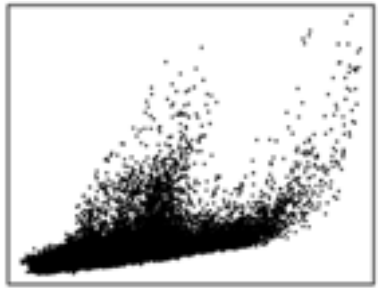


Temperature

Electricity price

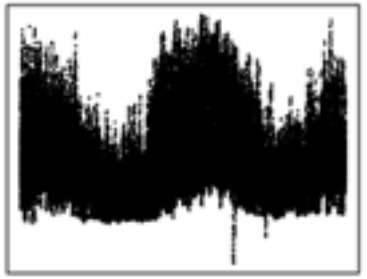


Annual position



Electric load

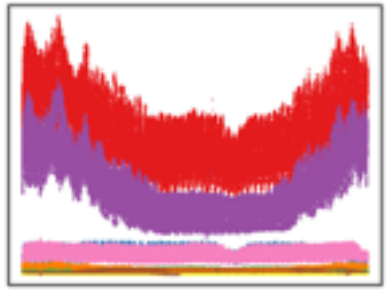
US electricity load



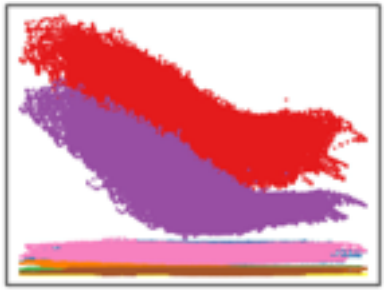
Annual position

Temperature

Groups of consumers



Annual position



Temperature

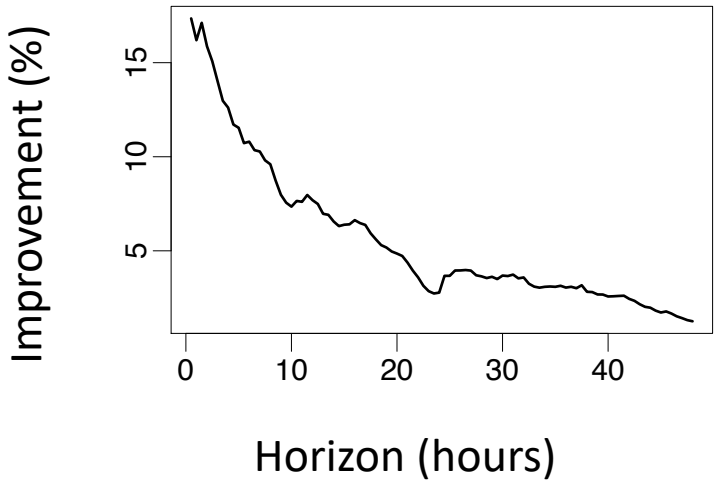
Challenge: use individual data

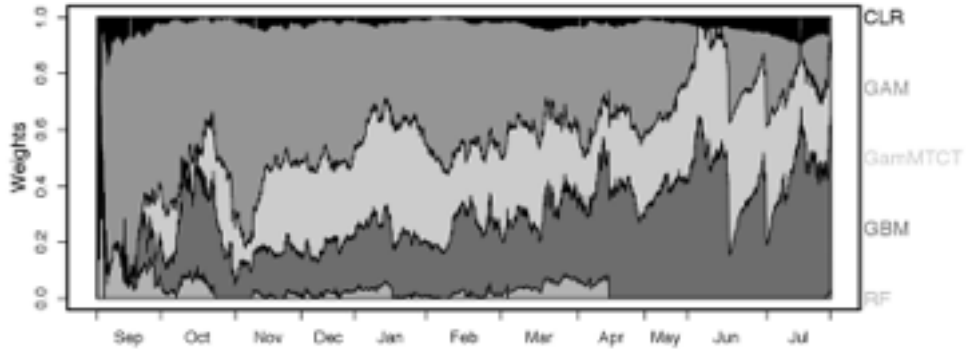
Universality – horizon of prediction

On a similar data set

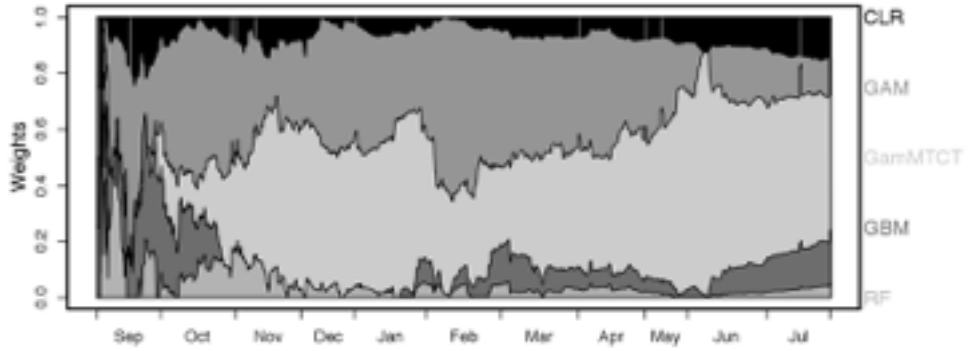
- we trained 5 experts making predictions at several horizons
- we adapted combining algorithm to multi-horizon forecasts

Gain of performance (RMSE) with respect of the best expert

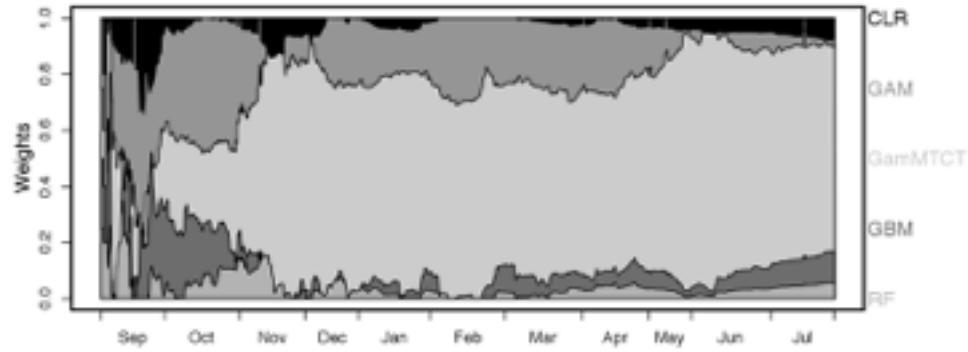




(a) 1 half-hour ahead



(b) 12 hours ahead

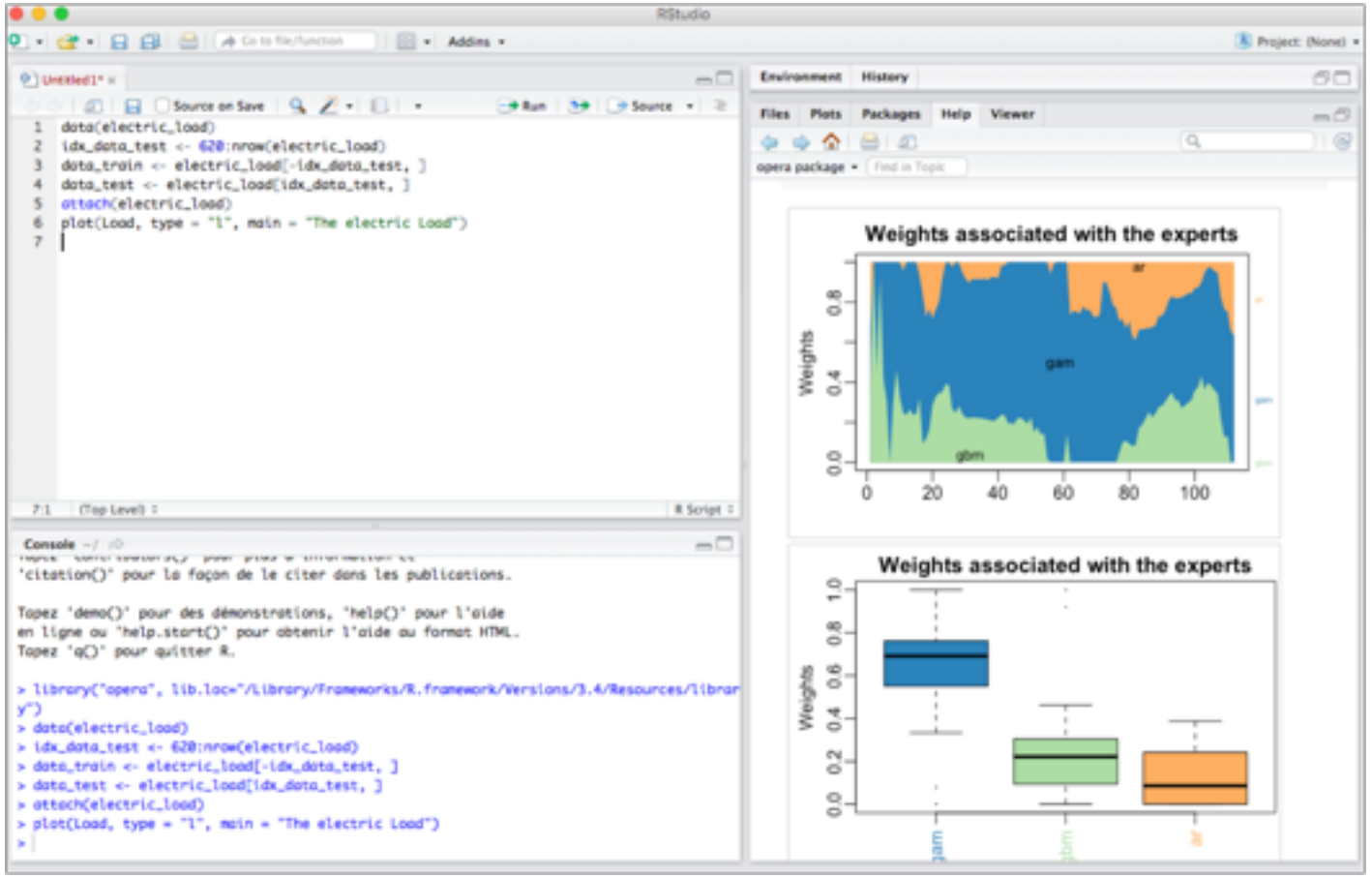


(c) 24 hours ahead

Most of these methods are implemented in the R-package opera

Example of usage on my webpage:

http://pierre.gaillard.me/opera.html



Conclusion

- Combining forecasts can greatly improve the performance
- Building good experts is important (do it automatically)
- Well-calibration and data-adaptive methods is important
- Works quite generally

Thank you!