Agrégation séquentielle d'experts

Comment tirer parti de l'embarras du choix de prévisions?

Exemples de collaborations recherche avec des entreprises

Gilles Stoltz

Attention!

Cet exposé est un exposé très prosélyte

1.

Qui décrit l'art et l'intérêt des collaborations avec des entreprises

Et finira par la présentation d'un sujet de thèse CIFRE...

2.

Qui vous donnera de (la moitié de) mon cours de M2 du S2

Mais il sera 100% théorique (et accessible aux probabilistes purs)

Introduction – part 1

Interested in forecasting a sequential quantitative phenomenon

What?	With whom?	When? (Time scale)
Exchange rates	HEC PARIS	Next month
Electricity consumption	EDF	Next hour or day
Sales volumes	Cdiscount	6 weeks from now
Ozone concentration	INERIS maîtriser le risque pour un développement durable	Next day
Oil and gas production	IFP Energies nouvelles	Next 5 - 10 years

Predict a time series $y_1, y_2, \ldots \in \mathbb{R}$ (= quantitative phenomenon)

Ensemble forecasts are available, e.g., given by some stochastic or machine-learning models (for us: black boxes)

At each instance t, forecasting black-box $j \in \{1, \ldots, N\}$ outputs

$$f_{j,t} \equiv f_{j,t}\left(y_1^{t-1}\right)$$

Observations and predictions are made in a sequential fashion:

The prediction $\widehat{y}_t$ of y_t is determined based

- on the past observations $y_1^{t-1} = (y_1, \ldots, y_{t-1})$,
- and the current and past ensemble forecasts $f_{j,s}$, where $s \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, N\}$

Our solution:

Convex / Linear combinations of the ensemble forecasts,
with adaptive weights $p_{1,t}, \ldots, p_{N,t}$

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

Our research work

- Most of the theory behind it is on how to pick the weights
- Most of the practice (strongly depends on the field) is how to build the base forecasts $f_{j,t}$
- Not only: practice indicates what theoretical extensions are to be designed (will provide examples thereof)

Performance guarantee:

What the best black-box can do (and beyond)

Introduction – part 2

Typical solution: convex (or linear) combinations of the ensemble forecasts, with adaptive weights $\mathbf{p}_t = \big(p_{1,t}, \ldots, p_{N,t}\big)$

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

The observations y_t will not be considered stochastic anymore at this stage; thus the performance criterion will be a relative one

Given a convex loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, e.g., the square loss $\ell(x, y) = \|x - y\|^2$:

The cumulative losses of the statistician and of the constant convex combinations $\mathbf{q} = (q_1, \ldots, q_N)$ of the forecasts equal

$$\widehat{L}_T = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) \quad \text{and} \quad L_T(\mathbf{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

The regret R_T is defined as the difference

$$\widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) = \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \min_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

We are interested in aggregation rules with (uniformly) vanishing per-round regret,

$$\limsup_{T\to\infty} \quad \frac{1}{T} \sup\left\{\widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q})\right\} \leqslant 0$$

The supremum is over all possible sequences of observations and of ensemble forecasts (not just over most of these sequences!)

Remarks:

- Hence the name "prediction of individual sequences" (or robust aggregation of ensemble forecasts)
- The best convex combination $\mathbf{q}^\star$ is known in hindsight whereas the statistician has to predict in a sequential fashion

This framework leads to a meta-statistical interpretation:

- ensemble forecasts are given by some statistical forecasting methods, each possibly tuned with a different given set of parameters
- these ensemble forecasts relying on some stochastic model are then combined in a robust and deterministic manner

The cumulative loss of the statistician can be decomposed as

$$\widehat{L}_T = \min_{\mathbf{q}} L_T(\mathbf{q}) + R_T$$

In words:

cumulative loss = approximation error + sequential estimation error

Disclaimer

We could also consider batch learning methods to aggregate forecasts, like

- BMA (Bayesian model averaging),
- CART (classification and regression trees),
- random forests, etc.,

or even selection methods, and apply them online, by running a batch analysis at each step

$\rightarrow$ We instead resort to "real" online techniques that, in addition, come up with theoretical guarantees even in non-stochastic scenarios

First study

Forecasting of the electricity load

Data source: EDF R&D

Authors: Pierre Gaillard and Yannig Goude

Reference: Proceedings of WIPFOR '2013

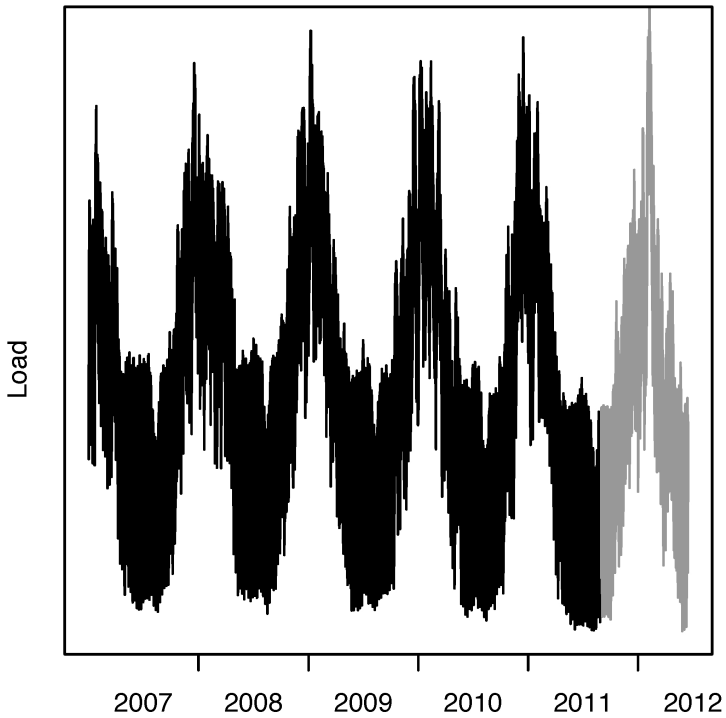
Some characteristics of one among the studied data sets:

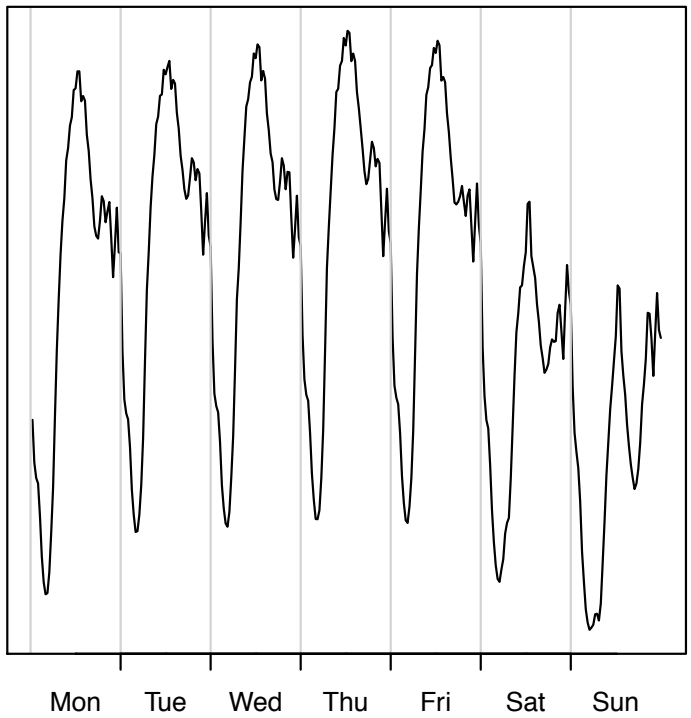
- January 1, 2008 – August 31, 2011 as a training data set
- September 1, 2011 – June 15, 2012 (excluding some special days) as testing set
- Electricity demand for EDF clients, at a half-hour step
- Typical values: median $= 43\,496$ MW
  maximum $= 78\,922$ MW
- Three forecasters: GAM, CLR, KWF

$\rightarrow$ Instead of trusting only one model/base forecaster ("selection"), we proceed in a more greedy way and consider ensemble forecasts, which we combine sequentially ("aggregation")

This leads to more accurate and more stable (meta-)predictions

Data looks like...





Convex loss functions considered:

- square loss $\ell(x, y) = (x - y)^2$ $\rightarrow$ RMSE
- absolute percentage of error $\ell(x, y) = |x - y|/|y|$ $\rightarrow$ MAPE

Operational constraint:

One-day ahead prediction at a half-hour step, i.e., 48 aggregated forecasts

Ensemble 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)

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

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}_t - y_t\right)^2} \qquad \text{and} \qquad \frac{1}{T}\sum_{t=1}^{T}\frac{\left|\widehat{y}_t - y_t\right|}{y_t}$$

How good are our building blocks? See the "oracles" below

	Uniform mean	Best forecaster	Best convex **p**	Best linear **u**
RMSE (MW)	725	744	629	629
MAPE (%)	1.18	1.29	1.06	1.06

In this article the focus is to create more base forecasting methods and to improve the oracles accordingly (and in turn, the performance of the aggregation methods)

A strategy to pick convex weights

Let's do some maths!

Reminder of the aim and setting:

Given a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Choose sequentially the convex weights $p_{j,t}$

To uniformly bound the regret with respect to all sequences of observations y_t and ensemble forecasts $f_{j,t}$:

$$\sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \min_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

When ℓ is convex and differentiable in its first argument:

For all $x, y \in \mathbb{R}^d$,

$$\forall x' \in \mathbb{R}^d, \qquad \ell(x, y) - \ell(x', y) \leqslant \nabla \ell(x, y) \cdot (x - x')$$

Assumption OK for RMSE, MAE, MAPE, etc.

To uniformly bound the regret with respect to all convex weight vectors $\mathbf{q}$, we write

$$\max_{\mathbf{q}} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t},\, y_t\right) - \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j\, f_{j,t},\, y_t\right)$$

$$\leqslant \max_{\mathbf{q}} \sum_{t=1}^{T} \nabla\ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot \left(\sum_{j=1}^{N} p_{j,t} f_{j,t} - \sum_{j=1}^{N} q_j f_{j,t}\right)$$

$$= \max_{\mathbf{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \sum_{j=1}^{N} q_j \widetilde{\ell}_{j,t}\right)$$

$$= \sum_{t=1}^{T}\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\dots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}$$

where we denoted

$$\widetilde{\ell}_{j,t} = \nabla\ell\left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\, y_t\right) \cdot f_{j,t}$$

Considering the (signed) pseudo-losses $\widetilde{\ell}_{j,t} = \nabla \ell \left(\sum_{k=1}^{N} p_{k,t} f_{k,t},\ y_t \right) \cdot f_{j,t}$

the regret is smaller than $\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}$

Exponentially weighted averages: $p_{j,1} = 1/N$ and

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

ensure that if all $\widetilde{\ell}_{j,t} \in [m, M]$, then

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t} \leqslant \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T$$

References: Vovk '90; Littlestone and Warmuth '94

Proof by mere calculus

Hoeffding's lemma: for all convex weights $(p_1, \ldots, p_N)$ and all numbers $u_1, \ldots, u_N$ with range $[b, B]$,

$$\ln \sum_{j=1}^{N} p_j e^{u_j} \leqslant \frac{(B-b)^2}{8} + \sum_{j=1}^{N} p_j u_j$$

For all $t = 1, 2, \ldots,$

$$\begin{aligned} -\eta \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} &= -\eta \sum_{j=1}^{N} \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)} \widetilde{\ell}_{j,t} \\ &\geqslant \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{t} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)} - \frac{\eta^2}{8}(M-m)^2 \end{aligned}$$

A telescoping sum appears and leads to

$$\begin{aligned} \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} &\leqslant \underbrace{-\frac{1}{\eta} \ln \frac{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{T} \widetilde{\ell}_{j,s}\right)}{N}} + \eta \frac{(M-m)^2}{8} T \\ &\leqslant \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t} + \frac{\ln N}{\eta} \end{aligned}$$

Obtained regret bound optimized over η:

$$R_T \leqslant \min_{\eta>0} \left\{ \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T \right\} = (M-m)\sqrt{\frac{T}{2}\ln N}$$

for the (theoretical) optimal choice $\eta^\star = \frac{1}{M-m}\sqrt{\frac{8\ln N}{T}}$

Issues:

- the parameters T and $[m, M]$ not always known beforehand
- even if they were, $\eta^\star$ leads to a poor performance

Solutions for the first issue (still poor performance):

- "doubling trick"
- adaptive learning rates η_t, picked according to some theoretical formulas

So, theoretically satisfactory solutions do not work well in practice!

This is what we do instead. (It is very different from techniques like cross-validation: we exploit the sequential fashion.)

The exponentially weighted average strategy $\mathcal{E}_\eta$ with fixed learning rate η picks

$$p_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

We denote its cumulative loss $\widehat{L}_t(\eta) = \sum_{s=1}^{t} \ell\left(\sum_{j=1}^{N} p_{j,s}(\eta) f_{j,s},\, y_s\right)$

Based on the family of the $\mathcal{E}_\eta$, we build a data-driven meta-strategy which at each instance $t \geqslant 2$ resorts to

$$\mathbf{p}_{t+1}(\eta_t) \qquad \text{where} \qquad \eta_t \in \operatorname*{arg\,min}_{\eta>0} \widehat{L}_t(\eta)$$

Reference: An idea of Vivien Mallet

Other natural variants: Focus on the most recent losses

Moving sums (with window of size H)

$$p_{j,t} = \frac{\exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=\max\{1,t-H\}}^{t-1} \widetilde{\ell}_{k,s}\right)}$$

Regret is $\geqslant \square\, T$ in the worst case

Discounted losses (with discounts given by a sequence $\beta_t \searrow 0$)

$$p_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \widetilde{\ell}_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} (1+\beta_{t-s}) \widetilde{\ell}_{k,s}\right)}$$

Sublinear regret bounds hold for suitable sequences (β_t) and (η_t):

$$t\,\eta_t \longrightarrow 0 \qquad \text{and} \qquad \eta_t \sum_{s \leqslant t} \beta_s \longrightarrow 0$$

(We often take $\beta_s = \square/s^2$ in the experimental studies.)

A strategy to pick linear weights

Some of you may know it in a stochastic setting!

Linear combinations: Ridge regression (and the LASSO?)

Ridge regression — introduced in the 70s by Hoerl and Kennard:

$$\mathbf{v}_t \in \operatorname*{arg\,min}_{\mathbf{u}\in\mathbb{R}^N} \left\{ \lambda \|\mathbf{u}\|_2^2 + \sum_{s=1}^{t-1} \left(y_s - \sum_{j=1}^{N} u_j \, f_{j,s} \right)^2 \right\}$$

It also exhibits a sublinear regret against individual sequences: for all $y_t \in [-B, B]$ and $f_{j,t} \in [-B, B]$, for all $\mathbf{u} \in \mathbb{R}^N$

$$\sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} v_{j,t} \, f_{j,t} \right)^2 - \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{N} u_j \, f_{j,t} \right)^2 \leqslant \lambda \|\mathbf{u}\|_2^2 + 4NB^2 \left(1 + \frac{NTB^2}{\lambda} \right) \ln \left(1 + \frac{TB^2}{\lambda} \right)$$

References: Vovk '01; Azoury and Warmuth '01; Gerchinovitz '11

The bound can be $O\big(\sqrt{T} \ln T\big)$ with λ of the order of $1/\sqrt{T}$

We do not know any such regret bounds for the LASSO (yet?)

These methods can compensate for biases in either direction (the weights do not need to sum up to 1)

Can even be used as a pre-treatment on each single forecaster (works well on some data sets):

- turn it into a forecaster with predictions $\gamma_t \, f_{j,t}$
- performing on average almost as well as the best forecaster of the form $\gamma \, f_{j,t}$ for some constant $\gamma \in \mathbb{R}$

This would improve greatly the predictions if there existed, for instance, an almost constant multiplicative bias of $1/\gamma$

First study, continued

Prediction of electricity load

Benchmark and oracles

(RMSE of the ensemble forecasts and of fixed combinations thereof)

Uniform mean	Best forecaster	Best convex **p**	Best linear **u**
725	744	629	629

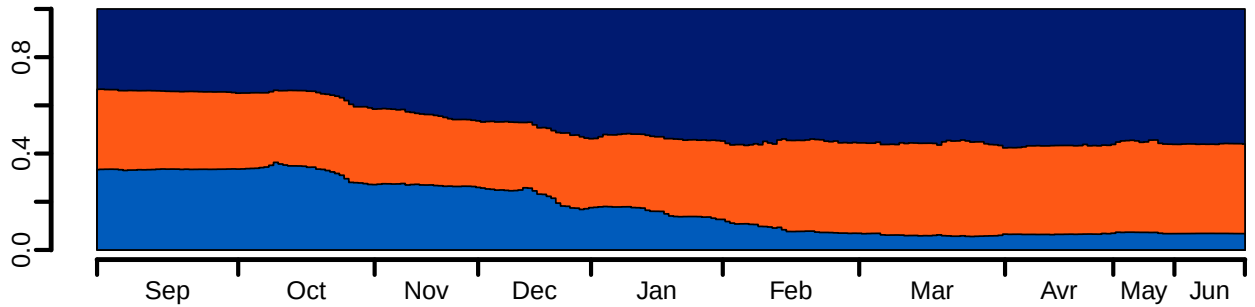
vs.

Aggregated forecasts with convex weights

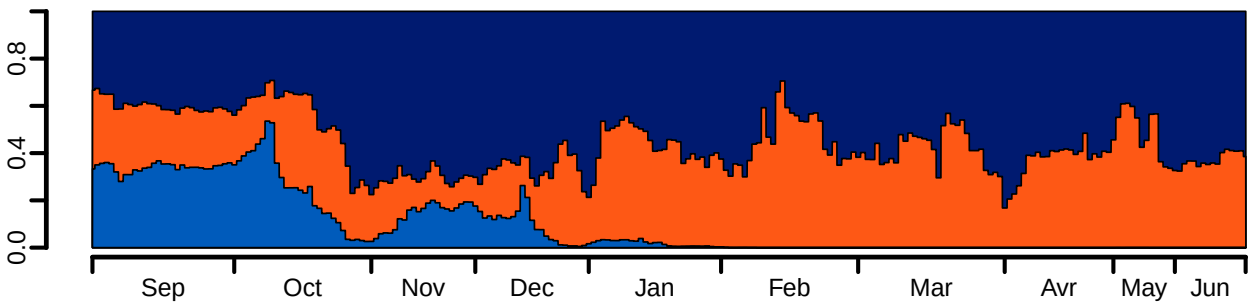
(No discount considered)

Exp. weights (best η for theory)	644
Exp. weights (best η on data)	619
Exp. weights (η_t tuned on data)	625
ML-Poly (tuned according to theory)	626

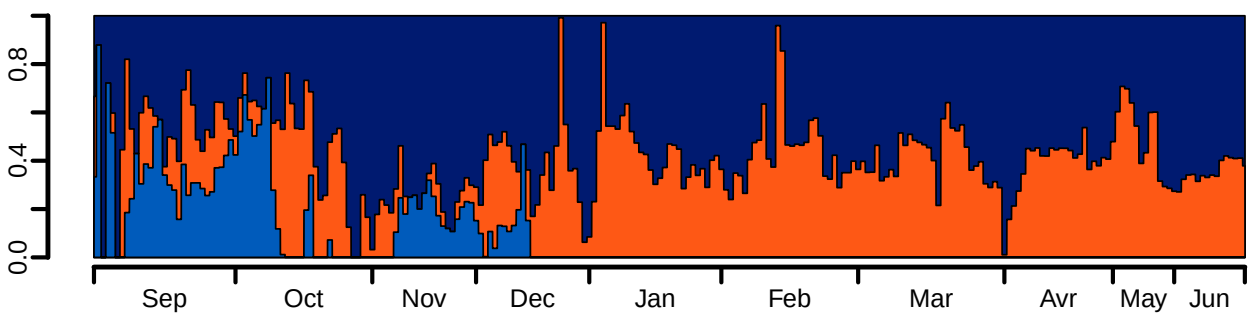
No focus on a single member! (See also the numerical performance.)



← Exp. weights (theory)



← Exp. weights (best η)



← ML-Poly (theory)

Weights change quickly and significantly over time and do not converge

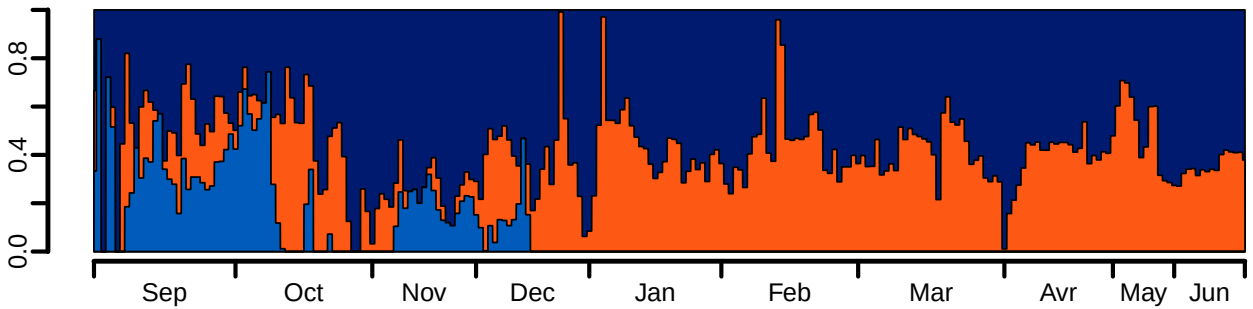
(illustrates that the performance of forecasters varies over time)

Are all forecasters useful? ... Definitely yes!

3 forecasters $\rightarrow$ only best 2

ML-Poly	626	$\rightarrow$	646
Exp. weights	625	$\rightarrow$	644

Forecasters not considered anymore can come back to life if needed



$\leftarrow$ ML-Poly

Benchmark and oracles

(RMSE of the ensemble forecasts and of fixed combinations thereof)

Uniform mean	Best forecaster	Best convex **p**	Best linear **u**
725	744	629	629

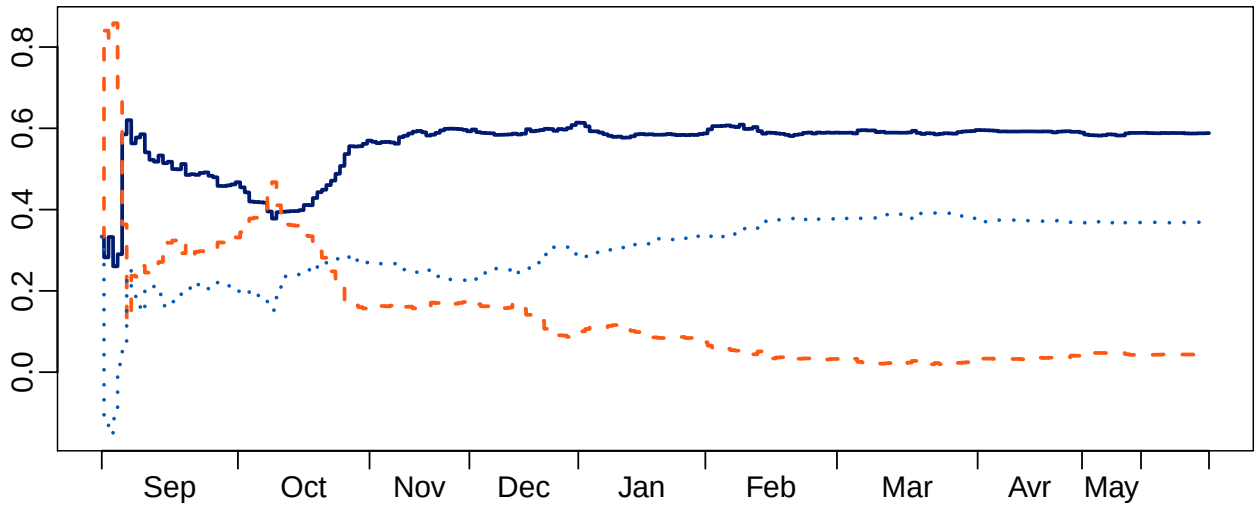
vs.

Aggregated forecasts with linear weights

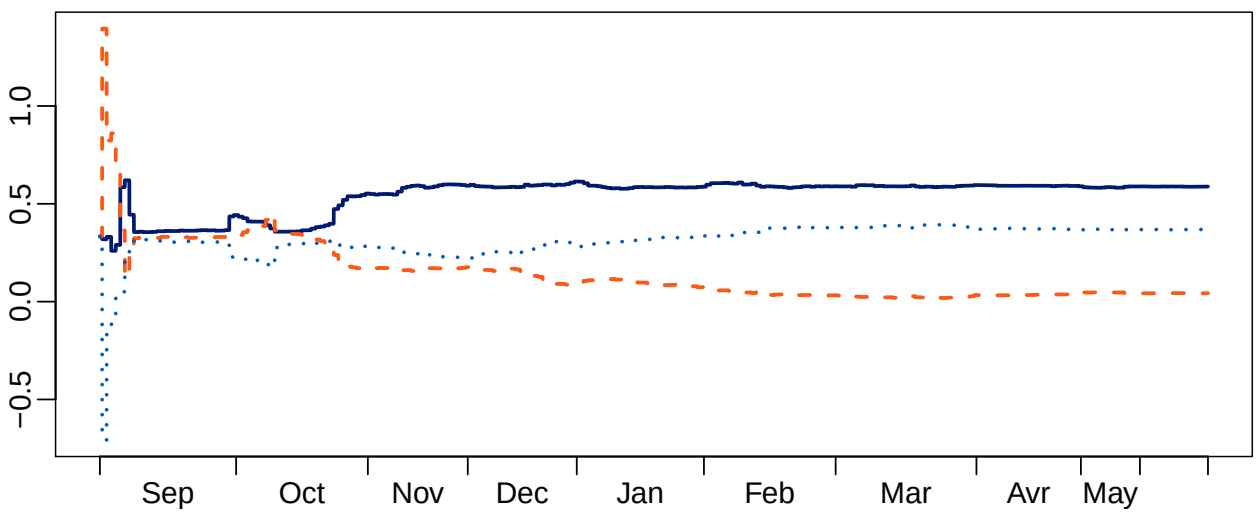
(No discount considered)

Ridge (best λ on data)	636
Ridge (λ_t tuned on data)	638

Weight vectors chosen by ridge regression



← Best λ



← λ_t on data

This was only a small glimpse into the work performed by Pierre Gaillard, Yannig Goude, Raphaël Nédellec, Côme Bissuel, and others, at EDF R&D

Other data sets studied include the forecasting of

- Slovakian demand for clients of an EDF subbranch
- GEFCom '2014 electricity price [Kaggle-ranked #1]
- GEFCom '2014 electricity load [Kaggle-ranked #1]
- Heat load of an Ukrainian co-generation plant
- Electricity demand of sub-groups of EDF clients

$\longrightarrow$ Universality of the aggregation methods!

Reference: Pierre Gaillard's Ph.D. dissertation, July 2015 (Paul Caseau Ph.D. award)

Methodological summary

Methodological summary

1. Build the N base forecasters, possibly on a training data set, and pick another data set for the evaluation, with T instances
2. Compute some benchmarks and some reference oracles
3. Evaluate our strategies when run with fixed parameters (i.e., with the best parameters in hindsight)
4. The performance of interest is actually the one of the data-driven meta-strategies

We typically expect $T \geqslant 5N$ or even $T \geqslant 10N$

Hope arises when the oracles are 10% or 20% better than the methods used so far (e.g., the best ensemble forecast when the latter is known in advance)

This usually requires the ensemble forecasters to be as different as possible!

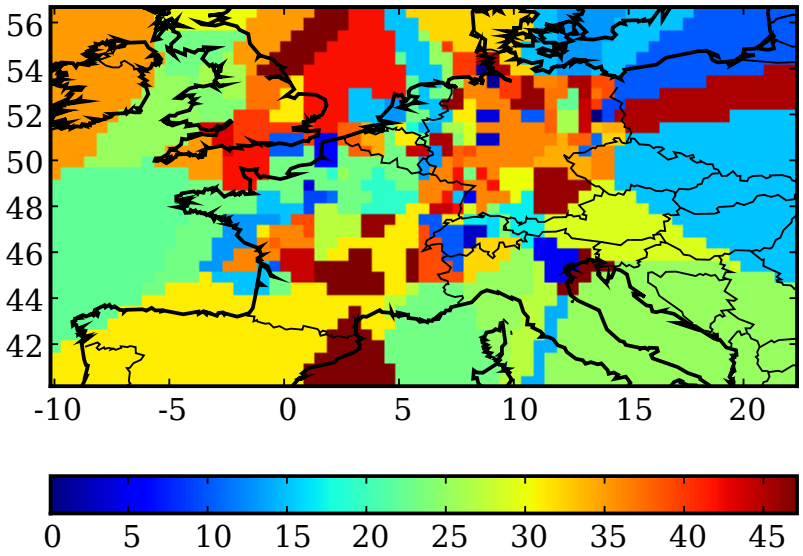
Air quality

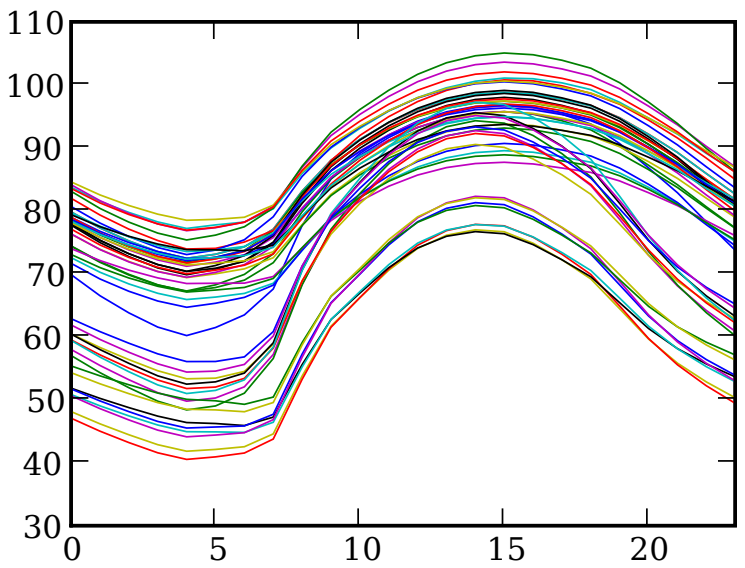
Ozone concentration

Network of 241 stations in France and Germany / summer 2001

48 forecasting black-boxes by Mallet and Sportisse [2006]

Exhibiting a strong diversity:





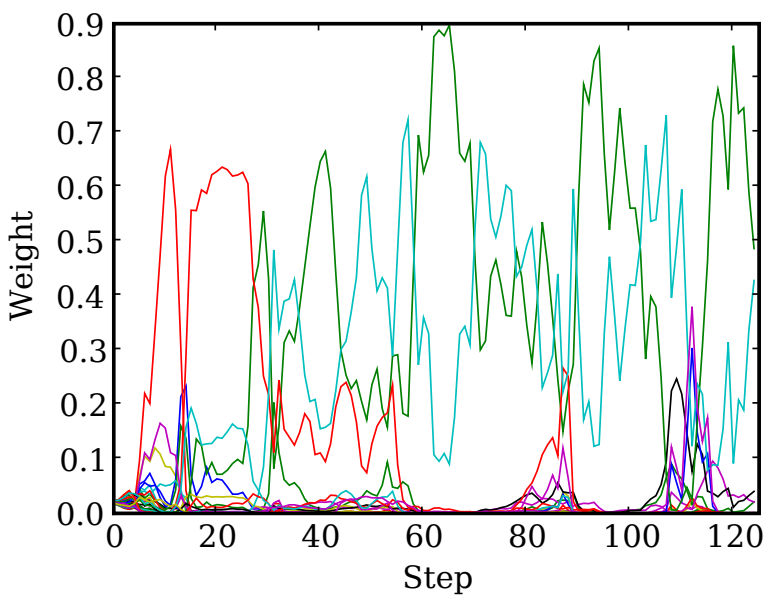
Left: Coloring of Europe according to the index of the locally best black-box

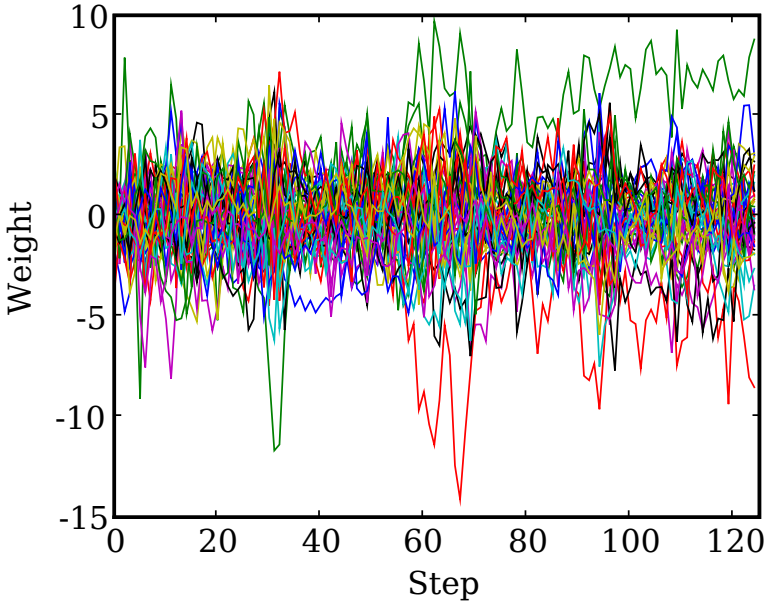
Right: Average forecasting profiles during a day (averages over time and space)

Performance in terms of RMSE

Uniform mean	Best black-box	Best convex	EWA	Ridge
24.41	22.43	21.45	21.77	20.81

Weights (y–axis) output by EWA (left) and Ridge (right) over the 120 days considered





Ref.: V. Mallet, B. Mauricette, G. Stoltz – Journal of Geophysical Research – 2009

Exchange rates

Forecasting of exchange rates on a monthly scale

April 1973 – May 2013 / Based on inflation or interest rates

Ensembles with two forecasts only (home, foreign values)

Forecasting equation on log-rates:

$$\widehat{y}_{t+1} - y_t = p_{1,t} i_t^{\text{home}} - p_{2,t} i_t^{\text{foreign}}$$

Meese and Rogoff [1983]: impossible to beat no-change
(because these fundamentals carry too little information)

Our stance: just revisit how weights are picked!
(do not change the fundamentals, change the aggregation layer)

Performance in terms of RMSE

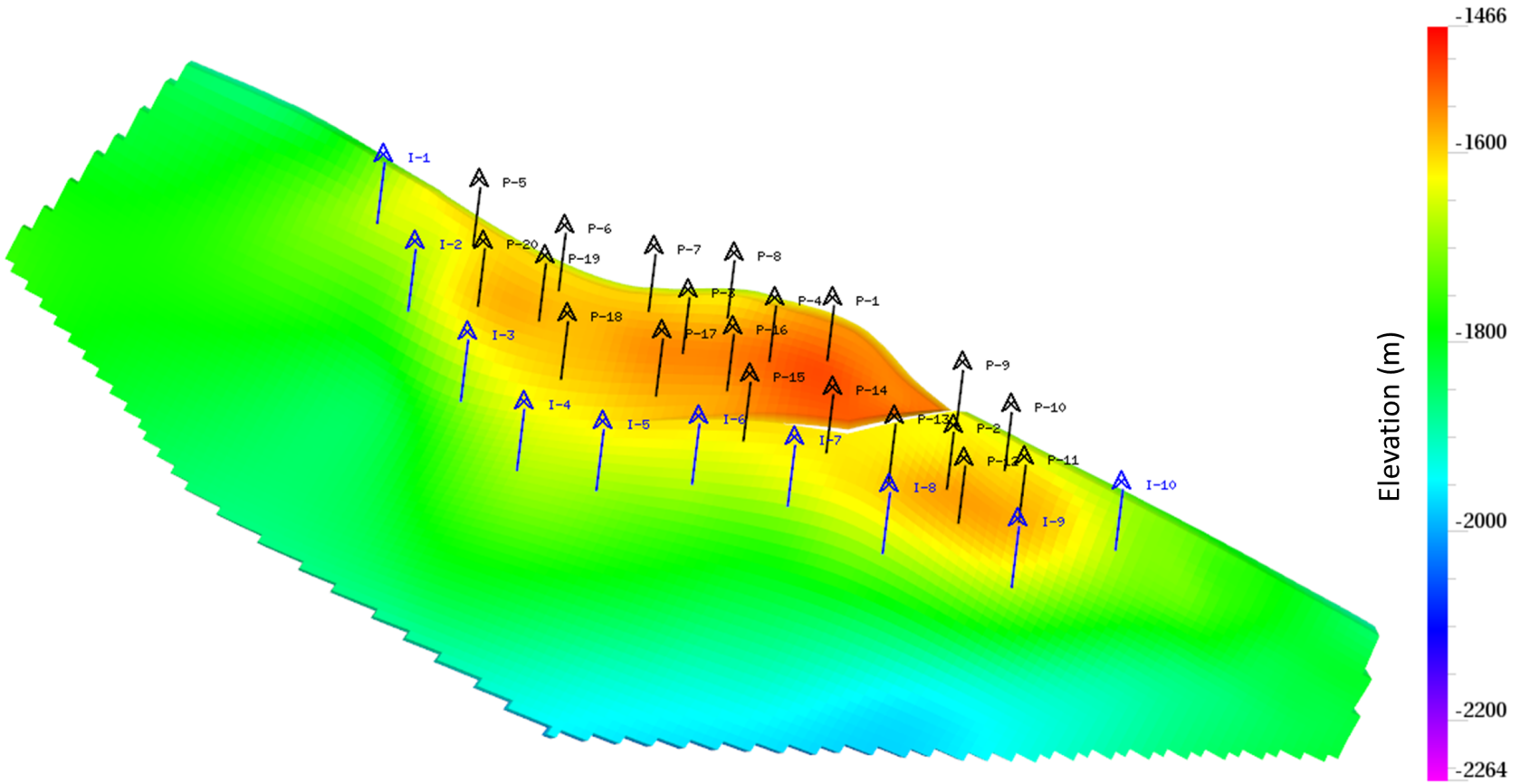
Currency pair	No change RMSE x 100	OLS Theil ratio	Ridge Theil ratio	EWA Theil ratio
Inflation fundamentals				
USD/GBP	2.4410	1.0005	**0.9735**	**0.9887**
JPY/USD	2.7039	1.0173	**0.9816**	**0.9862**
CHF/USD	2.7960	1.0116	**0.9848**	**0.9882**
CAD/USD	1.5091	**0.9867**	**0.9883**	**0.9964**
Interest rates fundamentals				
USD/GBP	2.4410	1.0241	**0.9688**	**0.9809**
JPY/USD	2.7039	1.0054	**0.9717**	**0.9809**
CHF/USD	2.8195	1.0135	**0.9790**	**0.9797**
CAD/USD	1.5091	1.0086	**0.9965**	**0.9986**

Ref.: C. Amat, T. Michalski, G. Stoltz – Journal of International Money and Finance

Oil and gas production: mid-term predictions

The Brugge field (synthetic but realistic data)

Reference: Peters et al. (2010), SPE



Can be decomposed into millions of grid blocks, in which petrophysical properties are unknown (= a model)

Classical approach for sequential forecasting:

Fluid-flow equations (and simulators) relate

- the production characteristics of the field (pressure, oil and water rates, etc.) over time
- to the model (to the petrophysical properties)

One may thus learn the model based on

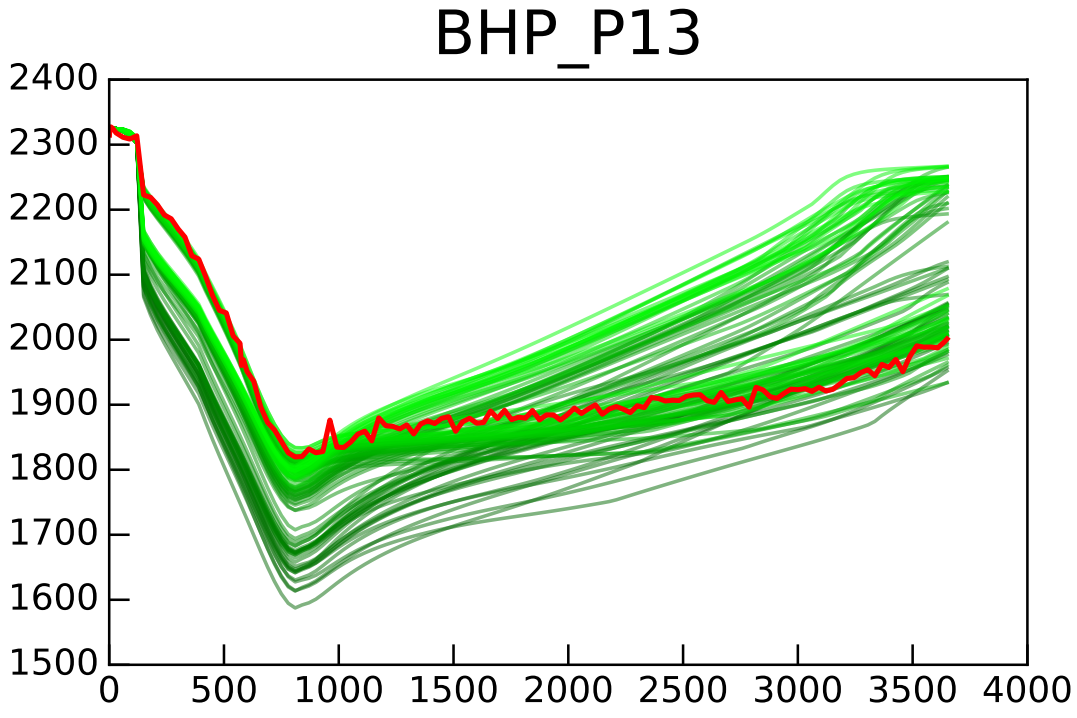
- estimates of the petrophysical properties (using some past measurements)
- constraints of closeness of their associated production characteristics to those actually observed over time

This is computationally heavy:

At each time step, many fluid-flow simulations must be performed (many models are tested)

Our approach:

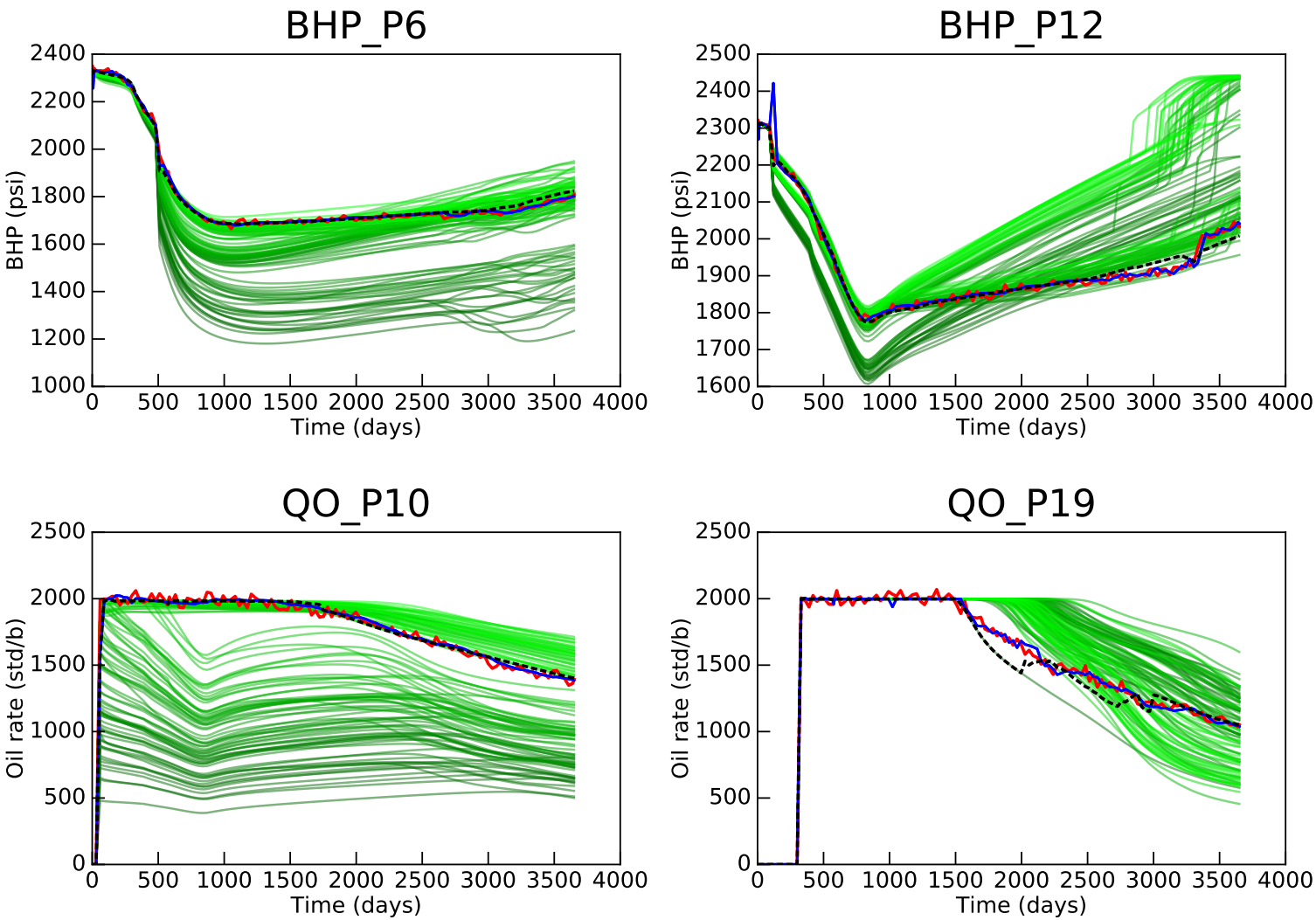
The Brugge data set comes with 104 geological models
(their petrophysical properties were chosen in some way)



BHP = pressure at the bottom of the hole; P = producer well; number 13

We will sequentially combine their outputs

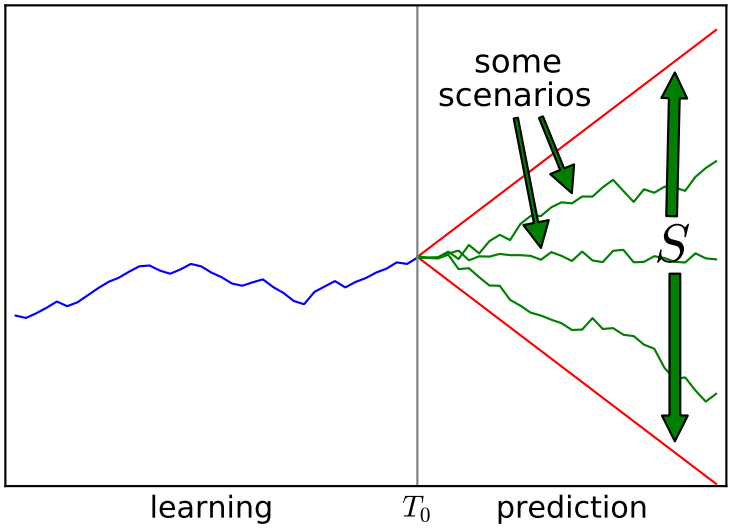
Ex.: model outputs / observations / predictions (EWA + Ridge)

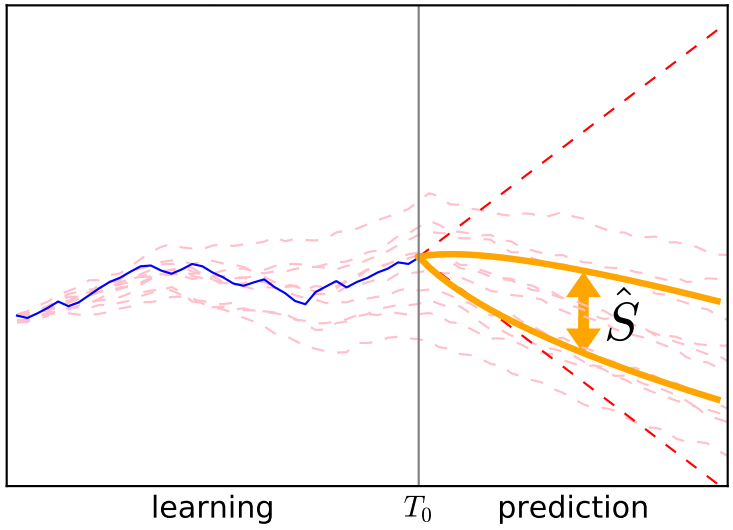


BHP = pressure at the bottom of the hole; QW = water flow rate; QO = oil flow rate

P = producer well; I = injection well; the numbers index the wells

Our individual-sequences approach for interval forecasts





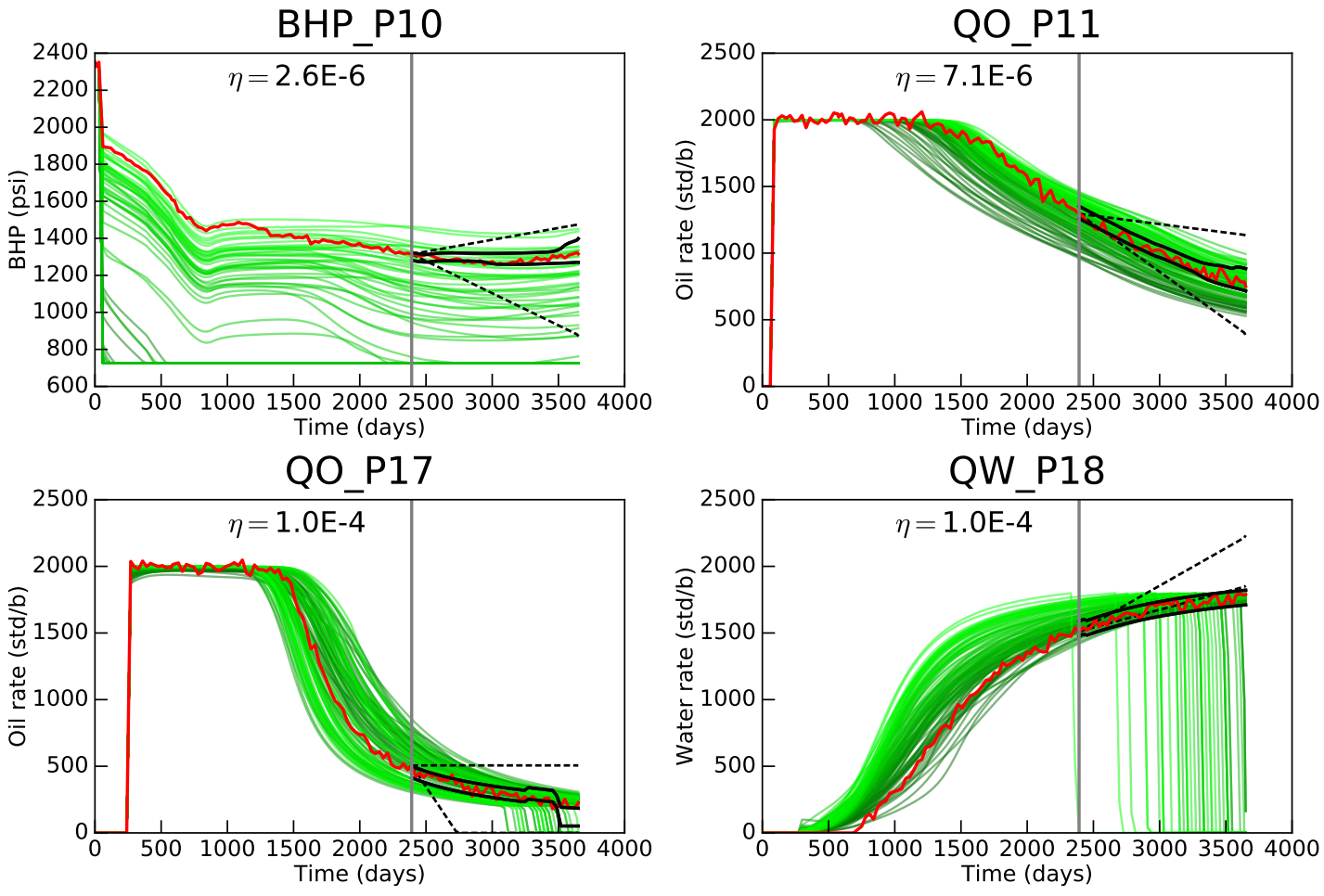
On the first part of the data set ("learning" or training part):

– Use the algorithms as explained above

On the second part of the data set ("prediction" part):

– Consider the set S of all plausible continuations ("scenarios")
– Compute all associated aggregated forecasts, using the model forecasts
– The interval forecasts $\widehat{S}$ are the convex hulls of the aggregates (with enlargement for noise level)

Interval forecasts with EWA



BHP = pressure at the bottom of the hole; QW = water flow rate; QO = oil flow rate

P = producer well; I = injection well; the numbers index the wells

Projet avec Cdiscount

Projet avec Cdiscount

374 K-produits, organisés en hiérarchie (= le verrou de ce sujet)
Prévoir volumes de ventes par semaine, 6 semaines à l'avance

Méthodes de lissage exponentiel, avec paramètre α

Au lieu de chercher à régler α
$\rightarrow$ agrégation de prévisions issues de $\alpha_1, \ldots, \alpha_K$

Pour prévoir la hiérarchie mais pas les produits individuels
Méthode plus fine pour déduire la répartition en produits d'une sous-sous-famille

Performance: déjà 15% meilleure qu'outil maison

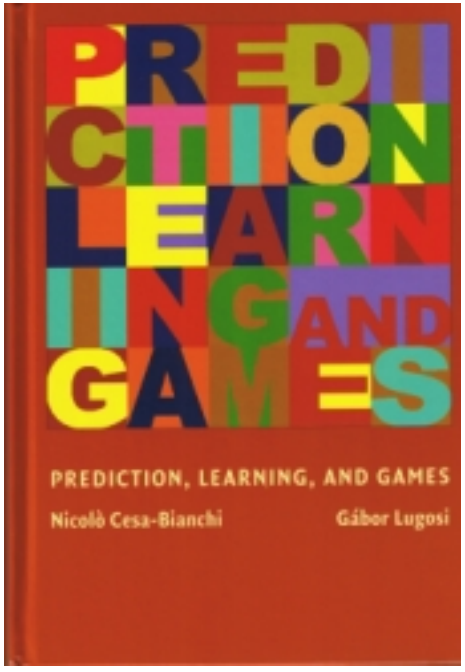
Contrat en cours (= chapitre de thèse de Malo Huard, LMO)

Bibliographie

Ouvrage de référence pour la théorie

(mais pas pour la pratique!)

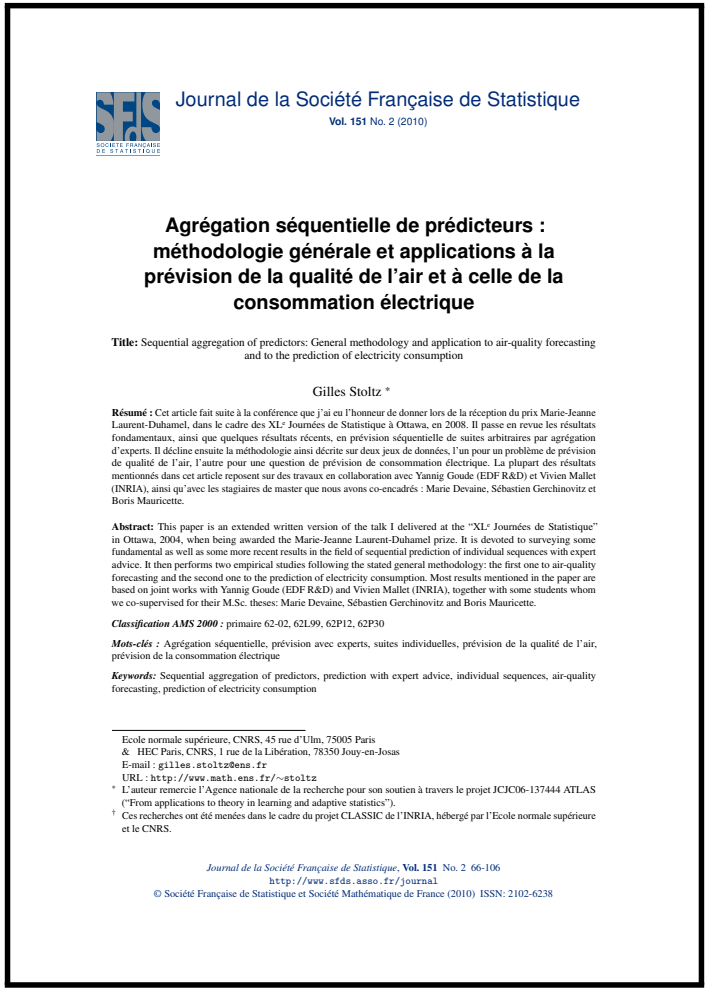
Le petit livre rouge



Prediction, Learning, and Games

Nicolò Cesa-Bianchi and Gábor Lugosi

Article de référence pour la pratique



Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique

Title: Sequential aggregation of predictors: General methodology and application to air-quality forecasting and to the prediction of electricity consumption

Gilles Stoltz *

Résumé : Cet article fait suite à la conférence que j'ai eu l'honneur de donner lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des XL^e Journées de Statistique à Ottawa, en 2008. Il passe en revue les résultats fondamentaux, ainsi que quelques résultats récents, en prévision séquentielle de suites arbitraires par agrégation d'experts. Il décline ensuite la méthodologie ainsi décrite sur deux jeux de données, l'un pour un problème de prévision de qualité de l'air, l'autre pour une question de prévision de consommation électrique. La plupart des résultats mentionnés dans cet article reposent sur des travaux en collaboration avec Yannig Goude (EDF R&D) et Vivien Mallet (INRIA), ainsi qu'avec les stagiaires de master que nous avons co-encadrés : Marie Devaine, Sébastien Gerchinovitz et Boris Mauricette.

Abstract: This paper is an extended written version of the talk I delivered at the "XL^e Journées de Statistique" in Ottawa, 2004, when being awarded the Marie-Jeanne Laurent-Duhamel prize. It is devoted to surveying some fundamental as well as some more recent results in the field of sequential prediction of individual sequences with expert advice. It then performs two empirical studies following the stated general methodology: the first one to air-quality forecasting and the second one to the prediction of electricity consumption. Most results mentioned in the paper are based on joint works with Yannig Goude (EDF R&D) and Vivien Mallet (INRIA), together with some students whom we co-supervised for their M.Sc. theses: Marie Devaine, Sébastien Gerchinovitz and Boris Mauricette.

Classification AMS 2000 : primaire 62-02, 62L99, 62P12, 62P30

Mots-clés : Agrégation séquentielle, prévision avec experts, suites individuelles, prévision de la qualité de l'air, prévision de la consommation électrique

Keywords: Sequential aggregation of predictors, prediction with expert advice, individual sequences, air-quality forecasting, prediction of electricity consumption

Ecole normale supérieure, CNRS, 45 rue d'Ulm, 75005 Paris
& HEC Paris, CNRS, 1 rue de la Libération, 78350 Jouy-en-Josas
E-mail : gilles.stoltz@ens.fr
URL : http://www.math.ens.fr/~stoltz

* L'auteur remercie l'Agence nationale de la recherche pour son soutien à travers le projet JCJC06-137444 ATLAS ("From applications to theory in learning and adaptive statistics").
† Ces recherches ont été menées dans le cadre du projet CLASSIC de l'INRIA, hébergé par l'Ecole normale supérieure et le CNRS.

Journal de la Société Française de Statistique, **Vol. 151** No. 2 66-106
http://www.sfds.asso.fr/journal
 ISSN: 2102-6238

En français, écrit par votre serviteur, mais qui date un peu (2010)

Présentation du sujet CIFRE avec la SNCF