Comment tirer parti de l'embarras du choix

face à plusieurs modèles de prévision concurrents

Gilles Stoltz

CNRS — École normale supérieure — HEC Paris

Qu'est-ce que la statistique ?

Deux branches ?

- Statistiques descriptives (bien représenter et résumer des données, mieux connaître l'échantillon)
- Statistique inférentielle (mieux connaître une population ou comprendre un phénomène à partir de ces données)

Dans les deux cas : connaître, comprendre, ...
Mais dans quel but ? Prévoir, prendre des décisions.

Prévoir, ce n'est pas prédire : ce qui distingue la statistique de la voyance, c'est qu'elle s'appuie sur des données.

Bon. Traiter des données de sondage, c'est facile à expliquer.

Mais expliquer comment construire des modèles de prévision statistique de la consommation électrique ?

C'est toute une industrie. Chacun y va de son (petit ou gros) modèle. Tous des usines à gaz. Et c'est pourquoi...

Cadre de cet exposé

Le mot-clé sera : «agrégation» (par opposition à «sélection»)

Un statisticien accepte la mission de prévoir une suite $y_1, y_2, \ldots$ d'observations vivant dans un intervalle $[0, B]$.

Ses prévisions $\widehat{y}_1, \widehat{y}_2, \ldots$ sont formées dans $[0, B]$ également, car B est supposé connu.

Les observations et prévisions (1) sont effectuées de manière séquentielle et (2) ne reposent sur aucun modèle stochastique.

(1) signifie qu'à chaque échéance, la prévision $\widehat{y}_t$ pour y_t est déterminée

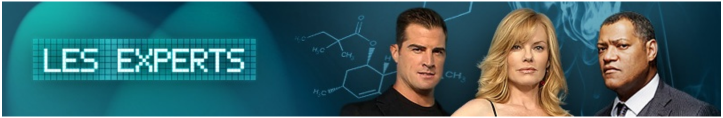
- sur le seul fondement du passé, $y_1^{t-1} = \big(y_1, \ldots, y_{t-1}\big)$,
- et avant que la vraie valeur y_t ne soit révélée au grand jour.

(2) indique qu'on ne se placera pas dans un cas où les y_t sont les réalisations d'un processus stochastique Y_t sous-jacent.

Tant mieux pour vous qui n'avez pas peut-être pas étudié la théorie des probabilités !

Pour que le problème ait un sens dans un cadre aussi général, on suppose avoir accès à des experts, en nombre fini N.

C'est le petit nom des fameux «modèles de prévisions concurrents».



A chaque échéance t, l'expert $j \in \{1, \ldots, N\}$ forme une prévision

$$f_{j,t} = f_{j,t}\left(y_1^{t-1}\right) \in [0, B]$$

Le statisticien fonde maintenant ses propres prévisions $\widehat{y}_t$ sur

- les observations passées : $y_1^{t-1} = (y_1, \ldots, y_{t-1})$,
- et sur les conseils passés et présents des experts : $f_{j,s}$ pour $s \in \{1, \ldots, t\}$ et $j \in \{1, \ldots, N\}$.

Méthode : combinaisons convexes à poids variables

Le statisticien va former à chaque tour des prévisions (dites agrégées) de la forme

$$\widehat{y}_t = \sum_{j=1}^{N} p_{j,t}\, f_{j,t} \qquad \text{où} \qquad \begin{cases} \forall j, \quad p_{j,t} \geqslant 0, \\ p_{1,t} + \ldots + p_{N,t} = 1 \end{cases}$$

C'est ce que l'on appelle une combinaison convexe selon $\mathbf{p}_t = \big(p_{1,t}, \ldots, p_{N,t}\big)$.

L'objectif du statisticien est de prévoir, en moyenne, presqu'aussi bien que la meilleure combinaison convexe constante $\mathbf{q}$ des prévisions des experts.

Mais il s'agit tout d'abord de quantifier la précision d'une (suite de) prévision(s)...

On introduit une fonction de perte $\ell : [0,B] \times [0,B] \to \mathbb{R}_+$.

Elle mesure l'écart entre une prévision $\widehat{y}$ et une observation y.

Des choix typiques sont

- la perte quadratique $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$;
- la perte absolue $\ell(\widehat{y}, y) = |\widehat{y} - y|$;
- le pourcentage d'erreur $\ell(\widehat{y}, y) = |\widehat{y} - y| / |y|$.

On retiendra la perte quadratique pour la suite (par exemple parce qu'elle permet de pénaliser davantage les erreurs de prévisions importantes).

On définit les perte cumulée et moyenne du statisticien comme

$$\widehat{L}_T = \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \, f_{j,t} - y_t \right)^2 \qquad \text{et} \qquad \widehat{\mathrm{EQM}}_T = \sqrt{\frac{\widehat{L}_T}{T}}$$

On a défini les perte cumulée et moyenne du statisticien :

$$\widehat{L}_T = \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t} - y_t \right)^2 \qquad \text{et} \qquad \widehat{\mathrm{EQM}}_T = \sqrt{\frac{\widehat{L}_T}{T}}$$

Pour une combinaison convexe constante $\mathbf{q} = (q_1, \ldots, q_N)$:

$$L_T(\mathbf{q}) = \sum_{t=1}^{T} \left(\sum_{j=1}^{N} q_j\, f_{j,t} - y_t \right)^2 \qquad \text{et} \qquad \mathrm{EQM}_T(\mathbf{q}) = \sqrt{\frac{L_T(\mathbf{q})}{T}}$$

EQM signifie «écart quadratique moyen».

Le regret est la différence entre les deux quantités cumulées

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

Comme $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$, il vient également

$$\widehat{\mathrm{EQM}}_T \leqslant \mathrm{EQM}_T(\mathbf{q}) + \sqrt{R_T/T}$$

Morale : il faut assurer que R_T/T tende vers 0 (ou soit négatif).

On rappelle que le regret R_T est défini comme la différence

$$\widehat{L}_T - \min_{\mathbf{q}} L_T(\mathbf{q}) = \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \, f_{j,t} - y_t \right)^2 - \min_{\mathbf{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} q_j \, f_{j,t} - y_t \right)^2$$

On veut construire des stratégies de prévision telles que le regret moyen converge vers 0, de manière uniforme :

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

où le supremum porte sur toutes les suites possibles d'observations et de prévisions des experts. (Pas juste sur la plupart d'entre elles !)

Remarques :

- On parle de prévision de suites individuelles ou d'agrégation robuste.
- La meilleure combinaison convexe $\mathbf{q}^\star$ ne peut être déterminée que rétrospectivement tandis que le statisticien est assujetti à une contrainte de prévision séquentielle.

Ce cadre admet une interprétation méta-statistique :

- chaque expert peut correspondre à une méthode de prévision statistique fondamentale ; des méthodes concurrentes peuvent être considérées ;
- on combine ensuite de manière robuste (et déterministe) ces prévisions fondamentales.

Or, la perte cumulée et l'écart quadratique moyen du statisticien peuvent être décomposés selon

$$\widehat{L}_T = \min_{\mathbf{q}} L_T(\mathbf{q}) + R_T \qquad \text{et} \qquad \widehat{\mathrm{EQM}}_T \leqslant \mathrm{EQM}_T(\mathbf{q}) + \sqrt{R_T/T}$$

On peut donc interpréter

- le terme de performance de la meilleure combinaison convexe des experts comme une erreur d'approximation,
- le terme de regret comme mesurant une difficulté d'estimation séquentielle.

Une stratégie simple

Que nous allons étudier de bout en bout !

Rappel de l'objectif de prévision :

Il s'agit donc de majorer uniformément le regret face à toutes les combinaisons convexes $\mathbf{q}$,

$$\sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \, f_{j,t} - y_t \right)^2 - \sum_{t=1}^{T} \left(\sum_{j=1}^{N} q_j \, f_{j,t} - y_t \right)^2$$

Inégalité utile : pour tous $u, v \in [0, B]$,

$$u^2 - v^2 \leqslant 2u \, (u - v)$$

Preuve : $u^2 - v^2 = (u+v)\,(u-v)$ et distinguer selon $u \geqslant v$ ou $u \leqslant v$ pour savoir s'il faut majorer ou minorer $u+v$ par $2u$.

On peut généraliser cette inégalité (dite «des pentes») à d'autres fonctions de pertes («convexes»).

Il s'agit de majorer uniformément le regret face à toutes les combinaisons convexes $\mathbf{q}$:

$$
\begin{aligned}
& \max_{\mathbf{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t}\, f_{j,t} - y_t \right)^2 - \sum_{t=1}^{T} \left(\sum_{j=1}^{N} q_j\, f_{j,t} - y_t \right)^2 \\
\leqslant \quad & \max_{\mathbf{q}} \sum_{t=1}^{T} 2 \left(\sum_{k=1}^{N} p_{k,t} f_{k,t} - y_t \right) \left(\sum_{j=1}^{N} p_{j,t} f_{j,t} - \sum_{j=1}^{N} q_j f_{j,t} \right) \\
= \quad & \max_{\mathbf{q}} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \sum_{j=1}^{N} q_j \widetilde{\ell}_{j,t} \right) \\
= \quad & \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \widetilde{\ell}_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \widetilde{\ell}_{i,t}
\end{aligned}
$$

où l'on a défini

$$
\widetilde{\ell}_{j,t} = 2 \left(\sum_{k=1}^{N} p_{k,t} f_{k,t} - y_t \right) f_{j,t} \quad \in \left[-2B^2,\, 2B^2 \right]
$$

Via la considération des pseudo-pertes (signées) $\widetilde{\ell}_{j,t}$, on vient de se ramener au cadre générique suivant d'apprentissage séquentiel.

A chaque tour $t = 1, 2, \ldots,$

- le statisticien choisit un vecteur de combinaison convexe $\mu_t = \bigl(\mu_{1,t}, \ldots, \mu_{N,t}\bigr)$;
- l'environnement détermine un vecteur de pertes $\ell_t = \bigl(\ell_{1,t}, \ldots, \ell_{N,t}\bigr)$, tel que $\ell_{j,t} \in [m, M]$ pour tout j ;
- le statisticien observe ℓ_t.

L'objectif est de contrôler uniformément le regret générique

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} \mu_{j,t} \ell_{j,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \ell_{i,t}$$

Pour tout $j \in \{1, \ldots, N\}$, on prend $\mu_{j,1} = 1/N$ et pour $t \geqslant 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

C'est la stratégie de pondération par poids exponentiels des pertes passées des experts (de vitesse d'apprentissage η).

Lemme. On fixe deux réels $m \leqslant M$.

Pour tout $\eta > 0$ et pour toute suite individuelle d'éléments $\ell_{j,t} \in [m, M]$, où $j \in \{1, \ldots, N\}$ et $t \in \{1, \ldots, T\}$,

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

Références : Vovk '90 et Littlestone et Warmuth '94

Eléments de preuve

1. Borne sous-optimale

$$\frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{2} T$$

à partir de $e^{-x} \leqslant 1 - x + x^2/2$, inégalité valable pour tout $x \geqslant 0$

2. Borne avec le facteur $1/8$ par le lemme de Hoeffding : pour toute variable aléatoire X à valeurs dans $[m, M]$, pour tout $s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{sX}\right] \leqslant s\,\mathbb{E}[X] + \frac{s^2}{8}(M-m)^2$$

Le diable est dans les détails...

On rappelle que $[m, M]$ est l'intervalle des pertes.

On peut optimiser la borne sur le regret en η :

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

pour le choix de

$$\eta^\star = \frac{1}{M-m}\sqrt{\frac{8 \ln N}{T}}$$

Ce choix dépend de M et m, qui sont connus à l'avance, mais également de T, qui est destiné à tendre vers l'infini.

Or, aucune valeur fixée de $\eta > 0$ n'assure que $R_T = o(T)$.

On n'a donc pas encore de stratégie totalement séquentielle.

Les solutions possibles sont, d'une part, de recourir à des re-démarrages à froid périodiques.

D'autre part, de faire varier les vitesses d'apprentissage au cours du temps, en fonction du passé : pour $t \geqslant 2$,

$$\mu_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

Un bon choix des η_t permet alors d'obtenir la borne de regret

$$R_T \leqslant (M - m)\sqrt{T \ln N}$$

Le prix à payer pour l'adaptation à T est donc un facteur multiplicatif $\sqrt{2}$.

Réferences : Auer, Cesa-Bianchi et Gentile '02

Le réglage théorique proposé pour les η_t conduit à de mauvaises performances pratiques.

La stratégie $\mathcal{E}_\eta$ de pondération par poids exponentiels des pertes, de vitesse d'apprentissage η, choisit la combinaison convexe $\mu_t(\eta)$:

$$\mu_{j,t}(\eta) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{k,s}\right)}$$

La perte cumulée de cette stratégie vaut $\widehat{L}_T(\eta) = \displaystyle\sum_{t=1}^{T} \sum_{j=1}^{N} \mu_{j,t}(\eta)\, \ell_{j,t}$

A partir de la famille des $\mathcal{E}_\eta$, on construit la méta-stratégie calibrée sur les données qui, à chaque échéance $t \geqslant 2$, utilise

$$\mu_t(\eta_t) \qquad \text{où} \qquad \eta_t \in \mathop{\arg\min}_{\eta > 0} \widehat{L}_{t-1}(\eta)$$

Problème à résoudre

Exhiber une borne théorique sur le regret de la méta-stratégie !

Référence : Idée de Vivien Mallet

Contre un environnement changeant la performance de toute combinaison convexe $\mathbf{p}$ fixée peut être mauvaise.

On voudrait plutôt pouvoir se comparer à des suites de combinaisons convexes de la forme

$$\underline{\mathbf{p}} = \left(\mathbf{p}^1, \ldots, \mathbf{p}^1,\ \mathbf{p}^2, \ldots, \mathbf{p}^2,\ \ldots,\ \mathbf{p}^{m+1}, \ldots, \mathbf{p}^{m+1}\right),$$

où parmi les T éléments, au plus m sauts peuvent survenir.

Il existe des stratégies d'agrégation séquentielle dont le regret moyen face à ce genre de suites est asymptotiquement négatif ou nul.

Elles sont dites par redistribution au vu de leur manière de procéder (mise à jour exponentielle des poids suivie d'un mélange avec la loi uniforme).

Elles dépendent de deux paramètres η et α.

Deux domaines d'application pratique

– Prévision de la qualité de l'air

– Prévision de la consommation électrique

Plan des études empiriques

(1) Construire des experts (sur données historiques, et utiliser un autre jeu de données pour les évaluer ainsi que nos méthodes)

(2) Evaluer leurs performances rétrospectives par calculs d'oracles : meilleur expert, meilleure combinaison convexe, meilleure combinaison linéaire

(3) Tabuler les performances de nos stratégies en fonction de valeurs fixées de leurs paramètres (η et α)

(4) Voir ce que leurs versions opérationnelles auraient valu (les méta-stratégies calibrées, qui choisissent séquentiellement les paramètres au vu des données)

Première étude : Prévision de la qualité de l'air

Date de début : Septembre 2005

Partenaire académique : Vivien Mallet, INRIA, équipe CLIME

Partenaire industriel : Edouard Debry, INERIS (Institut National de l'EnviRonnement Industriel et des RisqueS)

Stagiaires de M2 contributeurs :

- Boris Mauricette (6 mois en 2007)
- Sébastien Gerchinovitz (5 mois en 2008)
- Karim Drifi (4 mois en 2009)
- Paul Baudin (4 mois en 2012)

Publications : dans *Journal of Geophysical Research*

Quelques caractéristiques de l'un des jeux de données étudié :

- Eté 2001, période de 126 jours, prévision à 24h
- 241 stations en France et en Allemagne
- Concentrations typiques entre 40 $\mu g\,m^{-3}$ et 150 $\mu g\,m^{-3}$, seuils d'alerte à 180 $\mu g\,m^{-3}$ et 240 $\mu g\,m^{-3}$
- 48 experts, construits par Mallet et Sportisse '06, en faisant varier la formulation physico-chimique, le jeu de données d'entrée et le schéma de résolution numérique des EDPs

On mesure la qualité des prévisions avec une perte quadratique et on reportera des écarts quadratiques moyens (EQM).

Les stations du réseau sont indexées par $\mathcal{S}$.

Chaque modèle $j = 1, \ldots, 48$ prévoit $f_{j,t}^s$ pour le pic d'ozone à la station s et au jour t, une valeur à comparer à l'observation y_t^s.

Le statisticien choisit à chaque tour une unique combinaison convexe $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$ pour agréger les experts à toutes les stations ; on obtient ainsi des champs de prévision.

La perte associée à la combinaison convexe $\mathbf{p}_t$ au tour t est

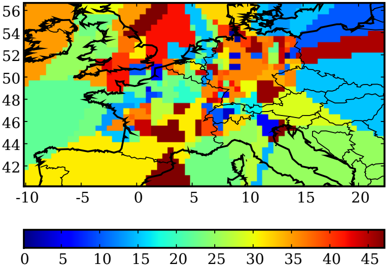
$$\ell_t(\mathbf{p}_t) = \sum_{s \in \mathcal{S}_t} \left(\sum_{j=1}^{48} p_{j,t}\, f_{j,t}^s - y_t^s \right)^2$$

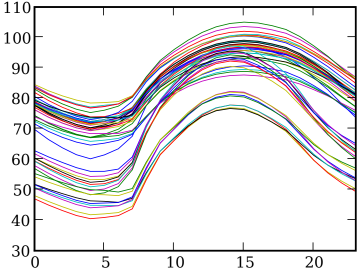
où $\mathcal{S}_t$ est le sous-ensemble des stations actives au jour t.

L'écart quadratique moyen vaut alors $\widehat{\mathrm{EQM}}_T = \sqrt{\dfrac{\sum_{t=1}^T \ell_t(\mathbf{p}_t)}{\sum_{t=1}^T |\mathcal{S}_t|}}$

Gauche : Plusieurs experts sont bons et utiles.

Droite : Leurs profils de prévision sont fort différents (les experts ne sont pas une armée de clones !).





Gauche : Coloriage de l'Europe en fonction de l'index du meilleur expert local

Droite : Profils de prévision moyens (moyennes en temps et en espace)

EQM / Performances des experts

Moyenne experts	Meilleur expert	Meilleur **p**
24.41	22.43	21.45

EQM / Performance de la pondération par poids exponentiels
(avec meilleurs paramètres rétrospectifs)

Version originelle	Version fenêtrée	Version escomptée
21.47	21.37	21.31

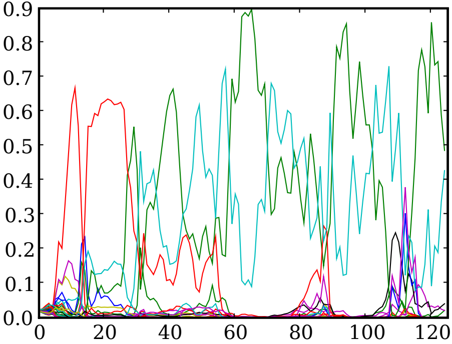
La version fenêtrée de fenêtre H utilise uniquement les pertes des H derniers tours.

La version escomptée accorde plus d'importance aux pertes les plus récentes (mais considère encore toutes les pertes).

Méta-stratégie calibrée (version originelle) : EQM de 21.77

Les méthodes d'agrégation séquentielle ne se concentrent pas sur un seul expert.

Les poids attribués aux modèles peuvent changer rapidement et de manière significative au cours du temps ; cela illustre la variabilité en temps des performances des experts.



Combinaisons convexes produites au cours du temps par la stratégie de pondération par poids exponentiels.

Deuxième étude : Prévision de la consommation électrique

Date de début : Mars 2009

Partenaire industriel : Yannig Goude, EDF R&D

Stagiaires de M2 contributeurs :

- Marie Devaine (5 mois en 2009)
- Pierre Gaillard (5 mois en 2011)

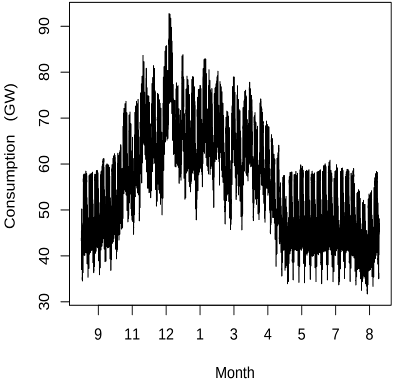
Publication : dans *Machine Learning*

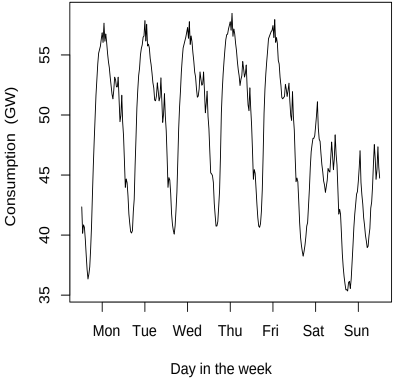
On dispose d'experts spécialisés, ne procurant des prévisions que dans certaines circonstances : par exemple, uniquement en semaine ou uniquement le week-end.

Il faut étendre les définitions et stratégies générales (détails omis).

Sur notre jeu de données de consommation globale en France, à pas demi-horaire,

- 24 experts sont disponibles, issus de 3 familles, chacune associée à un type de modèle (paramétrique, semi-paramétrique, non-paramétrique) ;
- il existe une contrainte opérationnelle de prévision à 24h (soit 48 échéances simultanées).





Consommation électrique en France

- Année 2007–08 (gauche)
- Semaine typique hors hiver (droite)

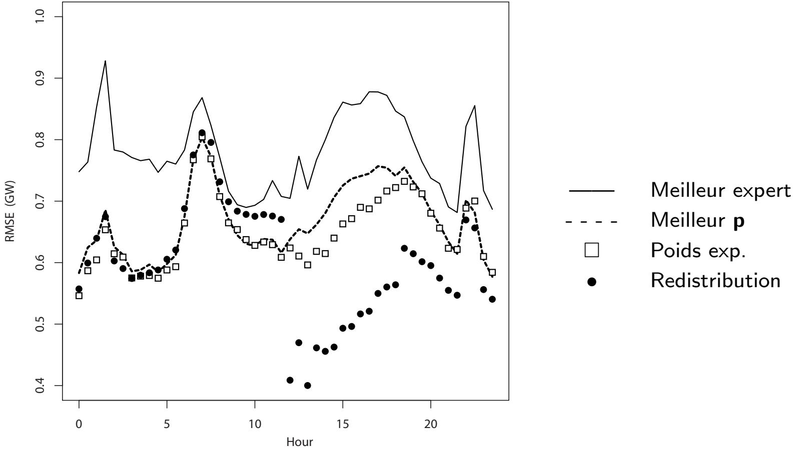
Quelques ordres de grandeur pour mieux saisir le problème.

Echéances	Toutes les 30 minutes
Nombre de jours D	320
Nombre d'échéances T	$15\,360$ $(= 320 \times 48)$
Nombre d'experts N	24 $(= 15 + 8 + 1)$
Médiane des y_t	$56\,330$ MW
Borne B sur les y_t	$92\,760$ MW

On précise là encore des EQM.

Meilleur expert	Moyenne experts	Meilleur **p**
782	724	658
Poids exp.	Meilleur param.	Opérat.
	629	637

Sauts	$m = T - 1 = 15\,359$	$m = 200$	$m = 50$
	223	414	534
Redistribution		Meilleur param.	Opérat.
		5999	629



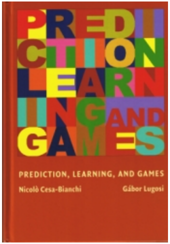
EQMs moyen (en GW / pas en MW) en fonction de la demi-heure

Les EQMs des deux stratégies étaient similaires mais leurs comportements diffèrent assez fortement au cours d'une journée.

Bibliographie

Juste au cas où vous n'en auriez pas déjà assez...

Le petit livre rouge



Prediction, Learning, and Games

Nicolò Cesa-Bianchi et Gábor Lugosi

J'ai publié un article de survol (en français !) dans le Journal de la Société Française de Statistique

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