Agrégation séquentielle d'experts

avec application à la prévision de consommation électrique

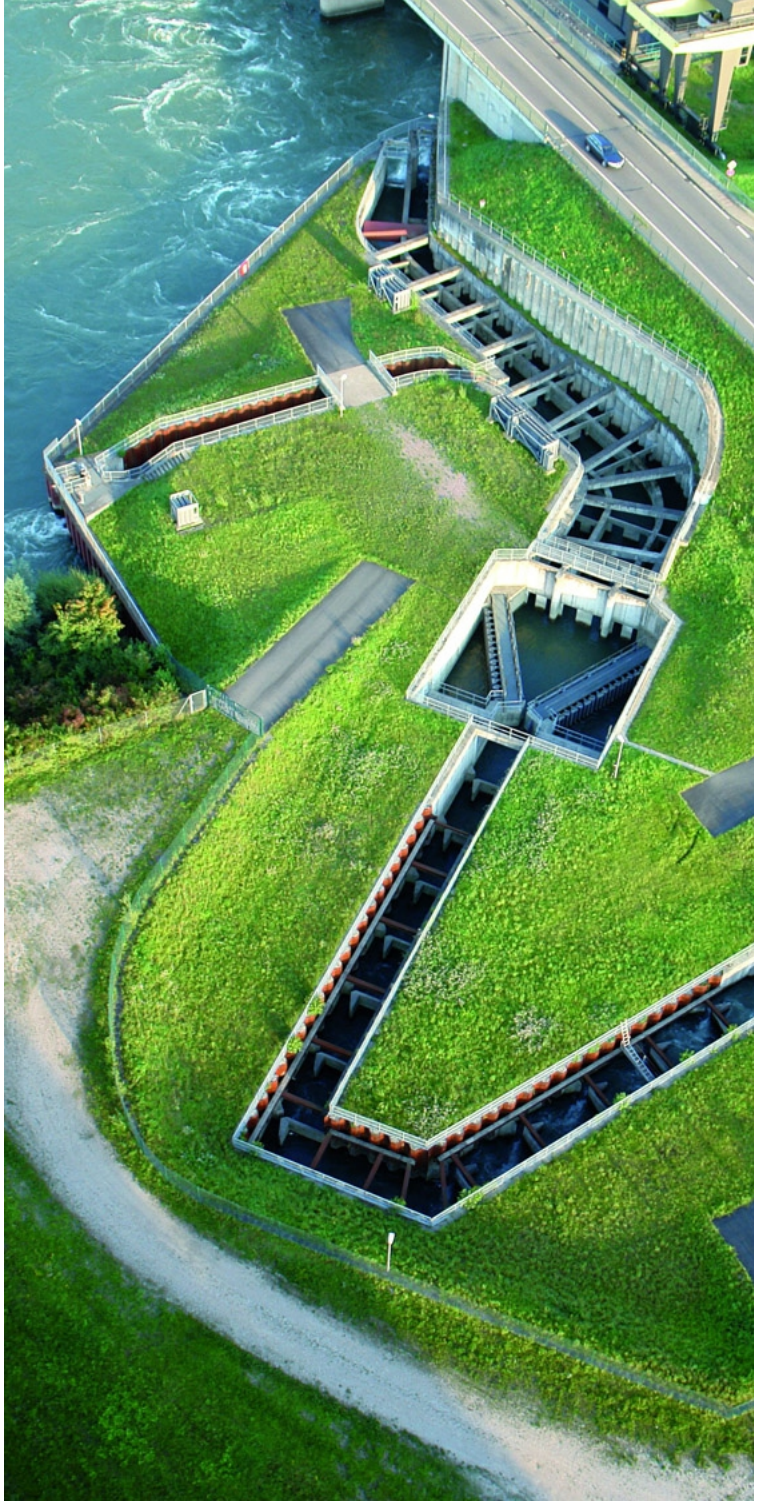
Pierre Gaillard

pierre-p.gaillard@edf.fr

avec Yannig Goude (EDF R&D) et Gilles Stoltz

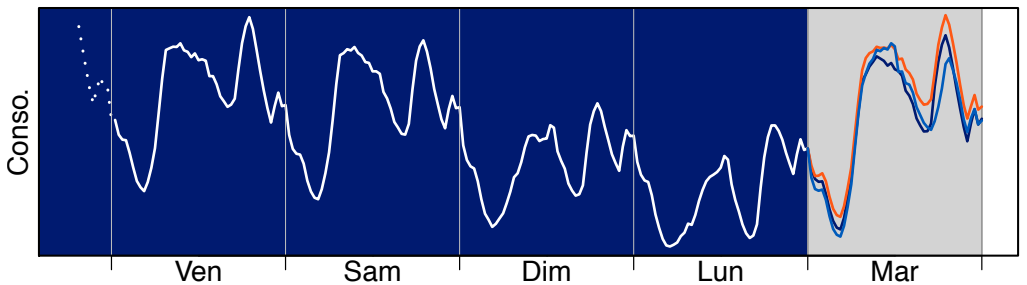
(CNRS, HEC Paris)

27 août 2014 – Journées MAS – Toulouse



Objectif industriel

Prévision à court terme (1 jour à l'avance) de la consommation électrique française



Parallèlement,

- EDF R&D développe de nombreux modèles de prévision
- Le paysage électrique français évolue

➥ remise en question des modèles historiques

Quel modèle utiliser ?

On se propose de les mélanger séquentiellement.

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on forme une prévision $\widehat{y}_t$ de la prochaine observation à partir
  - des observations passées $y_1, \ldots, y_{t-1}$
  - des prévisions actuelles et passées des experts $(x_{k,s})$
- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée

$$\sum_{t=1}^{T} \widehat{\ell}_t$$

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

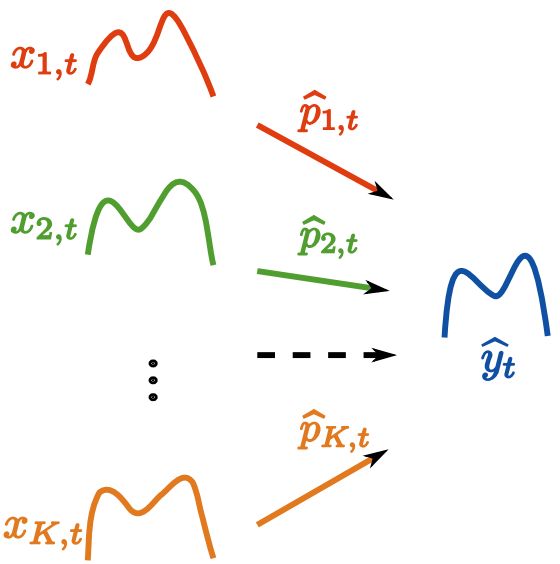
- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée

$$\sum_{t=1}^{T} \widehat{\ell}_t$$



Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée $\sum_{t=1}^{T} \widehat{\ell}_t$.

Exemple de fonction de perte $\ell(x, y)$

$\ell(x,y) =$	Perte carrée $(x-y)^2$	Perte absolue $\lvert x-y\rvert$	Perte relative $\frac{\lvert x-y\rvert}{y}$

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Notre but est de minimiser notre perte cumulée $\sum_{t=1}^{T} \widehat{\ell}_t$.

Remarques

- On ne fait aucune hypothèse stochastique sur la suite d'observations y_t.
- On observe les données et on prévoit séquentiellement.

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Si tous les experts sont mauvais, on ne peut pas espérer avoir une faible erreur.

On évalue donc notre performance relativement à celles des experts

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{k=1,\ldots,K} \sum_{t=1}^{T} \ell(x_{k,t}, y_t)}_{\text{Perte du meilleur expert}} + \underbrace{R_T^{\text{expert}}}_{\text{Regret}}$$

Notre but est d'avoir un regret moyen R_T^{expert} / T qui tend vers 0 quoi qu'il arrive.

Cadre – Prévision séquentielle à l'aide d'experts

On dispose d'une suite bornée d'observations $y_1, y_2, \ldots$ à prévoir instant par instant

À chaque instant t

- un nombre fini K d'experts nous proposent des prévisions $x_{k,t}$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts et on prévoit

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} x_{k,t}$$

- on observe y_t et on subit la perte $\widehat{\ell}_t = \ell(\widehat{y}_t, y_t)$

Si tous les experts sont mauvais, on ne peut pas espérer avoir une faible erreur.

On évalue donc notre performance relativement à celles des experts

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{\boldsymbol{q} \in \Delta_K} \sum_{t=1}^{T} \ell(\boldsymbol{q} \cdot \boldsymbol{x}_t, y_t)}_{\substack{\text{Perte du meilleur} \\ \text{mélange constant}}} + \underbrace{R_T^{\text{convex}}}_{\text{Regret}}$$

Notre but est d'avoir un regret moyen R_T^{convex} / T qui tend vers 0 quoi qu'il arrive.

Meilleur expert $\longrightarrow$ meilleur mélange constant

Si la fonction de perte $\ell(\cdot, y)$ est convexe pour tout y, alors on peut se ramener au cadre linéaire suivant

À chaque instant t

- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\widehat{\ell}_t = \widehat{\boldsymbol{p}}_t \cdot \ell_t \qquad \left(= \sum_{k=1}^{K} \widehat{p}_{k,t} \ell_{k,t}\right)$

Notre but étant de minimiser notre perte cumulée

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} = \underbrace{\min_{k=1,\dots,K} \sum_{t=1}^{T} \ell_{k,t}}_{\text{Perte du meilleur expert}} + \underbrace{R_T}_{\text{Regret}}$$

On peut montrer que $R_T^{\text{expert}} \leqslant R_T^{\text{convex}} \leqslant R_T$ pour ℓ_t bien choisi.

Meilleur expert $\longrightarrow$ meilleur mélange constant

Si la fonction de perte $\ell(\cdot, y)$ est convexe pour tout y, alors on peut se ramener au cadre linéaire suivant

À chaque instant t

- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\widehat{\ell}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{\ell}_t \qquad \left(= \sum_{k=1}^{K} \widehat{p}_{k,t} \ell_{k,t}\right)$

De nombreuses stratégies permettent de contrôler notre perte cumulée

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} \leqslant \underbrace{\min_{k=1,\dots,K} \sum_{t=1}^{T} \ell_{k,t}}_{\text{Perte du meilleur expert}} + \underbrace{\square\sqrt{T \log K}}_{\text{Regret}}$$

On peut montrer que $R_T^{\text{expert}} \leqslant R_T^{\text{convex}} \leqslant R_T$ pour $\boldsymbol{\ell}_t$ bien choisi.

Stratégie ML-prod

Paramètres: $\eta_1, \dots, \eta_K > 0$ Initialisation: $\widehat{\boldsymbol{w}}_0 = (1/K, \dots, 1/K)$

À chaque instant t

- on attribue à l'expert k le poids $\widehat{p}_{k,t} = \eta_k \widehat{w}_{k,t-1} \Big/ \left(\sum_j \eta_j \widehat{w}_{j,t-1} \right)$
- on met à jour les poids

$$\widehat{w}_{k,t} = \widehat{w}_{k,t-1} \left(1 + \eta_k \left(\widehat{\ell}_t - \ell_{k,t} \right) \right)$$

Si $\eta_k \leqslant 1/2$ et $\ell_{k,t} \in [0,1]$, notre erreur cumulée est alors majorée par

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} \leqslant \min_{k=1,\dots,K} \left\{ \underbrace{\sum_{t=1}^{T} \ell_{k,t}}_{\text{Perte de l'expert } k} + \underbrace{\frac{\log K}{\eta_k} + \eta_k \sum_{t=1}^{T} \left(\widehat{\ell}_t - \ell_{k,t} \right)^2}_{\text{Regret}} \right\}$$

Stratégie ML-prod

Paramètres: $\eta_1, \dots, \eta_K > 0$ Initialisation: $\widehat{\boldsymbol{w}}_0 = (1/K, \dots, 1/K)$

À chaque instant t

- on attribue à l'expert k le poids $\widehat{p}_{k,t} = \eta_k \widehat{w}_{k,t-1} \Big/ \left(\sum_j \eta_j \widehat{w}_{j,t-1} \right)$
- on met à jour les poids

$$\widehat{w}_{k,t} = \widehat{w}_{k,t-1} \left(1 + \eta_k \left(\widehat{\ell}_t - \ell_{k,t} \right) \right)$$

Si on peut optimiser $\eta_k = \sqrt{(\log K) / \sum_t (\widehat{\ell}_t - \ell_{k,t})^2}$

$$\underbrace{\sum_{t=1}^{T} \widehat{\ell}_t}_{\text{Notre perte}} \leqslant \min_{k=1,\dots,K} \left\{ \underbrace{\sum_{t=1}^{T} \ell_{k,t}}_{\text{Perte de l'expert } k} + \underbrace{2\sqrt{(\log K) \sum_{t=1}^{T} \left(\widehat{\ell}_t - \ell_{k,t} \right)^2}}_{\text{Regret}} \right\}$$

Borne de regret de ML-prod – Faibles pertes

Si une stratégie vérifie la borne de regret

$$\sum_{t=1}^{T} \widehat{\ell}_t \quad \leqslant \quad \min_{k=1,\dots,K} \left\{ \quad \sum_{t=1}^{T} \ell_{k,t} \quad + \quad 2\sqrt{\log K \sum_{t=1}^{T} \left(\widehat{\ell}_t - \ell_{k,t}\right)^2} \quad \right\}$$

Alors elle vérifie aussi

$$\sum_{t=1}^{T} \widehat{\ell}_t \quad \leqslant \quad \min_{k=1,\dots,K} \left\{ \quad \sum_{t=1}^{T} \ell_{k,t} \quad + \quad 2\sqrt{\log K \sum_{t=1}^{T} \ell_{k,t} + 16 \log K} \quad \right\}$$

Le regret par rapport à l'expert k est d'autant plus faible que l'expert est bon !

Borne de regret de ML-prod – Pertes i.i.d.

Pour l'instant toutes nos bornes étaient déterministes (dans le pire des cas).
Que se passe-t-il si on suppose nos pertes i.i.d?

Hypothèse

Si les vecteurs de pertes $(\ell_{1,t}, \dots, \ell_{K,t})$ sont
- indépendant et identiquement distribués
- avec un écart entre l'espérance de la perte du meilleur expert $k^\star$ et les autres

Si une stratégie vérifie

$$\sum_{t=1}^{T} \widehat{\ell}_t \leqslant \sum_{t=1}^{T} \ell_{k^\star,t} + 2\sqrt{\log K \sum_{t=1}^{T} \left(\widehat{\ell}_t - \ell_{k^\star,t}\right)^2}$$

alors son regret est constant avec grande probabilité

$$\sum_{t=1}^{T} \widehat{\ell}_t \leqslant \sum_{t=1}^{T} \ell_{k^\star,t} + O(1)$$

Experts spécialisés

À chaque instant t

- chaque expert donne une mesure de confiance en sa prévision $I_{k,t} \in [0,1]$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\widehat{\ell}_t = \widehat{\boldsymbol{p}}_t \cdot \ell_t \qquad \left(= \sum_{k=1}^{K} \widehat{p}_{k,t} \ell_{k,t}\right)$

On souhaite minimiser notre confidence regret par rapport à chaque expert

$$R_{k,T} = \sum_{t=1}^{T} I_{k,t}\left(\widehat{\ell}_t - \ell_{k,t}\right)$$

Le cas particulier $I_{k,t} = 0$ exprime l'inactivité de l'expert k à l'instant t. Les jour où l'expert pense être meilleur sont plus pondérés.

Experts spécialisés

À chaque instant t

- chaque expert donne une mesure de confiance en sa prévision $I_{k,t} \in [0,1]$
- on attribue un poids $\widehat{p}_{k,t}$ à chacun des experts
- chaque expert subit la perte $\ell_{k,t}$
- on subit la perte linéaire $\widehat{\ell}_t = \widehat{\boldsymbol{p}}_t \cdot \ell_t \qquad \left(= \sum_{k=1}^K \widehat{p}_{k,t}\ell_{k,t}\right)$

Si on applique l'algorithme ML-prod avec les pseudo pertes

$$\tilde{\ell}_{k,t} = I_{k,t}\ell_{k,t} + (1 - I_{k,t})\widehat{\ell}_t \,,$$

alors

$$R_{k,T} = \sum_{t=1}^{T} I_{k,t}\left(\widehat{\ell}_t - \ell_{k,t}\right) \leqslant 2\sqrt{\log K \sum_{t=1}^{T} I_{k,t}^2}$$

Application – Le jeu de données

Il inclut 1696 jours du 1er janvier 2007 au 15 juin 2012 et contient

- la consommation électrique des clients EDF
- De l'information exogène
  - météo: température, nébulosité, vent
  - temporelle: date, tarif spécial
  - perte de clients

On retire les jours fériés ± 2.

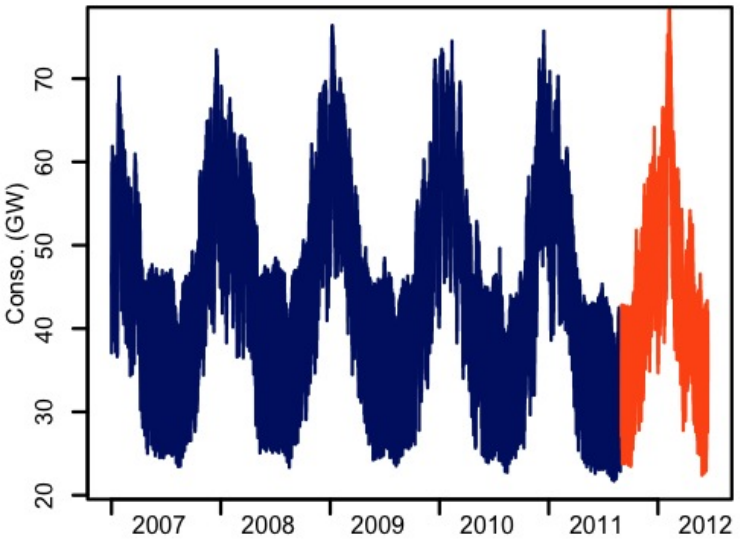
On divise le jeu de données en deux

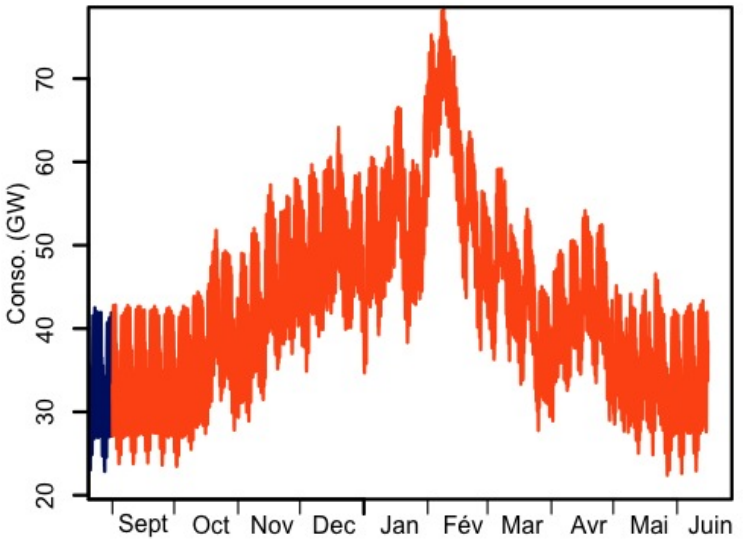
- Jan. 2007 – Août. 2011: ensemble d'entrainement $\rightarrow$ construire les experts
- Sept. 2011 – Juin. 2012: ensemble de test $\rightarrow$ tester les experts et le mélange

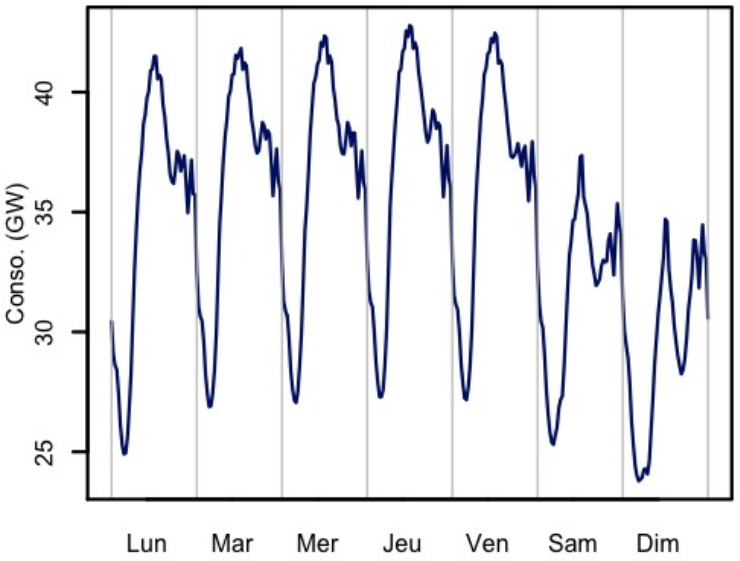
On considère 3 experts construit à partir de trois modèles statistiques différents

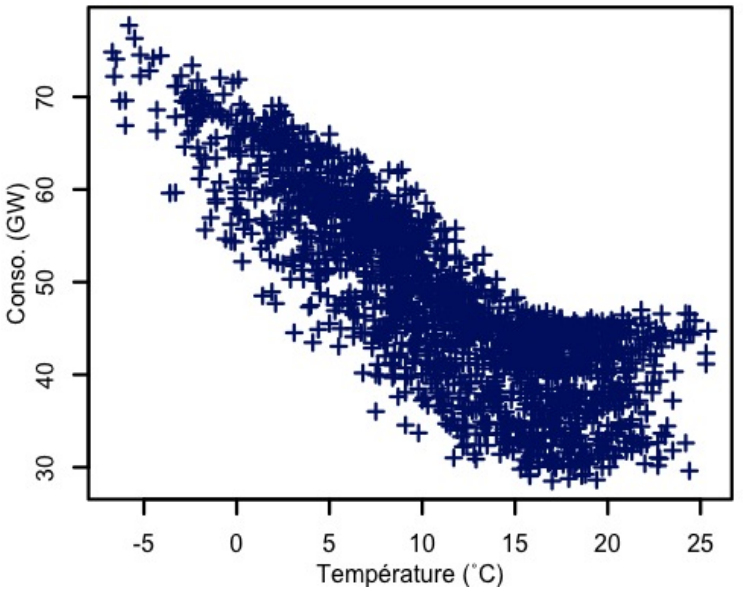
- Régression régularisée sur une base de splines
- Régression linéaire sur un espace fonctionnel
- Clustering de données fonctionnelles à partir d'une base d'ondelettes

Les données en images...







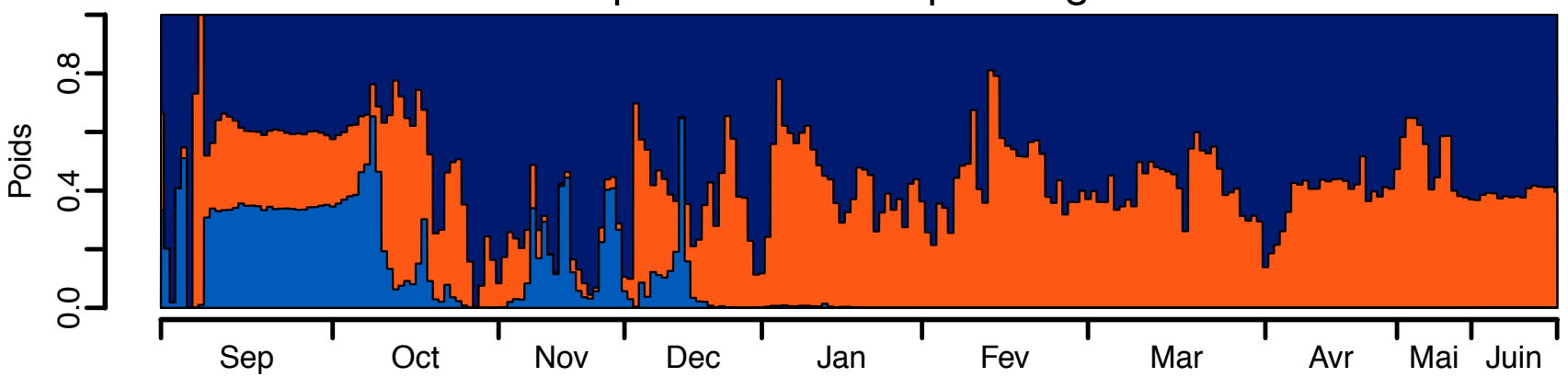


Performances des experts et du mélange

On évalue les performances à l'aide du RMSE (root mean square error) sur l'année test. Plus c'est faible, mieux c'est !

Method	RMSE (MW)
Meilleur expert	744
Meilleur mélange constant	629
ML-prod	626

Évolution des poids attribués par l'algorithme



Références

- Application EDF :
  - Gaillard, Goude. Forecasting the electricity consumption by aggregating experts; how to design a good set of experts, 2014
  - Devaine, Gaillard, Goude, Stoltz. Forecasting electricity consumption by aggregating specialized experts, 2013.
- Théorie (borne du second ordre et conséquences) :
  - Gaillard, Stoltz, van Erven. A second-order bound with excess losses, 2014
  - Wintenberger. Optimal learning with Bernstein Online Aggregation, 2014.

Merci !