Data Challenge

Nazih Benoumechiara, Nicolas Meyer et Taïeb Touati

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université, Paris

Mardi 29 mai 2018

Journées de Statistique 2018

SORBONNE UNIVERSITÉ

LPSM
LABORATOIRE DE PROBABILITÉS
STATISTIQUE & MODÉLISATION

Introduction

Données

- Consommation électrique sur l'île d'Ouessant du 13 septembre 2015 au 13 septembre 2016 (maille horaire)
- Données météorologiques du 13 septembre 2015 au 13 septembre 2016 (maille tri-horaire)
- Prévisions météorologiques pour la semaine du 14 au 20 septembre 2016 (maille tri-horaire)

Introduction

Données

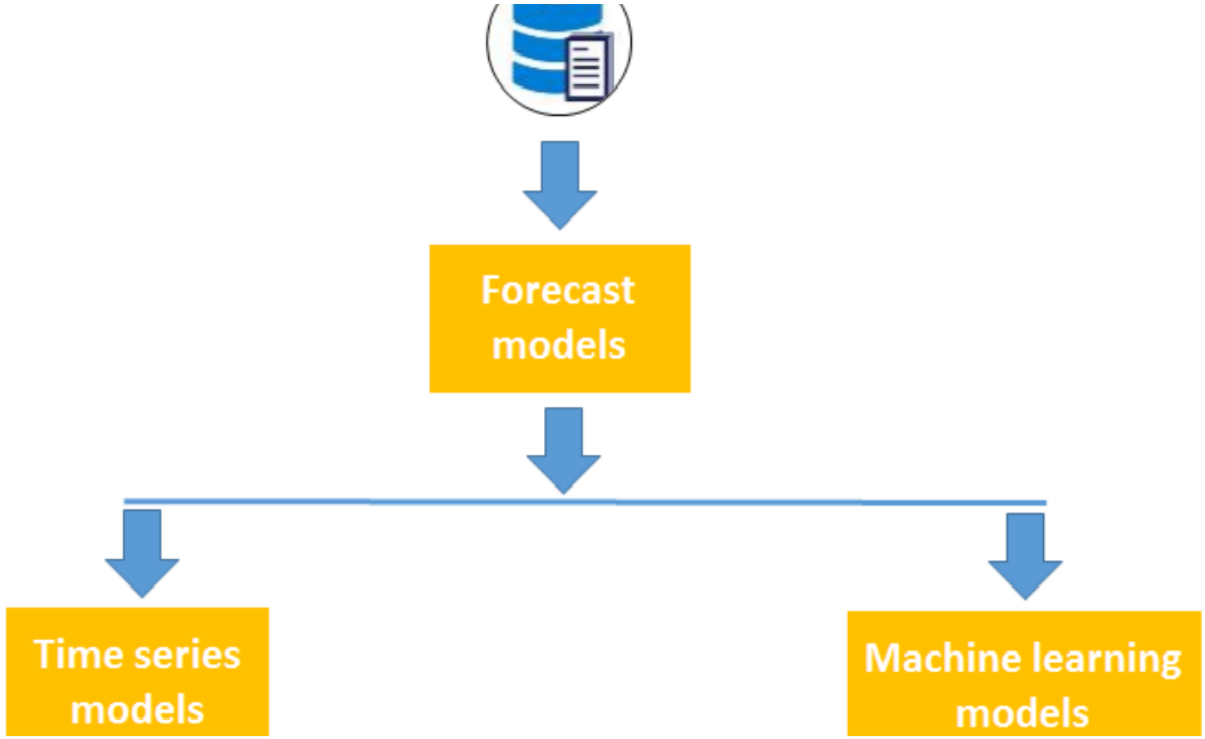
- Consommation électrique sur l'île d'Ouessant du 13 septembre 2015 au 13 septembre 2016 (maille horaire)
- Données météorologiques du 13 septembre 2015 au 13 septembre 2016 (maille tri-horaire)
- Prévisions météorologiques pour la semaine du 14 au 20 septembre 2016 (maille tri-horaire)

Objectif

- Prédire la consommation électrique sur la semaine du 14 au 20 septembre 2016
- Critère d'évaluation : erreur absolue moyenne en pourcentage (MAPE)

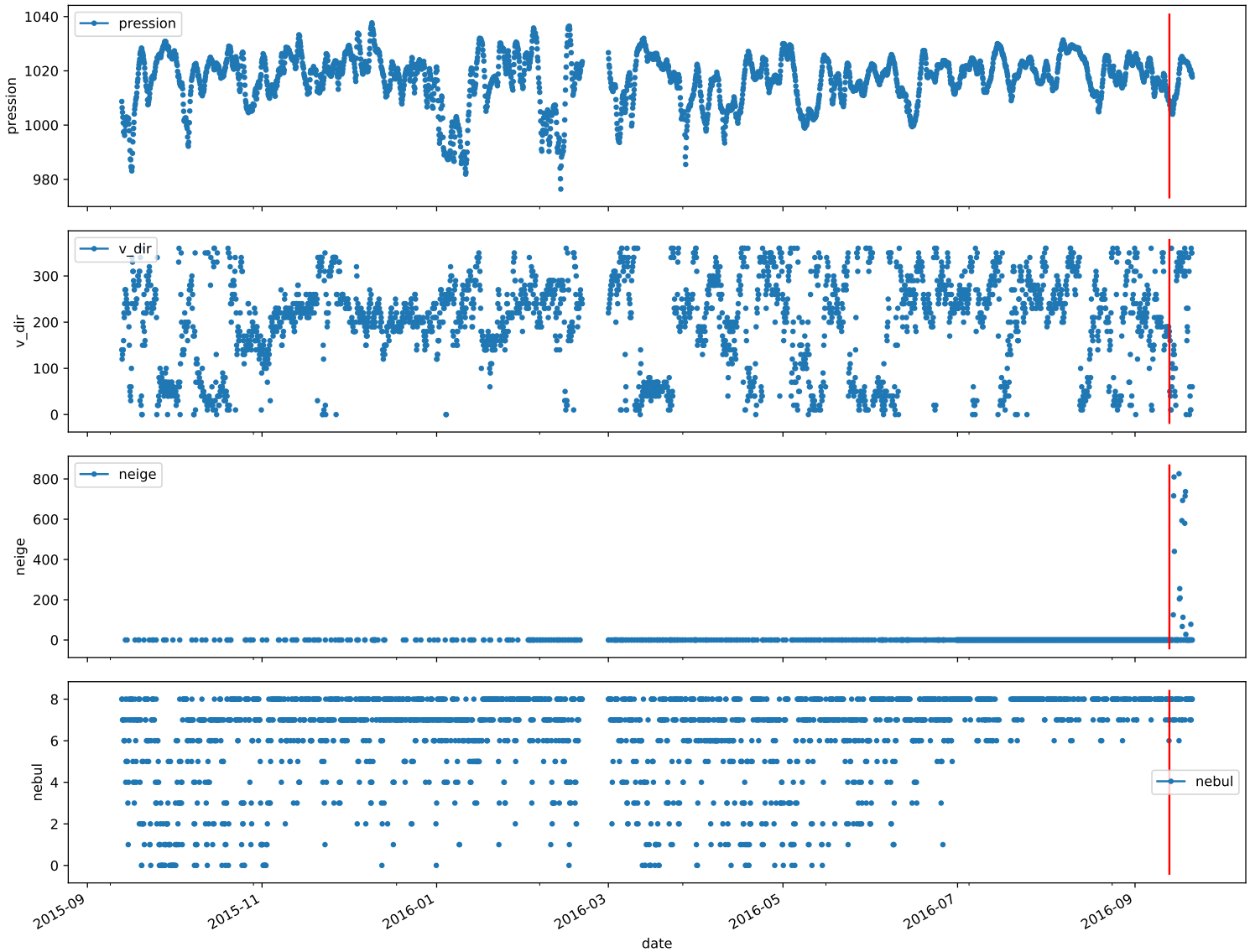
Méthode générale

Mélange de modèles de séries temporelles et de machine learning

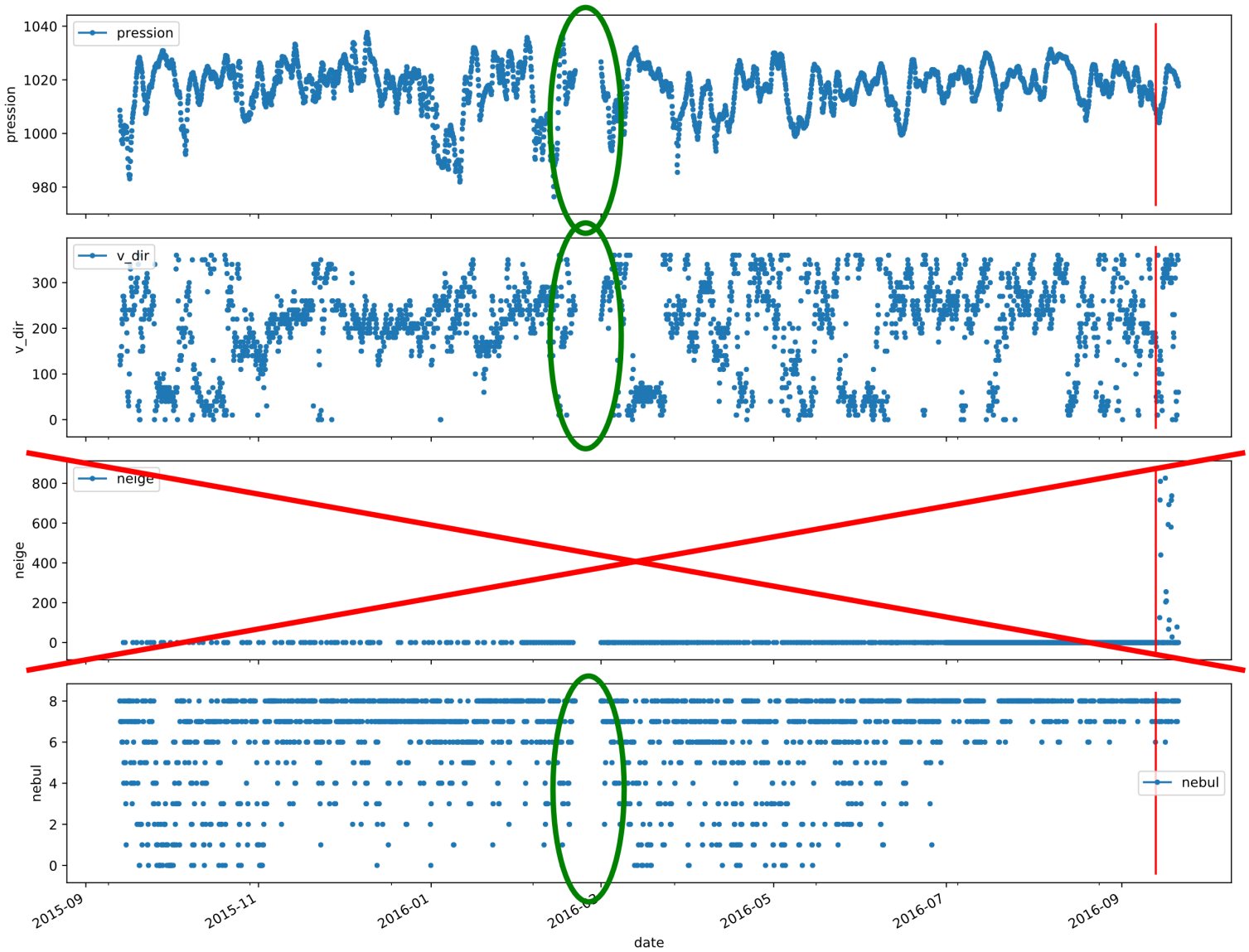


Contents

Données météorologiques



Données météorologiques



Données météorologiques

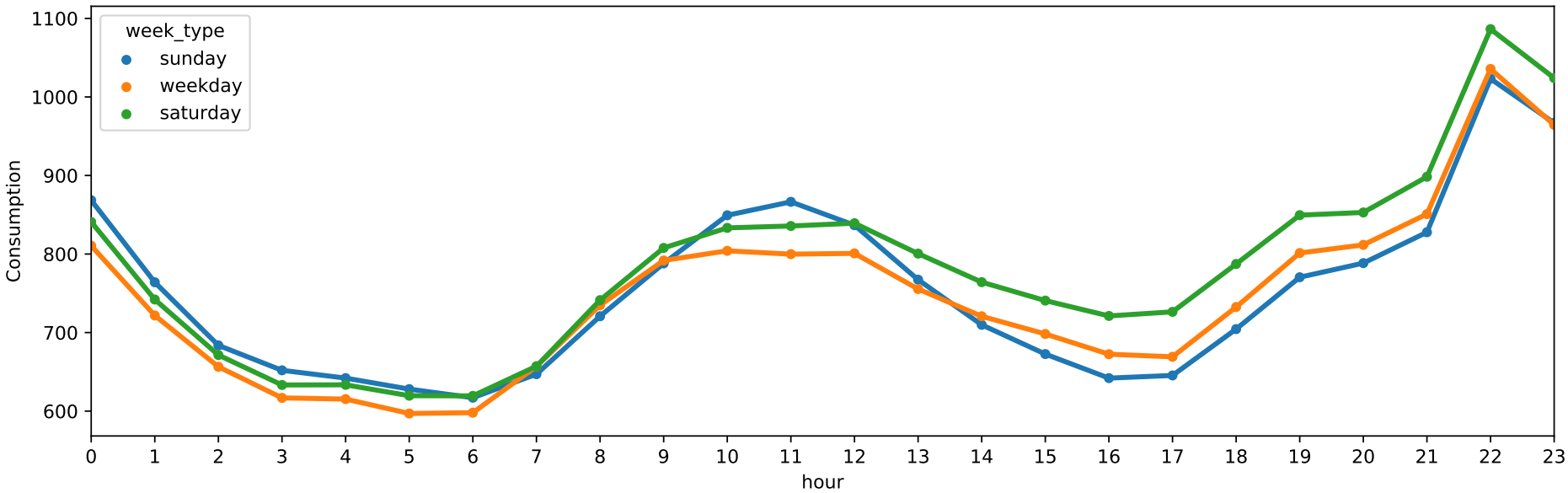
- Pas de neige dans l'échantillon d'entraînement, mais dans la prévision
- Absence de données météo la semaine du 21 au 28 février 2016
- Données horaires (électriques) vs tri-horaires (météorologiques)

	date	y	heure_ete	type	temperature	pression	HR	rosee	visibilite	v_moy	v_raf	v_dir	RR3h	neige	nebul	national_holiday
0	2015-09-13 00:00:00	NaN	True	train	12.5	1008.7	81.0	9.3	40.0	9.260	18.520	140.0	0.0	NaN	8.0	False
1	2015-09-13 01:00:00	526.166667	True	train	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False
2	2015-09-13 02:00:00	495.000000	True	train	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False
3	2015-09-13 03:00:00	446.166667	True	train	12.3	1006.4	83.0	9.5	40.0	11.112	16.668	120.0	0.0	NaN	8.0	False
4	2015-09-13 04:00:00	365.833333	True	train	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False

- Interpolation des valeurs météorologiques manquantes : quadratique, linéaire ou valeur la plus proche

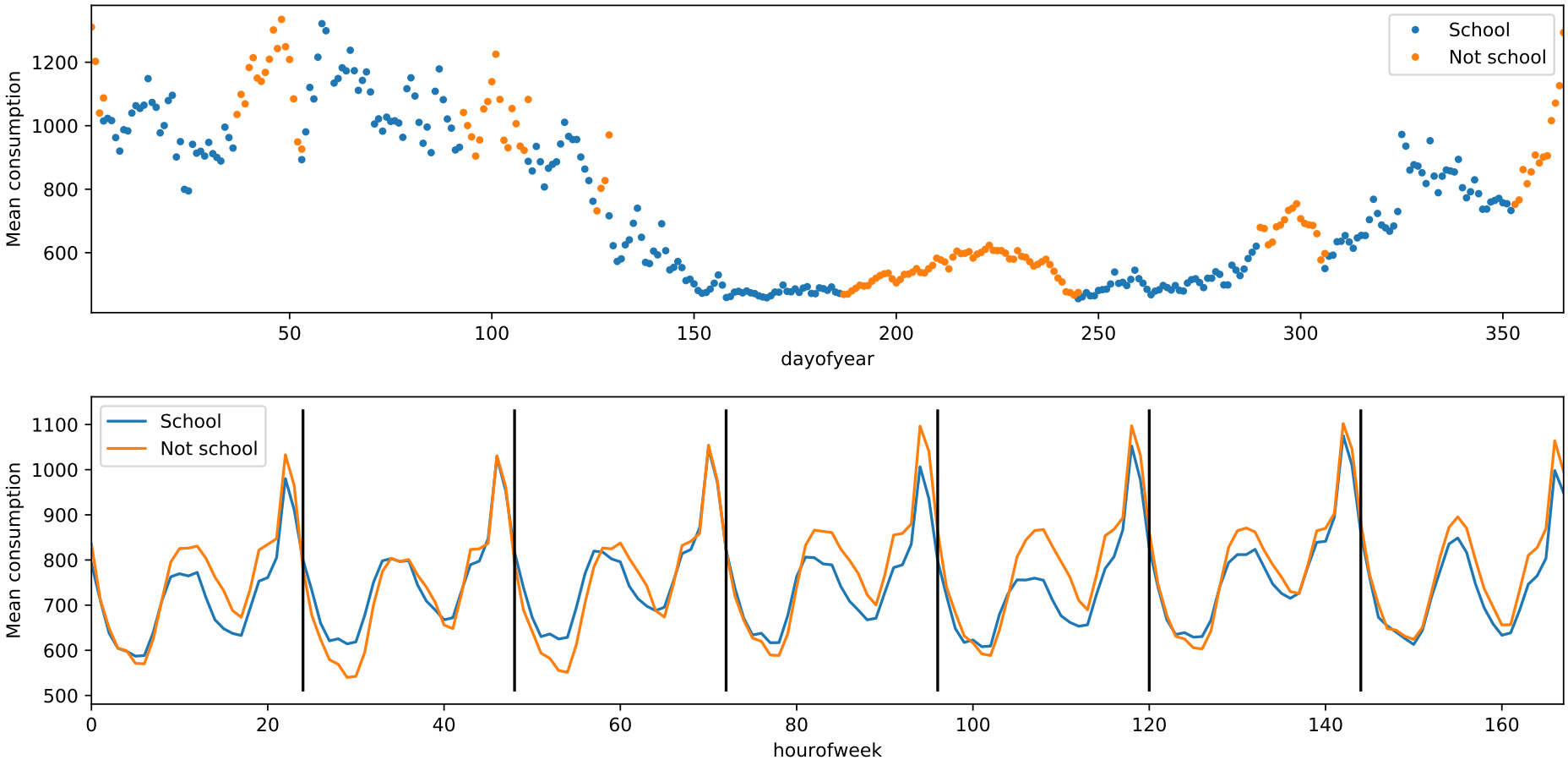
Données de consommation

Influence du jour de la semaine (jour ouvré vs samedi vs dimanche):



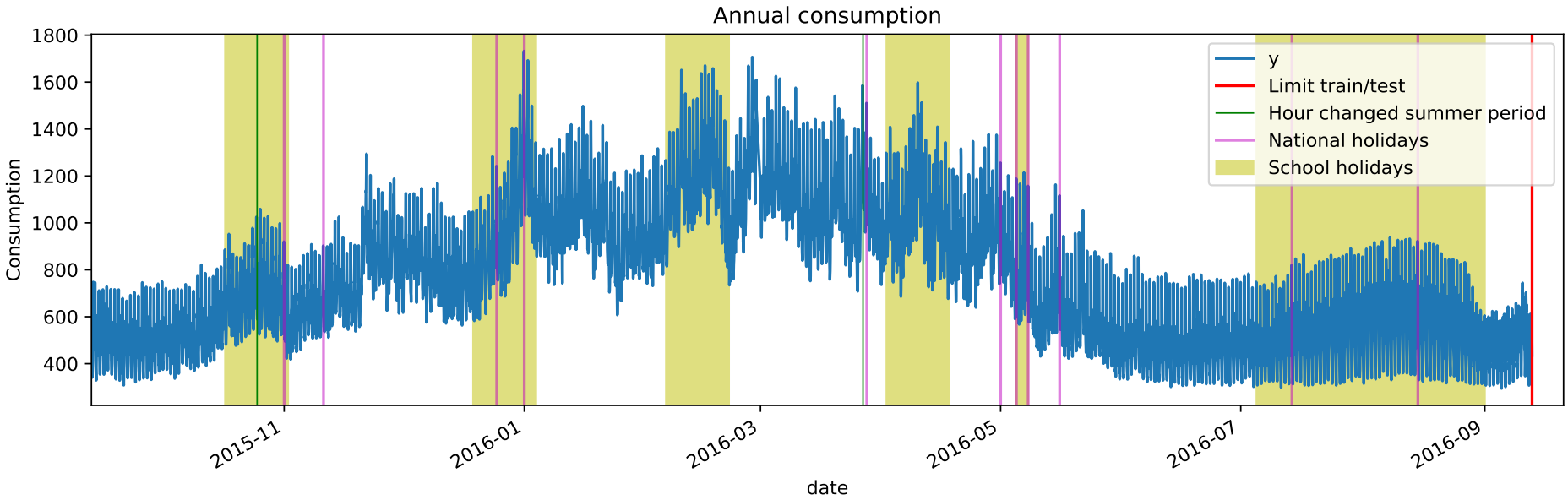
Données de consommation

Influence des vacances scolaires:



Données de consommation

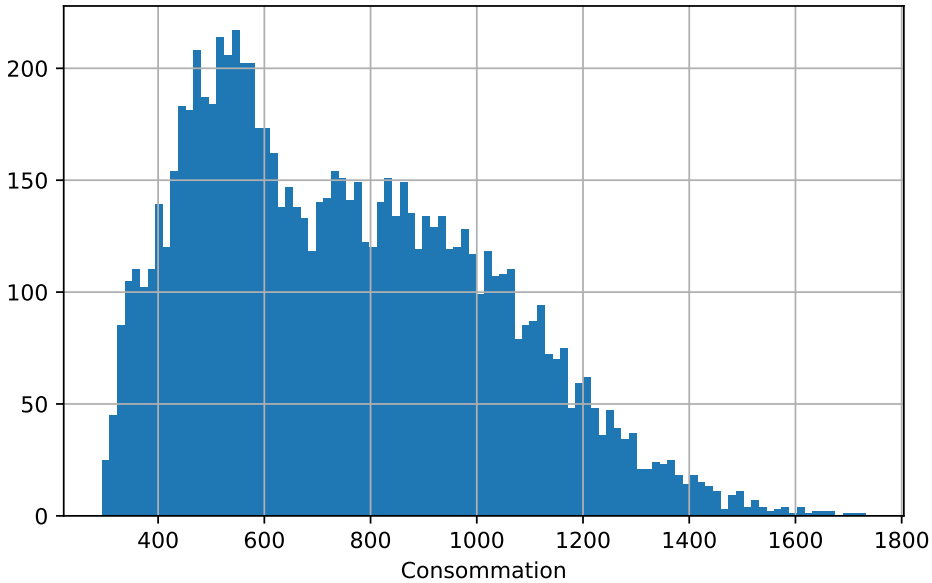
Sur toute l'année:

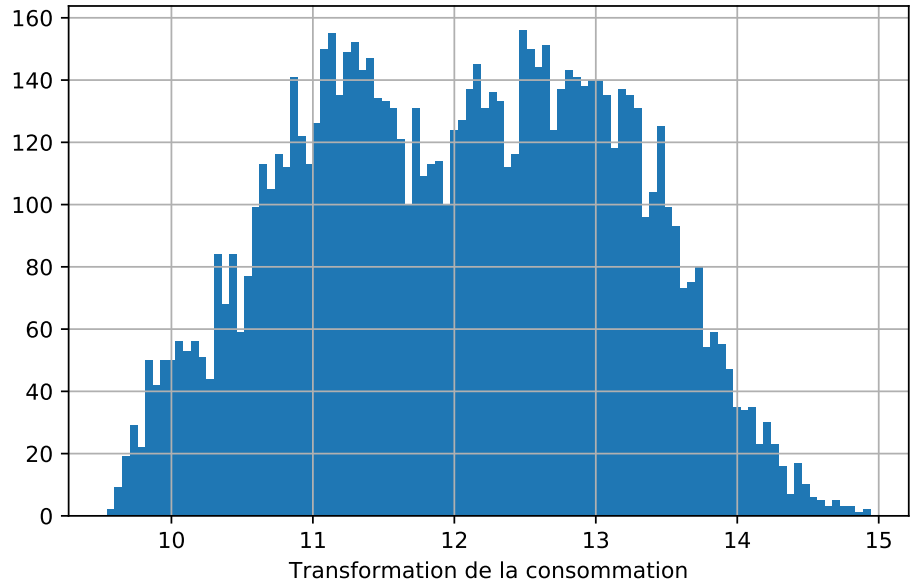


Contents

Apprentissage / Test local et transformation de la consommation

- Train local sur la période 13/09/2015 - 04/09/2016 ($\approx$ 1 an),
- Test local sur la période 05/09/2016 - 12/09/2016 (8 jours).
- Transformation Boxcox de la consommation.

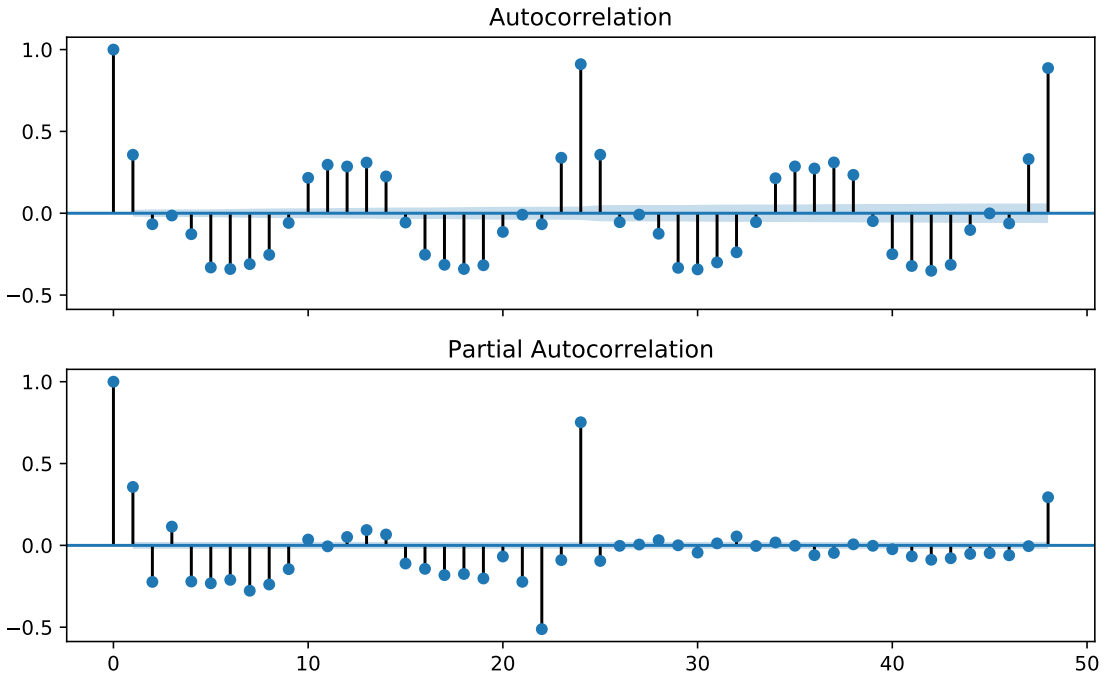




ARMA: choix de (p, q)

La prévision est effectuée sur la différence de consommation, puis retransformée.

- Via les ACF et PACF:

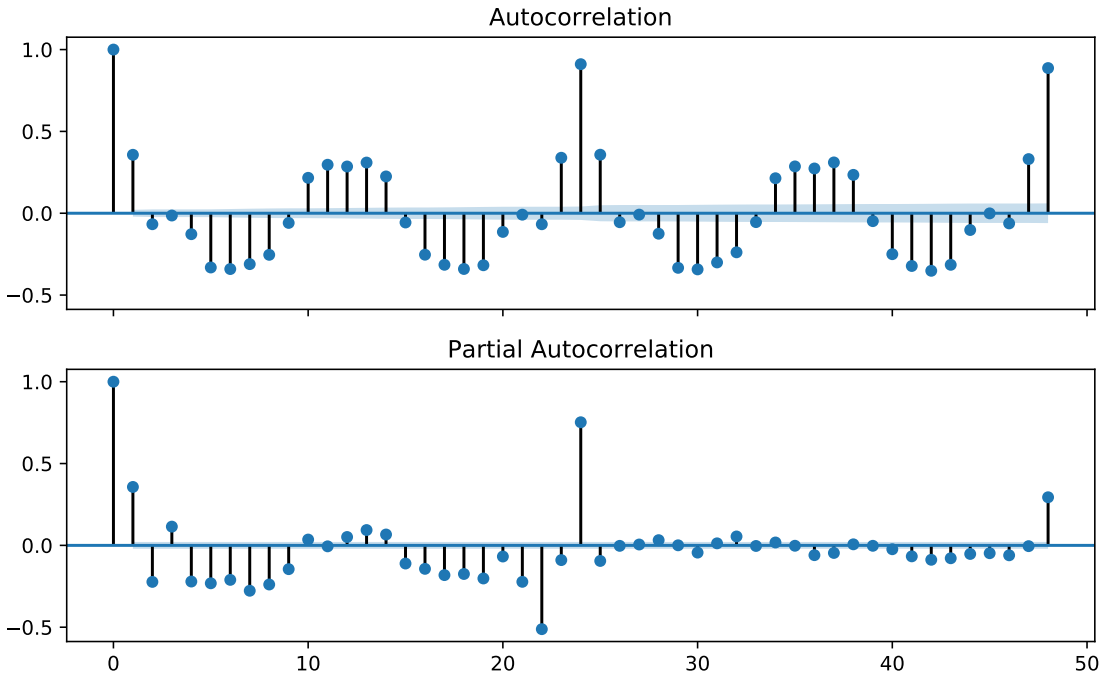


- Par un grid search sur p et q et validés sur 5 train/test locaux différents (soit 3125 modèles testés) et en prenant la combinaison qui minimise le MAPE.

ARMA: choix de (p, q)

La prévision est effectuée sur la différence de consommation, puis retransformée.

- Via les ACF et PACF:



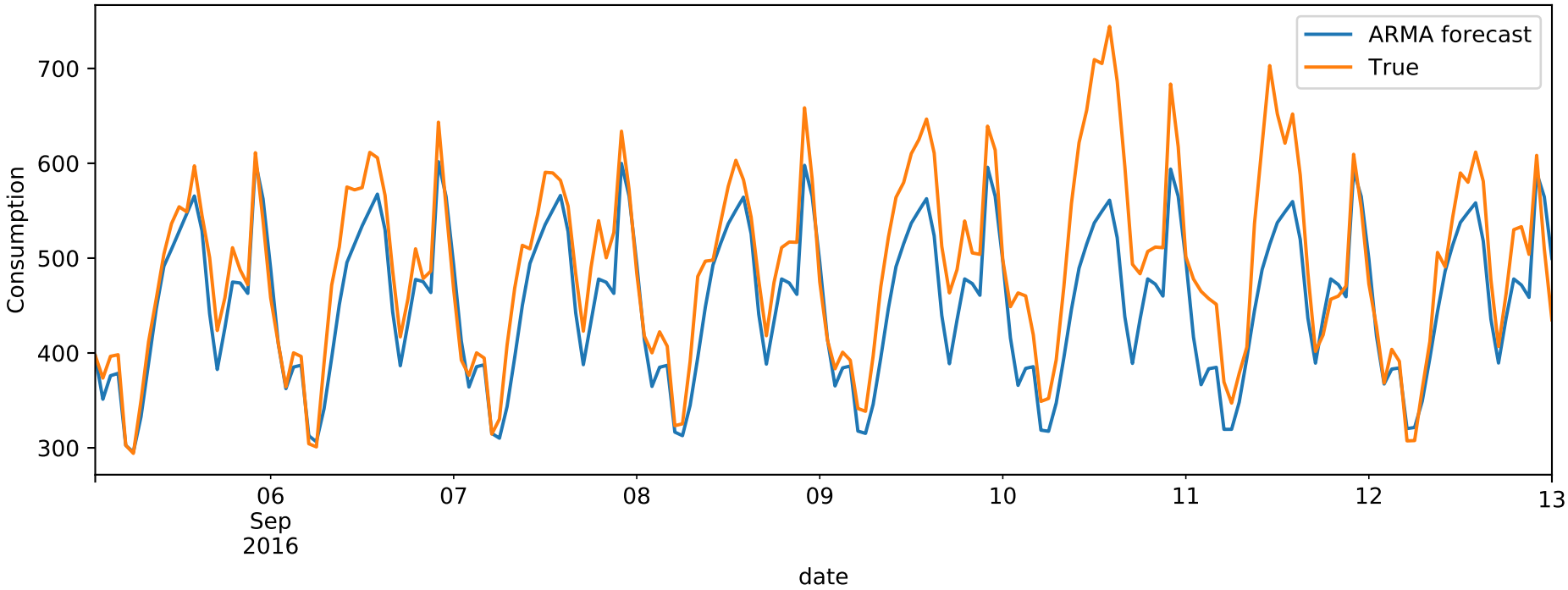
- Par un grid search sur p et q et validés sur 5 train/test locaux différents (soit 3125 modèles testés) et en prenant la combinaison qui minimise le MAPE.

Choix optimal: $p = 24, q = 24$

ARMA

Avec ce choix de paramètres :

- erreur MAPE à 8.21%
- mauvaise prédiction pour les 10 et 11 septembre (week-end)



Gradient Boosting Model: Feature Engineering

- Température ressentie décrite par[2]

$$T_{felt} = (A - T) * \sqrt{V},$$

où T est la température, V est la vitesse du vent et A une température t.q. T_{felt} est le plus corrélé à la consommation.

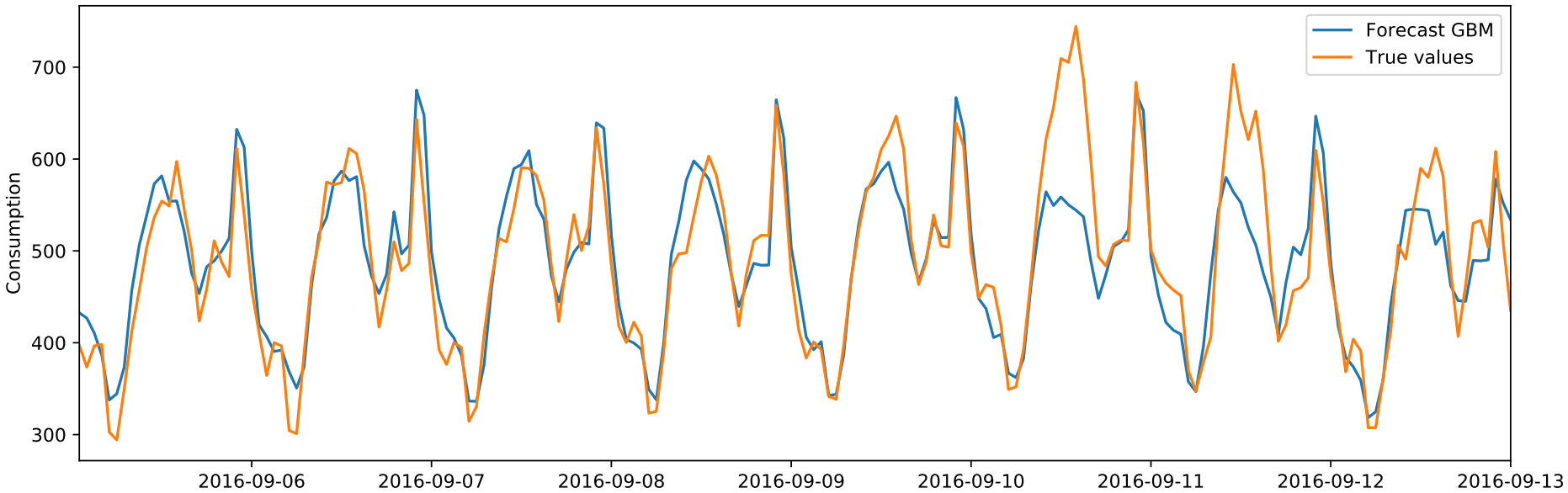
- Décalage des valeurs pour
  - certaines données météo (3, 6, 12 et 24h),
  - la consommation (1, 2, 3 et 4 semaines).
- Target encoding[3] sur le train local. Moyenne sur :
  - le jour de la semaine, jour de l'année
  - l'heure conditionnellement
    - Jour de la semaine
    - Vacance
    - Jour férié
  - etc...

[2] Peter Lusis et al. "Short-term residential load forecasting: Impact of calendar effects and forecast granularity". In: *Applied Energy* 205 (2017), pp. 654–669.

[3] Daniele Micci-Barreca. "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems". In: *ACM SIGKDD Explorations Newsletter* 3.1 (2001), pp. 27–32.

Gradient Boosting Model

- Tuning des hyper-paramètres (vitesse d'apprentissage, régularisation, nombres de feuilles, etc...)
- Erreur MAPE : 6.46 %



- La prédiction reste mauvaise le weekend du 10 et 11 septembre.

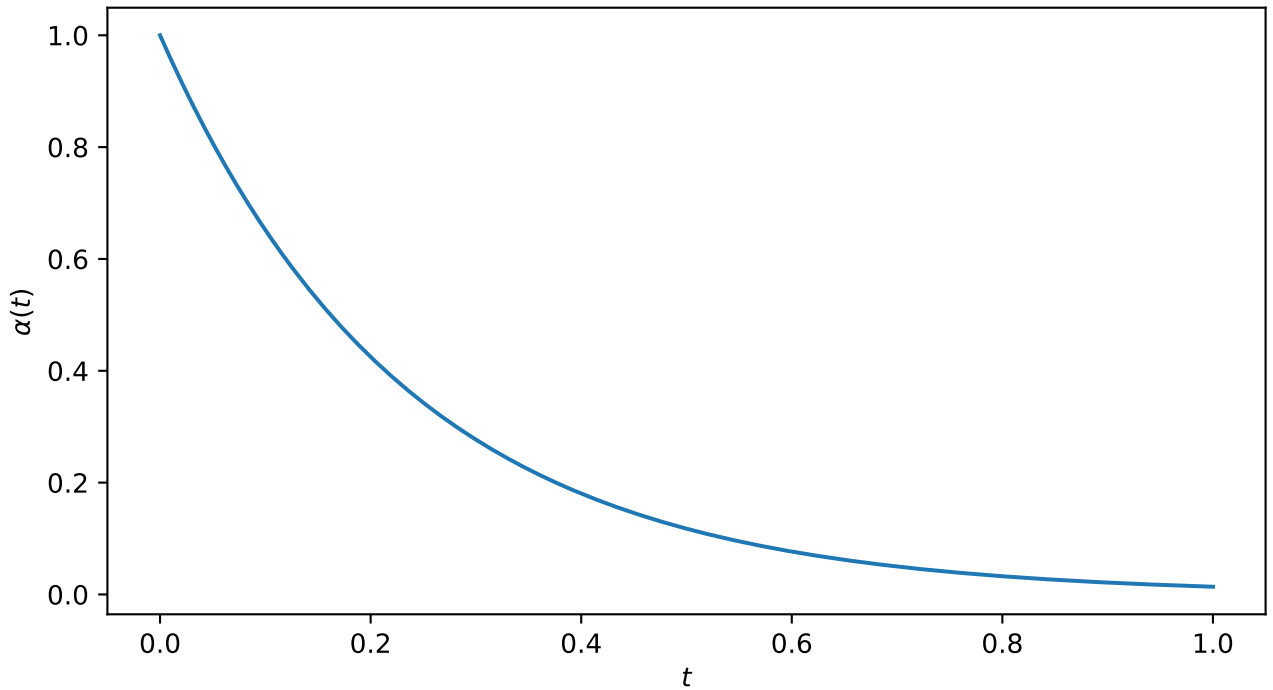
Mélange entre ARMA et GBM

- ARMA : bonnes prédictions au début de la semaine, moins sur la fin
  $\hookrightarrow$ utiliser un mélange entre ARMA et GBM
- Modèle de mélange:

$$\hat{y}_{\text{pred}} = \alpha(t)\hat{y}_{\text{ARMA}} + (1 - \alpha(t))\hat{y}_{\text{GBM}}, \tag{1}$$

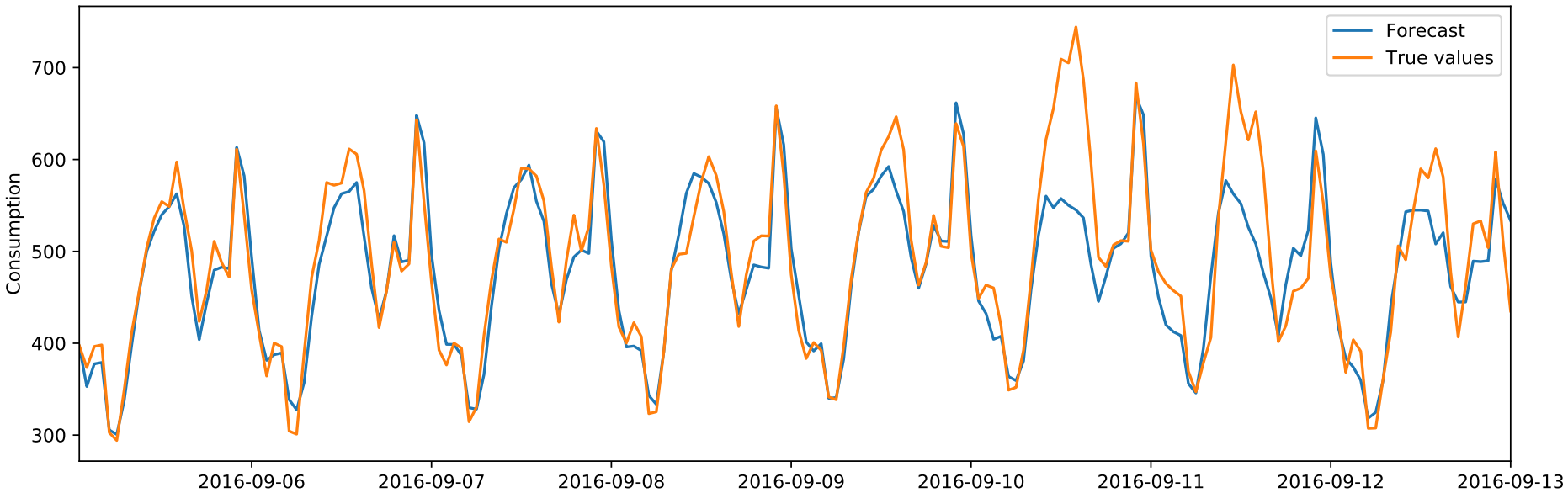
  avec $\alpha(t) = \exp(-t/\lambda), \quad \lambda > 0.$
- Le paramètre λ est choisi de sorte à minimiser l'erreur $\hat{y}_{\text{pred}}$

$\hookrightarrow$On trouve $\lambda_{\text{optimal}} = 0.29$, pour une erreur MAPE de 5.84% sur le test local.



Résultats sur le test local

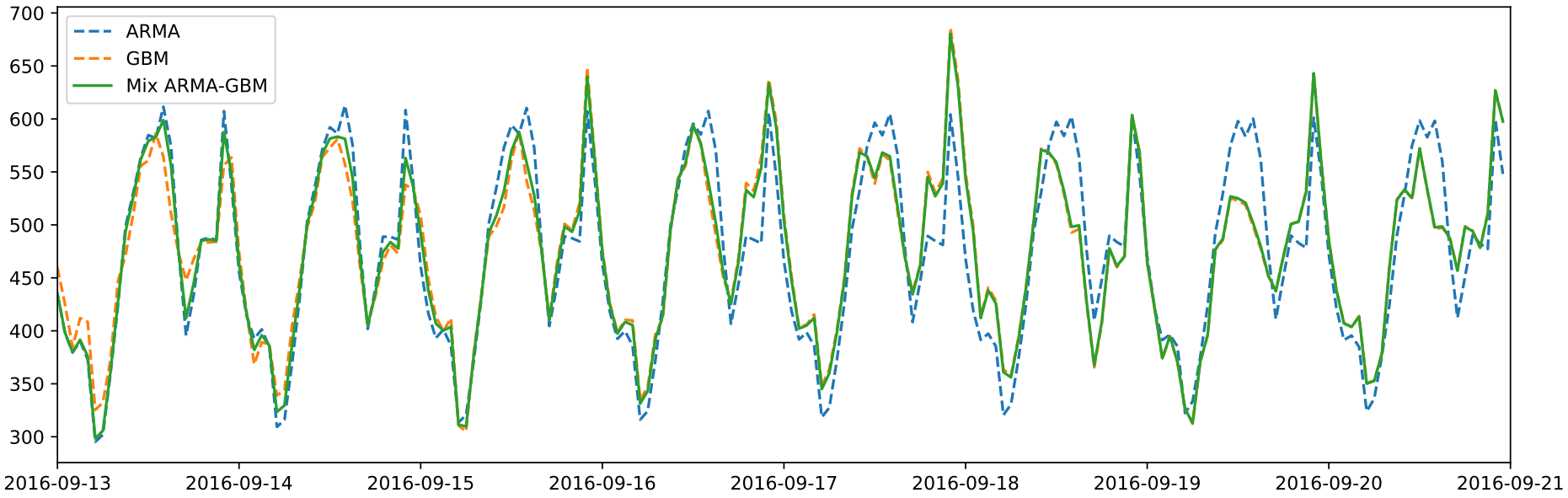
On obtient la prédiction suivante sur la semaine du 05/09/2016 au 12/09/2016



Prédiction sur le test

La pipeline est la suivante:

- Faire une prévision ARMA
- Prévision par GBM
- Mélanger les prévisions ARMA-GBM avec le λ_{optimal}



Contents

Perspectives

- Améliorer la méthode ARMA par une stabilisation de la variance[1]
- Tester la robustesse de la méthode sur un échantillon de séries temporelles basées sur des lieux différents
- Adapter la problématique à des cas de consommations individuelles après clustering
- Codes, notebook et slides : https://github.com/NazBen/solution-challenge-jds18

[1] Kianoosh G Boroojeni et al. "A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon". In: *Electric Power Systems Research* 142 (2017), pp. 58–73.

Merci de votre attention

Nazih Benoumechiara, Nicolas Meyer, Taïeb Touati
(prénom.nom@upmc.fr)
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université

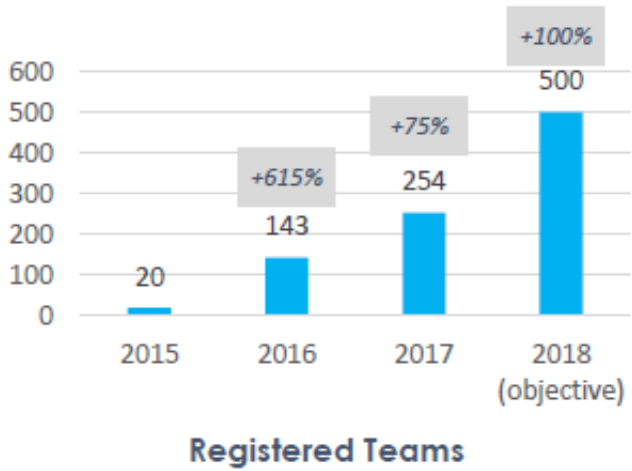
DATA SCIENCE GAME

1 Qualification *Online*

April-June 2018

2 Finals *In Paris*

September 2018



Registered Teams

20

Finalist Teams