Prévision de pertes RTE

Pia CHANCEREL, Louis HEMADOU, François MEDINA

Mai 2020

Section 1

Introduction

Contexte

- RTE : gestion du réseau haute tension français
- 2,5% de la consommation perdue : 500 M€
- Objectif : modèle de prédiction de pertes

Problématique

- 35 variables explicatives
- Prédiction à long terme (1 an)

Ressources

- Base de données : 35 variables et pertes horaires
- Rapport de stage sur la prédiction de pertes

Objectif

- Identifier les variables significatives
- Identifier et paramétrer un algorithme de prédiction efficace

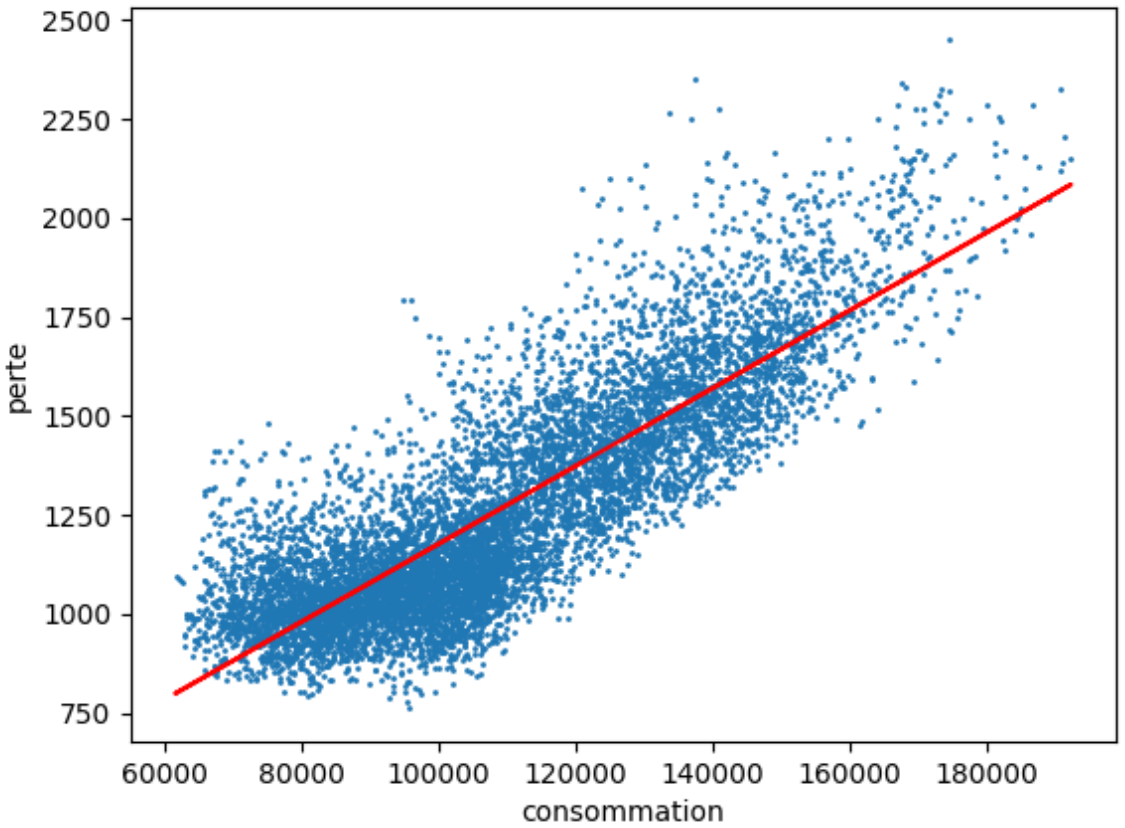
Section 2

Récupération des données

Réception

- Activité du réseau (consommation, production, énergie) : eco2mix
- Pertes relevées : portail client RTE

Visualisation



Description

- date/heure : représentatif de l'activité et du climat
- consommation, prévisions : charge et imprévus
- production : régimes d'activation et de charge du réseau
- échanges : charge supplémentaire sur des points individuels

Traitement des fichiers

- encodage utf-8, comma separated values
- colonnes en `snake_case`
- dates/heures numériques

pertes au même format que les données de consommation/production/échanges :

- élimination des lignes parasites (commentaires)
- un fichier par an
- une ligne par heure (colonnes jour/mois pour accès facile en observation)

Section 3

Algorithmes de prédiction

Régression linéaire

Dépendance linéaire à déterminer :

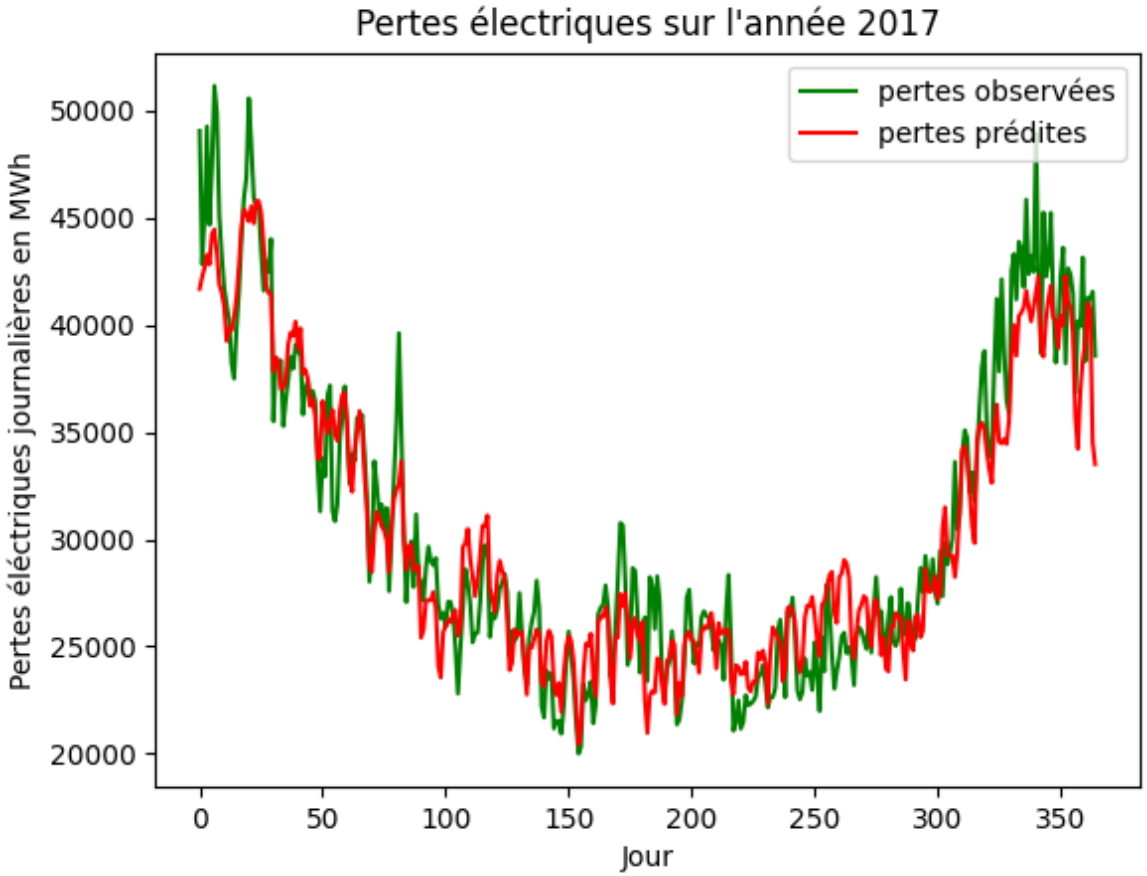
$$f(x, \epsilon) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \epsilon \tag{1}$$

Résultats selon le pré-traitement des données :

traitement	R^2
normalisées	0.80
standardisées	0.83
orthogonalisées	-1.9

Régression linéaire

De bons résultats en standardisant les données :



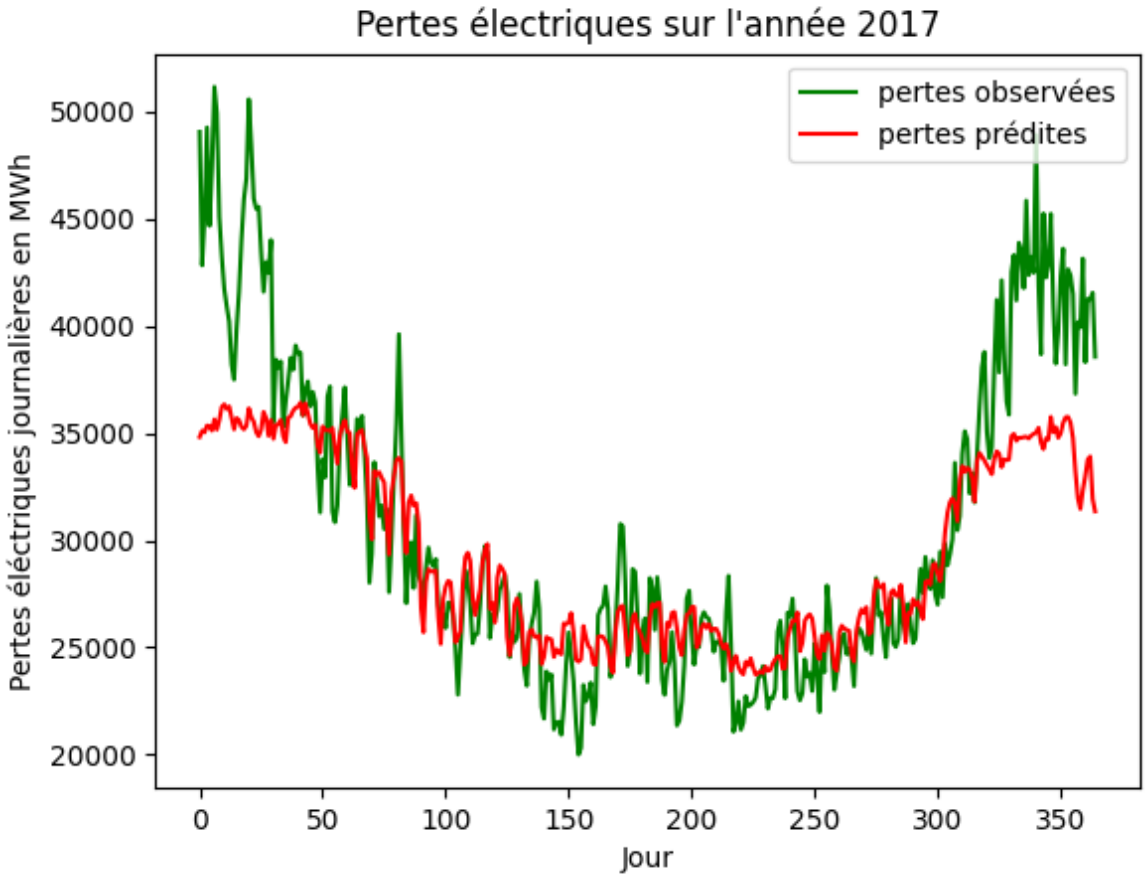
Machine à noyaux

Coefficient de détermination selon le prétraitement :

traitement	R^2
normalisées	0.63
standardisées	0.61
orthogonalisées	-0.17

Machine à noyau

Résultats avec la machine à noyau, données standardisées :



Réseau de neurones

- utilisation de Tensorflow
- structure du réseau :

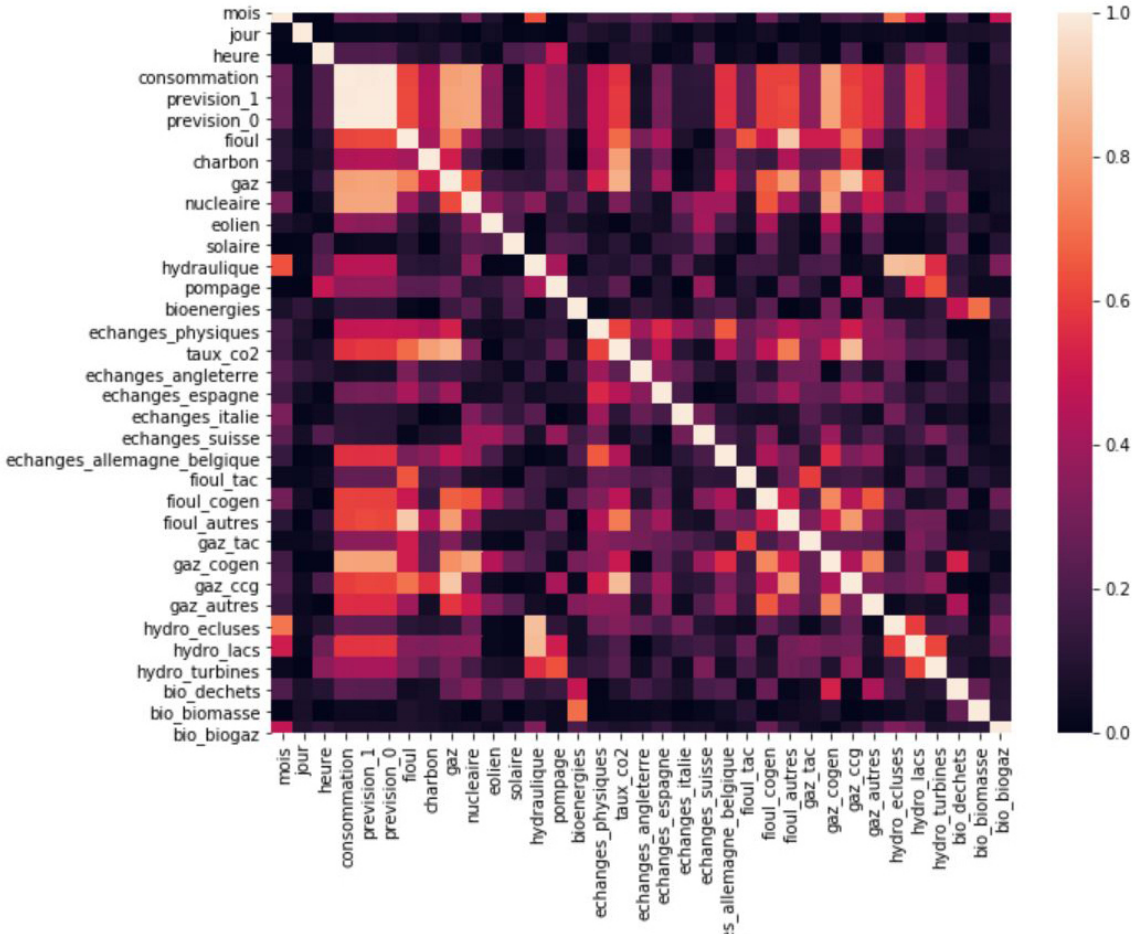
neurones	**activation**
35	(entrée)
400	sigmoïde
400	sigmoïde
100	ReLU
1	linéaire (sortie)

Section 4

Sélection de variables

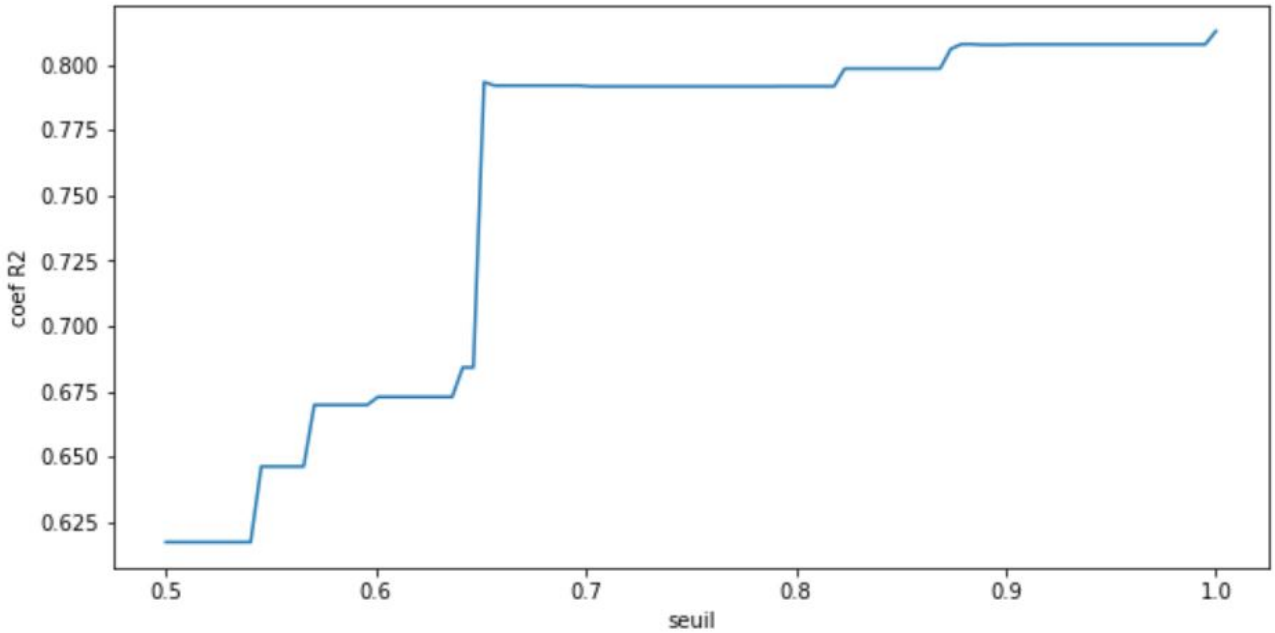
Élimination des doublons

Corrélation de Pearson : $\rho(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$



Élimination des doublons

Coefficient de détermination selon le seuil d'élimination :



seuil à $0.8 > 0.65$, supprimant `consommation`, `prevision_0`, `fioul`, `gaz`, `hydraulique`, `hydro_lacs`, `taux_co2`, 1.4% de perte

`prevision_1` moins bruitée que `consommation`

Élimination des doublons

- Redondance prévision/consommation ($\rho > 0.99$)
- Filtrage des doublons $0.5 < |\rho| < 0.9$, suppression d'une variable par doublon.

Sélection des variables explicatives

La corrélation de Pearson ne suffit plus pour l'explication des pertes :

- sensibilité aux valeurs extrêmes
- relations non linéaires

On cherche donc une meilleure méthode : détermination d'un score pour chaque variable

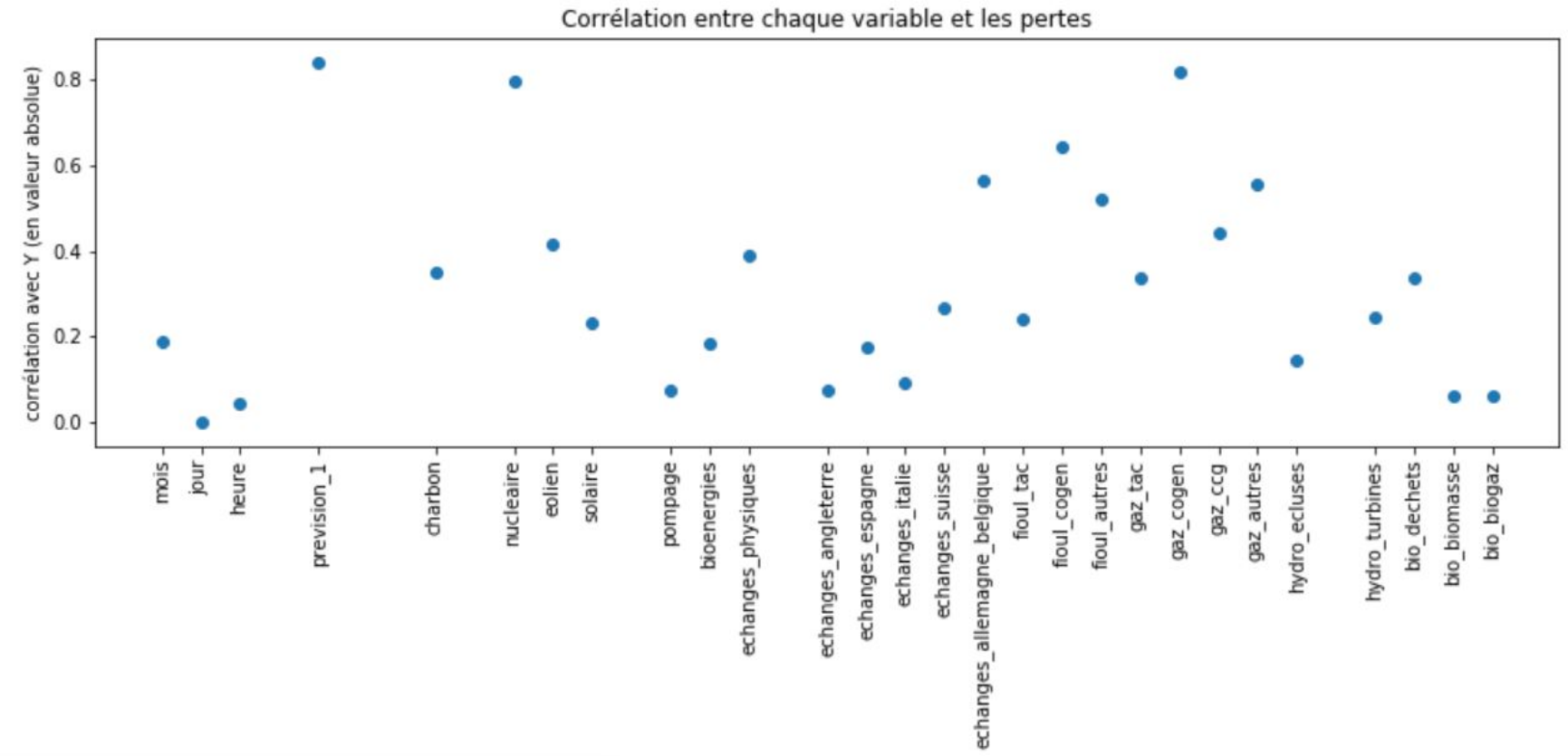
Sélection des variables explicatives

Méthodes de filtrage :

- matrices de corrélation (de Pearson)
- PCA (Principal Component Analysis)

Corrélation avec les pertes

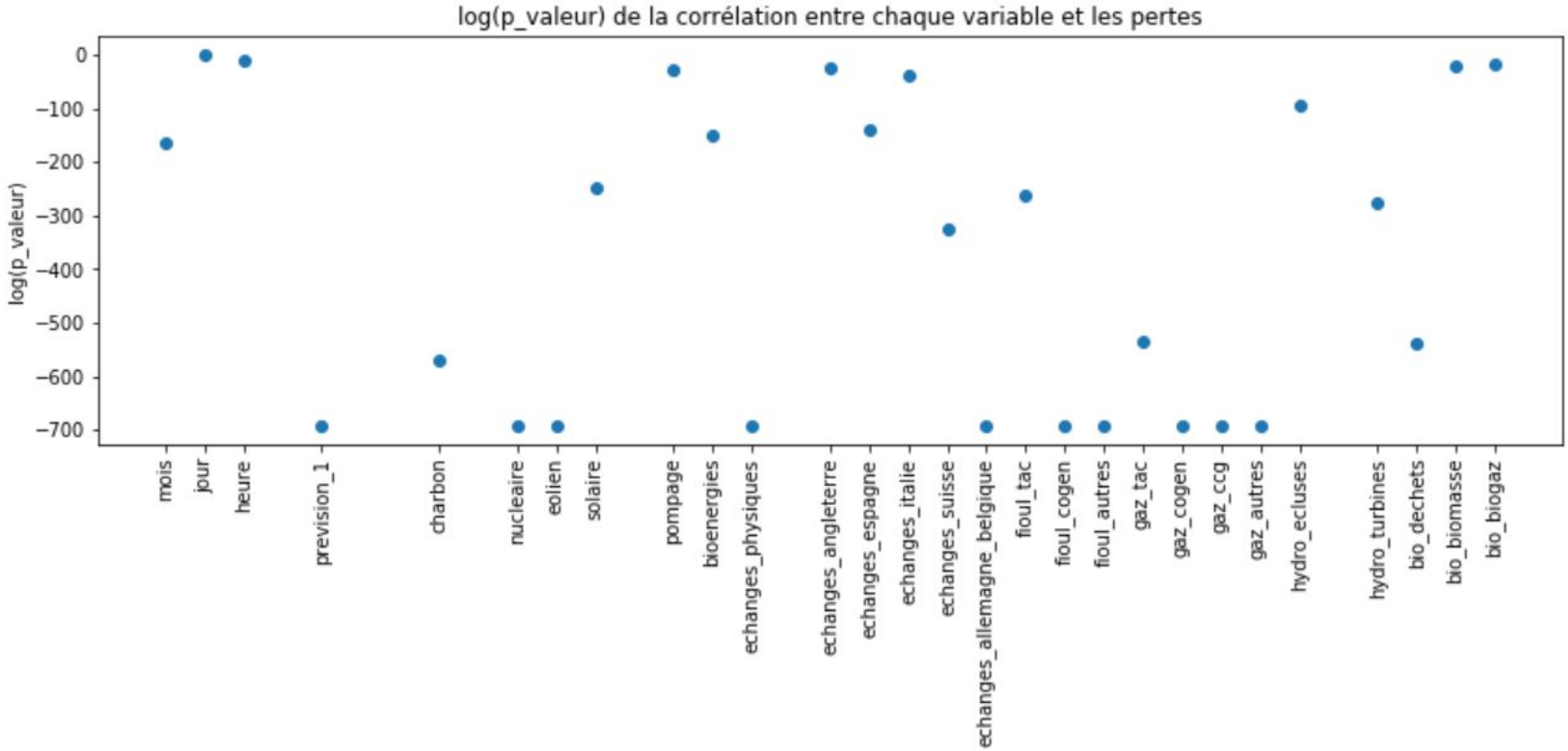
Corrélation de chaque variable avec les pertes :



Beaucoup de variables peu significatives ($\rho \approx \pm 0.5$)

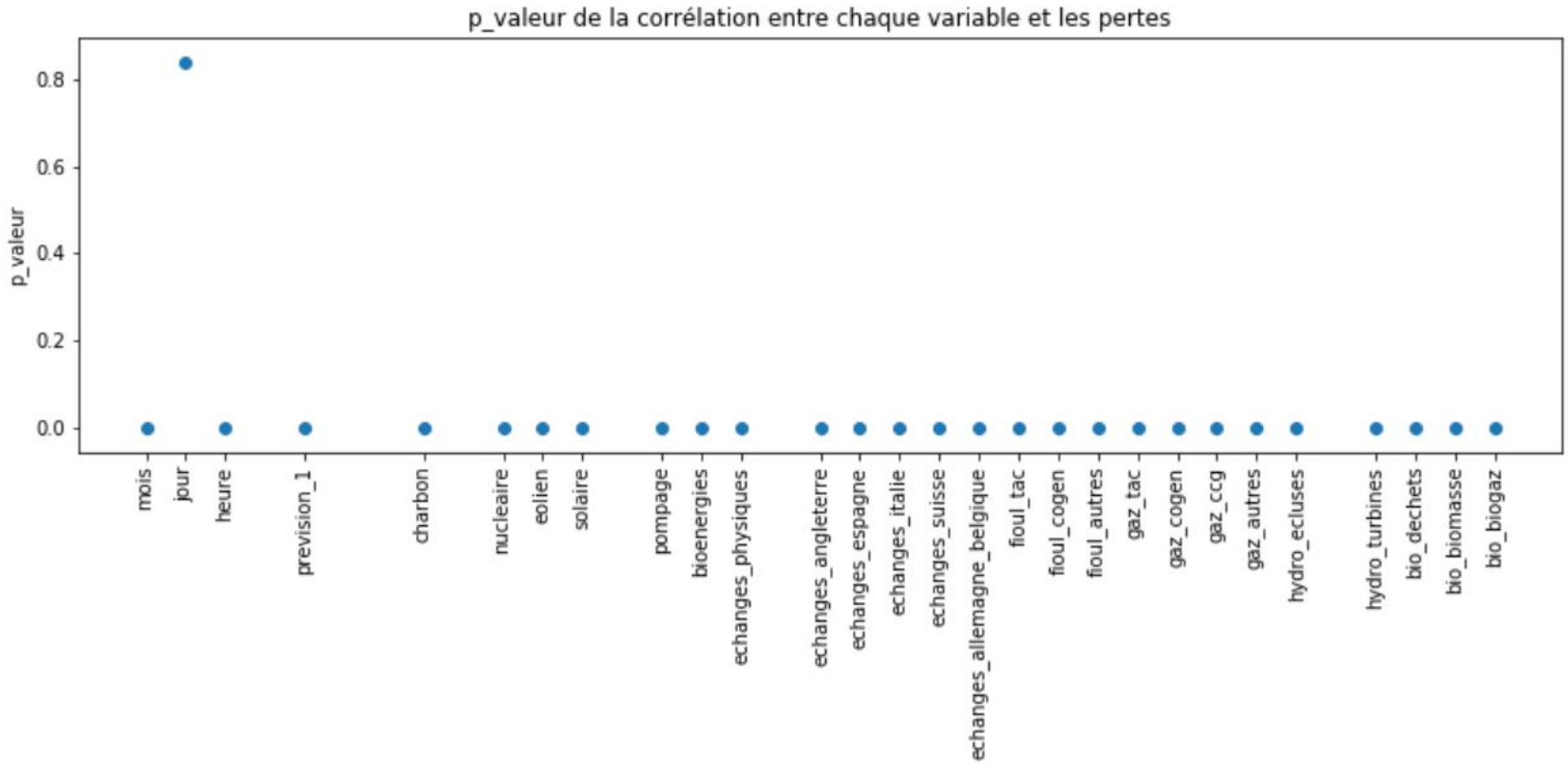
P-value

p-valeur avec `scipy.stats` : probabilité du même ρ dans un système décorrélé



P-value

Sans le log on trouve un intrus :



En effet, peu de variations périodiques significatives à l'échelle d'un mois

Corrélation avec les pertes

`prevision_1`, `nucleaire` et `gaz_cogen`

- $\rho \approx 0.8$: forte corrélation
- `nucleaire` et `gaz_cogen` très corrélées : on peut n'en garder qu'une
- p-value très basse

Corrélation avec les pertes

`mois`, `jour`, `solaire` ainsi que des données `bio` spécifiques (7 featues) :

- Corrélation à 0
- Pas de relation linéaire : exclues pour la régression linéaire
- En les excluant : R^2 de 03805 à 0.799, soit 0.7%
- Comparé au doublons : 2 fois moins de perte, même nombre d'éliminés
- On élimine aussi les fortes p-values (probablement décorrélées)

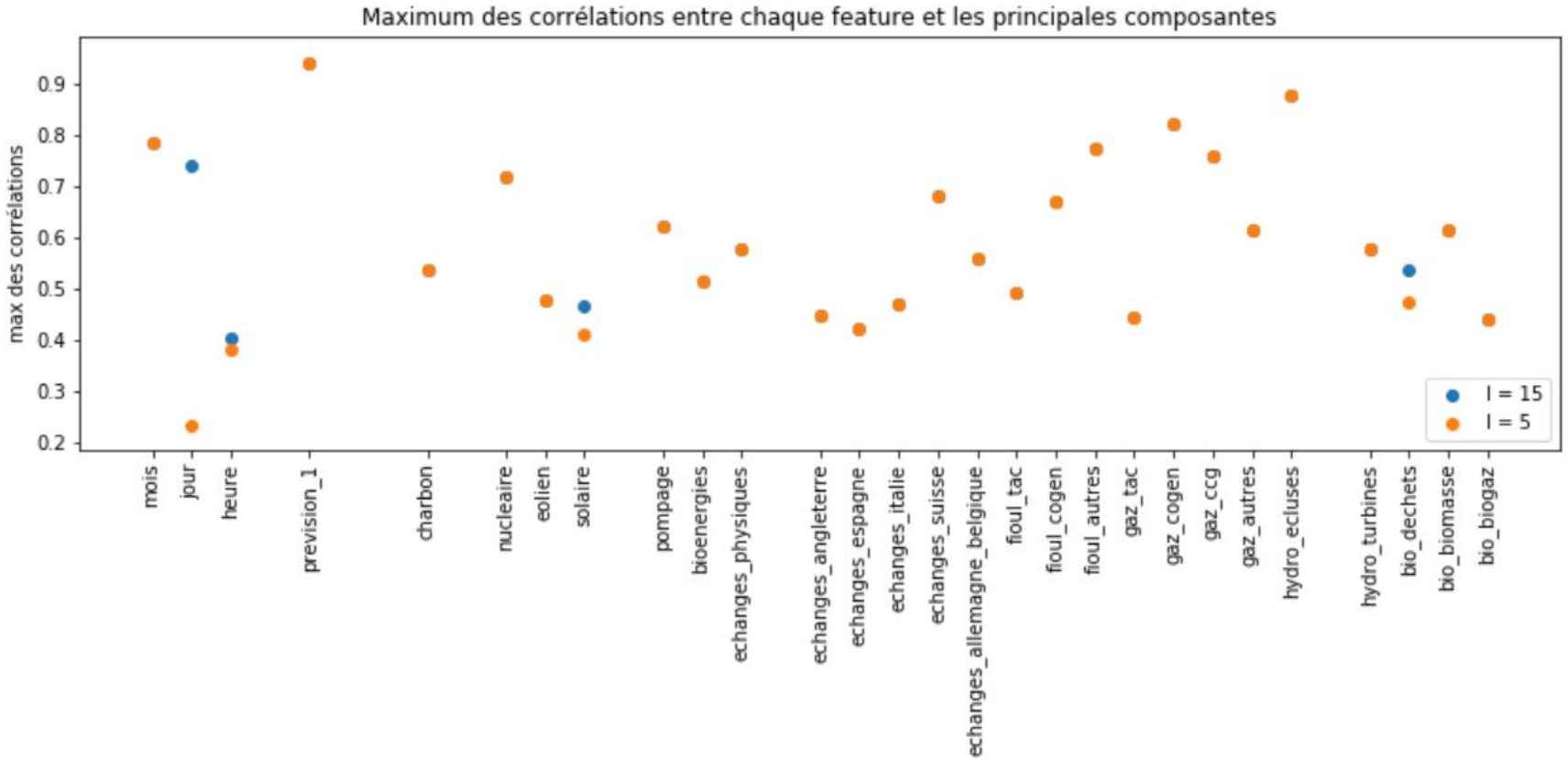
Analyse en composantes principales

Variables x_i, covariance K, composantes orthogonales c et valeurs propres λ associées aux vecteurs e. Corrélation entre variable et composante principale :

$$\text{Corr}(c^l, x^j) = \frac{\sqrt{\lambda_l} e_l^j}{K_{ij}} \tag{2}$$

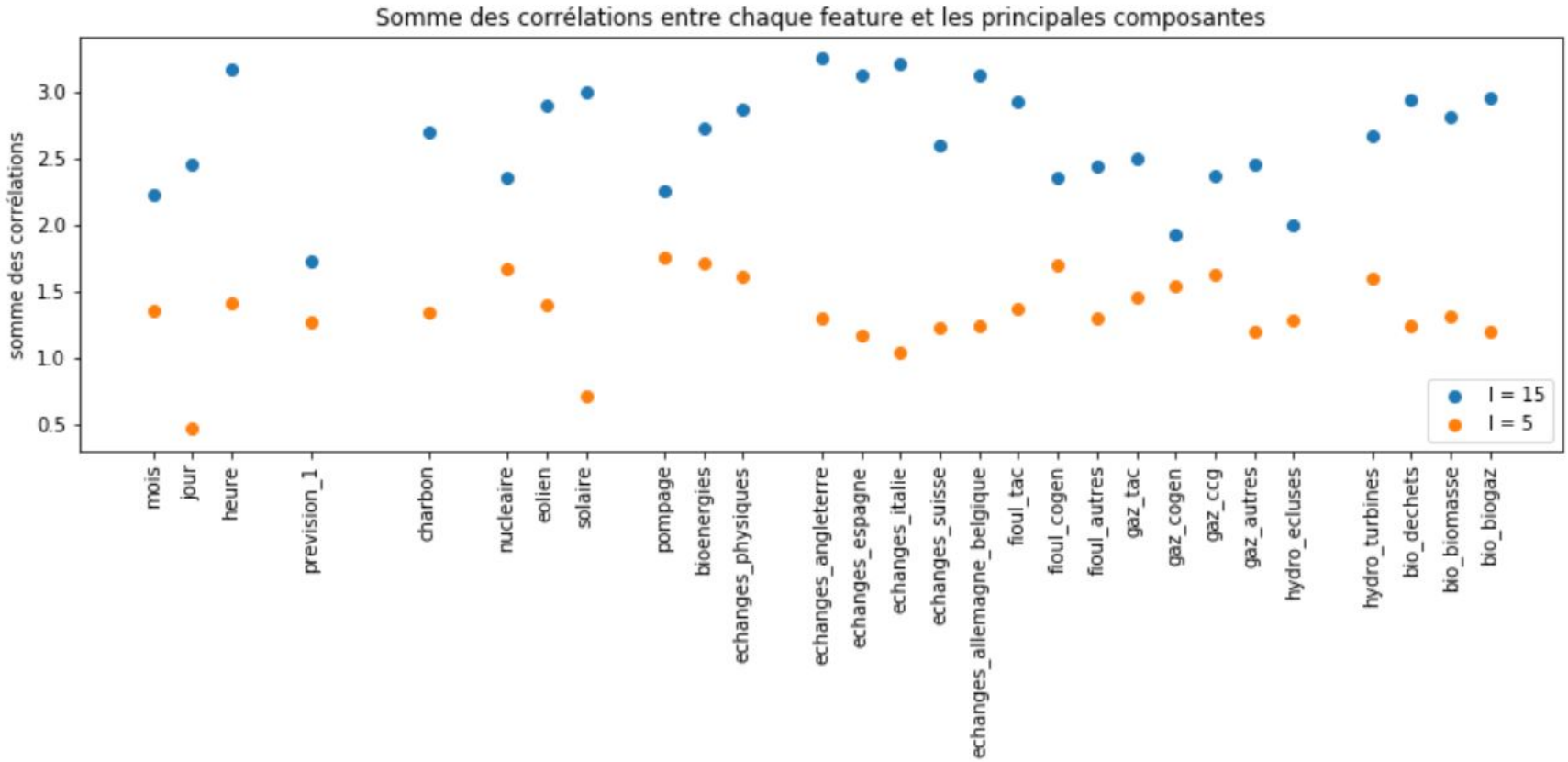
PCA

On observe ces corrélations avec 5 puis 15 composantes principales, en prenant le maximum pour chaque variable :



PCA

Puis la somme des corrélations par variable :

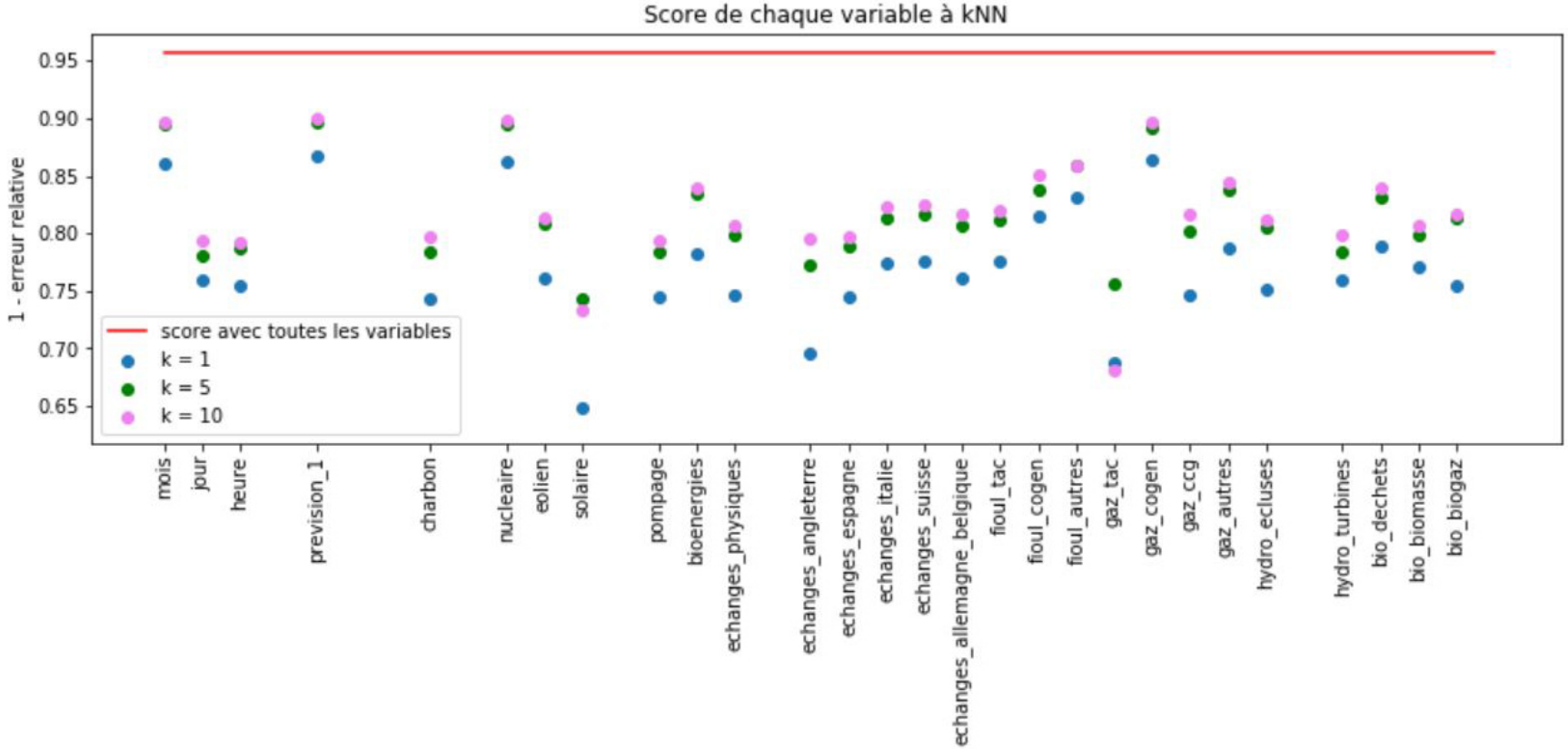


PCA

- La consommation a un maximum très élevé pour une somme faible : elle est presque à elle seule une composante principale
- `heure` `hydro_ecluses` et `solaire` sont réparties sur plusieurs composantes
- `heure` `solaire` et `echanges_*` disparaissent en restreignant le nombre de composantes.

k Nearest Neighbors

On lance
k-NN avec k=1, 5 puis 10, en mesurant le conjugué de l'erreur relative.

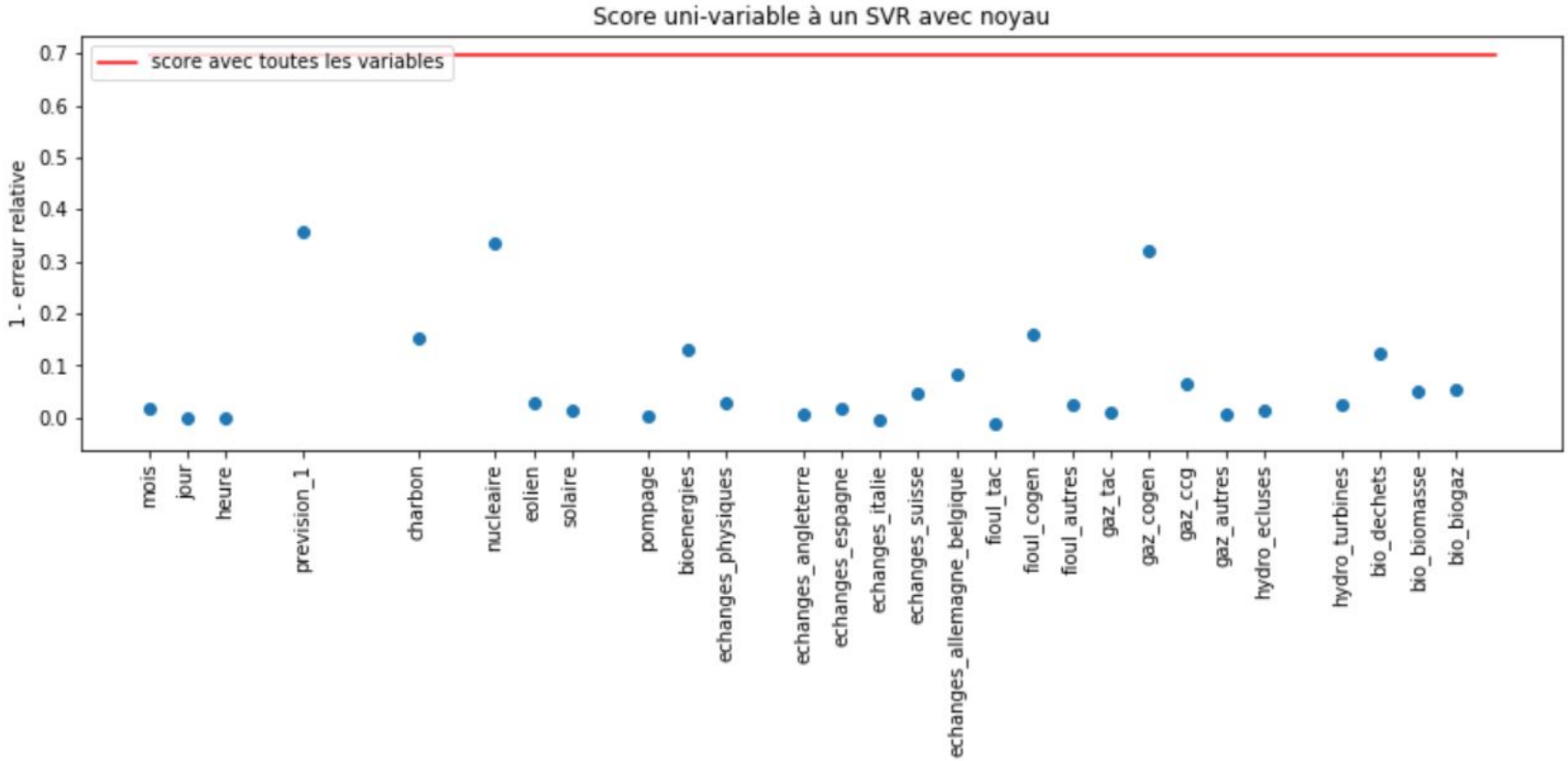


k-NN

- `mois`, `prevision_1`, `nucléaire` et `gaz_cogen` sont performantes, même avec peu de voisins
- `solaire` et `gaz_tac` sont peu performantes
- Les autres dans une bande moyennée entre 0.75 et 0.9 : non concluant

Support Vector Regressor

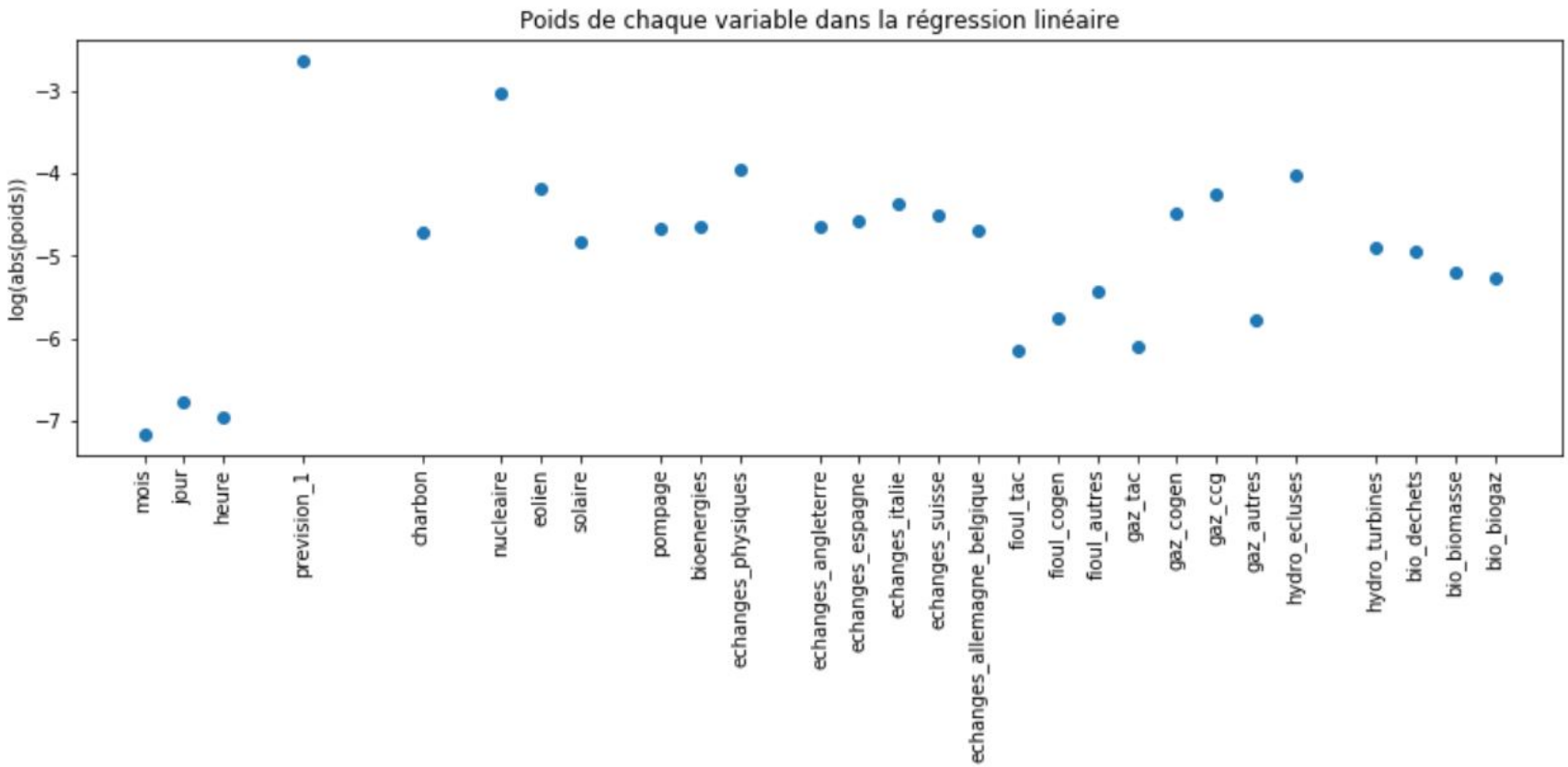
Noyau gaussien de paramètre 0.001 et constante de tradeoff 1, entraîné sur 10 époques et 100 observations.



SVR

- On retrouve `prevision_1`, `nucleaire` et `gaz_cogen` performants
- Quelques R^2 négatifs : susceptibles de fausser les prédictions

Poids de la régression linéaire



Les différences de poids peuvent être dues aux différences d'échelle

Test de student sur les poids

- Hypothèse nulle pour chaque poids : il est nul (donc variable inutile)
- Test à 5% : élimination de la moitié des variables (sans compter les doublons qui restent à éliminer)

Section 5

Conclusion

Difficultés

- différentes installation python : contournement avec jupyter notebook, organisation du code
- factorisation difficile : beaucoup de paramètres entrent en jeu
- travail à distance

Résultats

- Élimination de nombreuses variables inutiles ou redondantes, en lien avec intuitions
- Résultats très satisfaisants avec certains modèles

Variables conservées	R^2 *de la prédiction obtenue*
Toutes	0.865
Test de Student	0.858
Test de Student inverse	0.501
Doublons	0.849
Toutes méthodes considérées	0.833

Pour aller plus loin

- Évolution de la relation entre variables explicatives potentielles et pertes
- Transformation préalable des variables (périodicité notamment)

Remerciements

- Aboubakr MACHRAFI (stagiaire RTE)
- Valentin CADORET, Virginie DORDONNAT (RTE)
- Gabriel STOLTZ (ENPC)
- David PICARD (ENPC)