Fouille de données : Etude et analyse de thèse

Nabil Redha BELKHOUS

Numéro d'étudiant : 16705491

Email : nbelkhous@gmail.com

Université Paris 8 Vincennes-Saint-Denis
UFR : MIASHS
Master 2 Big Data et fouille de données

Année Universitaire : 2016 - 2017

Sommaire

Choix de la thèse : thèse cifre

Méthodes de fouilles de données pour la prédiction de l'évolution du prix d'un billet et application au conseil à l'achat en ligne

Présentée et soutenue publiquement par

Till WOHLFARTH

Introduction

Collecte de données

Sources de données

- Sites d'agence de voyages (Expedia, GoVoyages, ...)
- Sites de compagnies aériennes (Air France, EasyJet, ...)

Informations exploités

- Recherches utilisateurs
- Résultats de recherches
- Alertes

Structure de données

Structure de la base de données

- Vol :
  - aéroport de départ, aéroport d'arrivée
  - horaire de départ et de retour
  - un code transporteur aller et retour
- Séries de prix
  - vol, vendeur et prix
- Attributs
  - attributs définissant un vol
  - attributs dérivés des attributs définissant un vol
  - attributs liés au site marchand
  - attributs contextuels qui évoluent avec les points de la série temporelle

Étude du comportement des prix

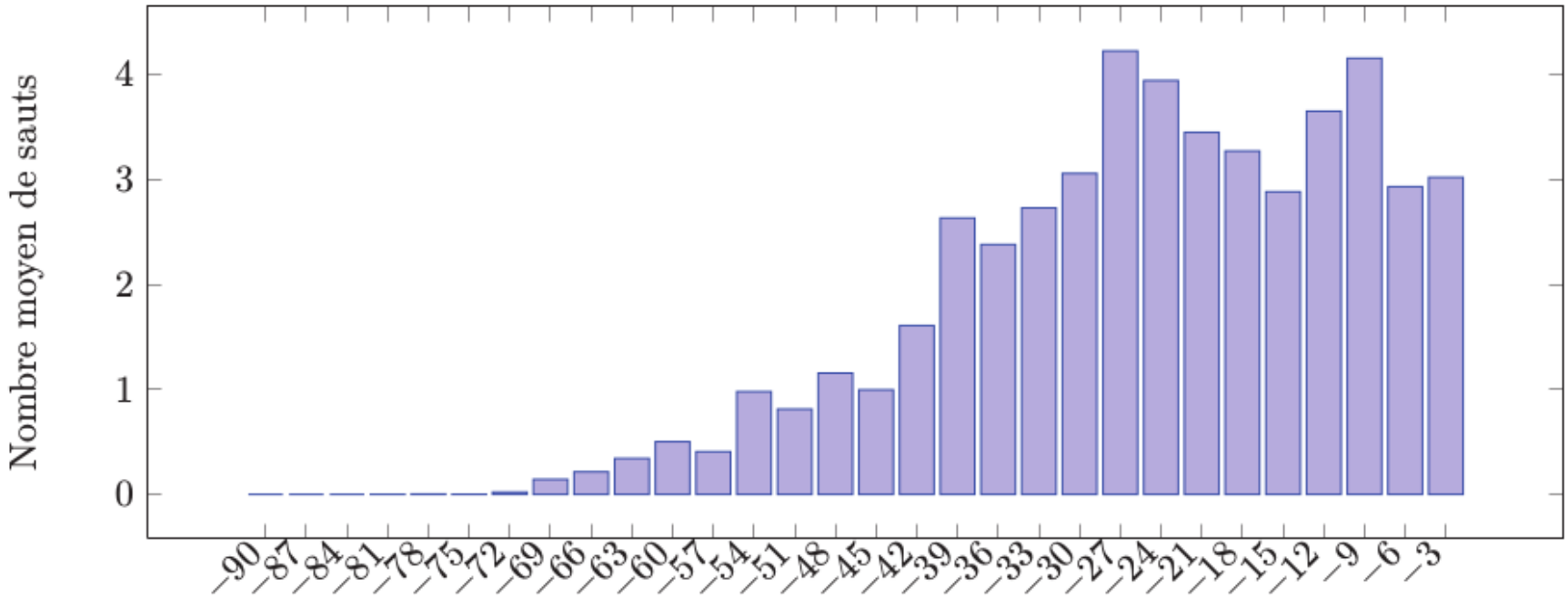


Figure: Nombre de sauts par rapport au nombre de jours avant le départ (tranches de 3 jours) **[1]**

Série temporelle de prix

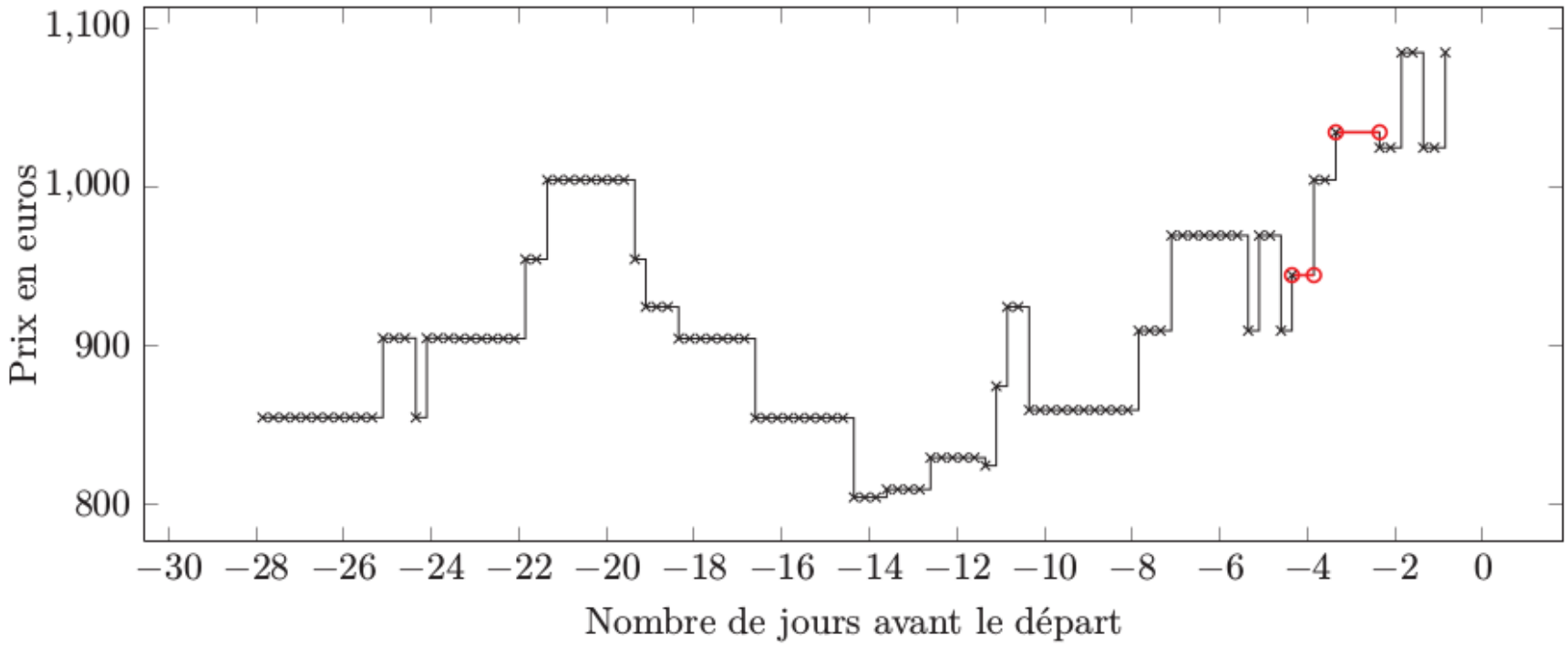


Figure: Time Series : Paris-Bangkok départ le 11/01/2013 pour 14 jours par Qatar **[1]**

Série de rendements

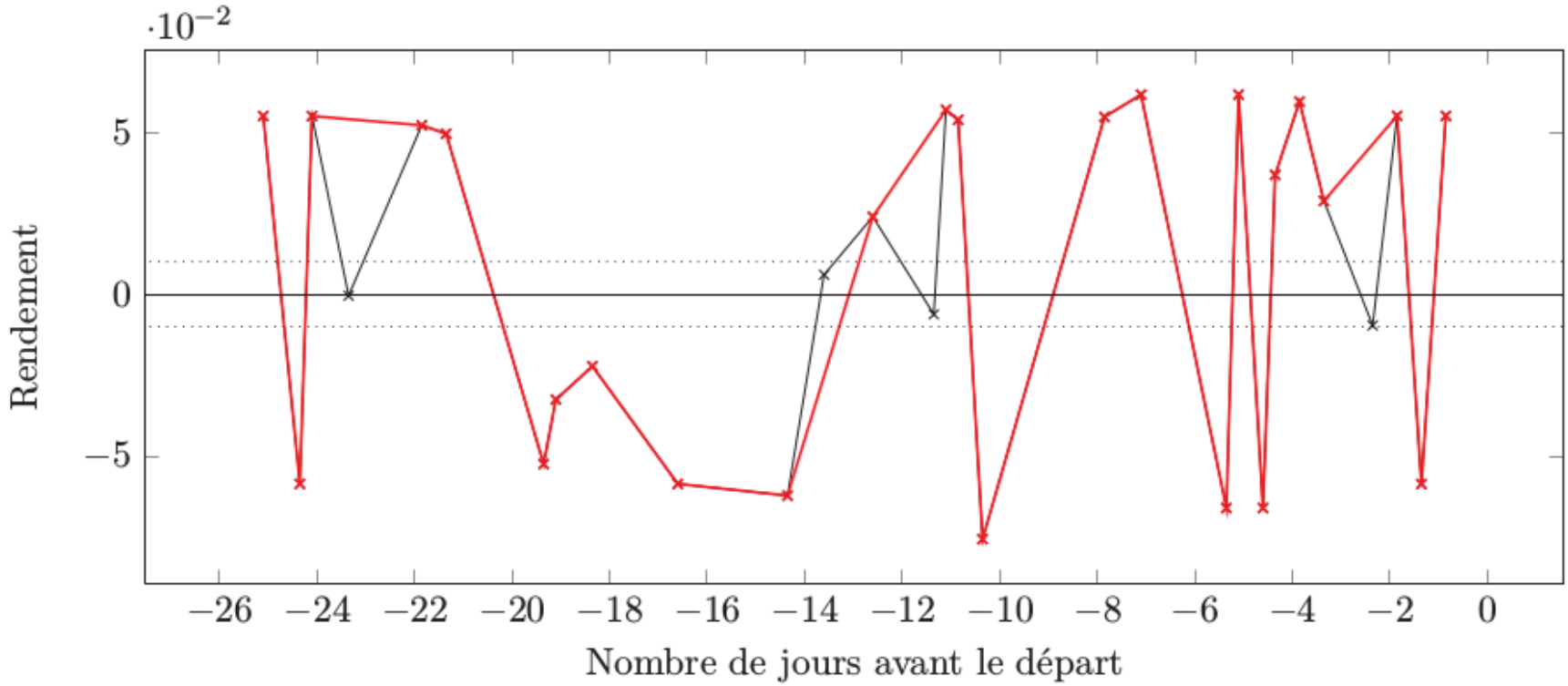


Figure: Série de rendements après filtre : Paris-Bangkok départ le 11/01/2013 pour 14 jours parQatar **[1]**

Niveau de gris

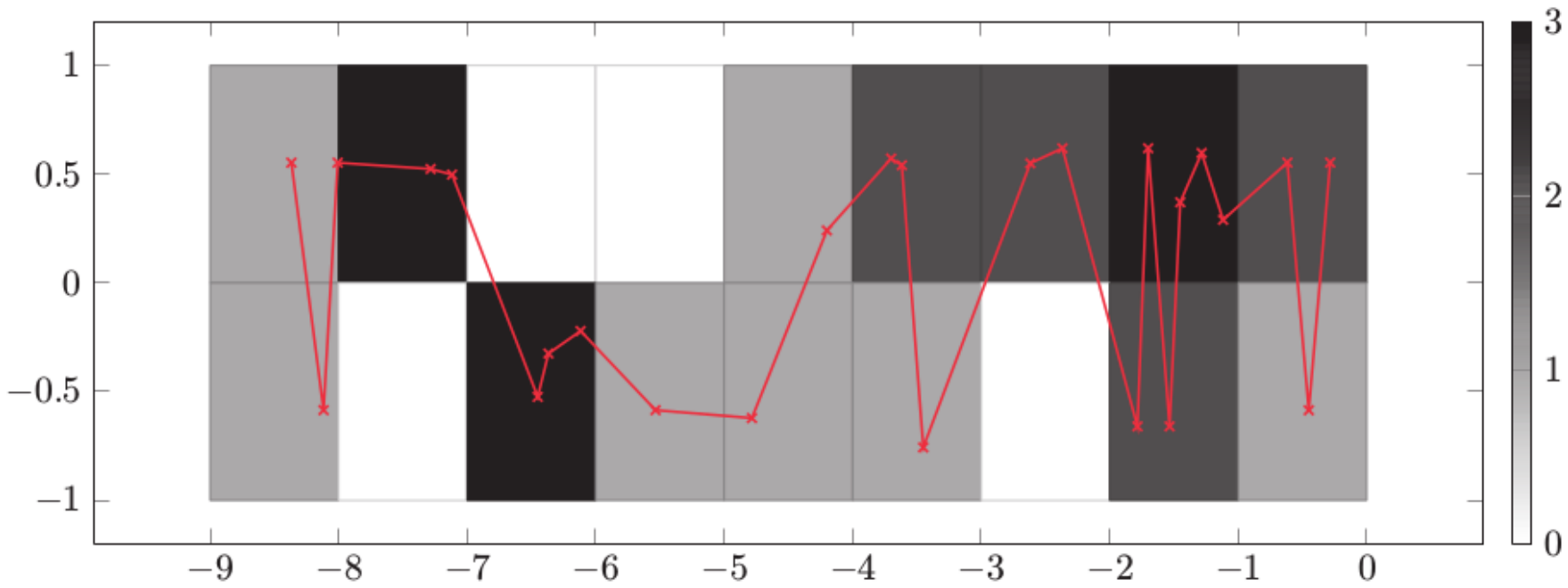


Figure: Niveau de gris après filtre : Paris-Bangkok départ le 11/01/2013 pour 14 jours par Qatar **[1]**

Clustering

1. K-means
2. Bagged K-means
3. EM

Classification

1. CART
2. C4.5
3. Adaboost
4. Forêts aléatoires

Résultat de segmentation

1. K-means

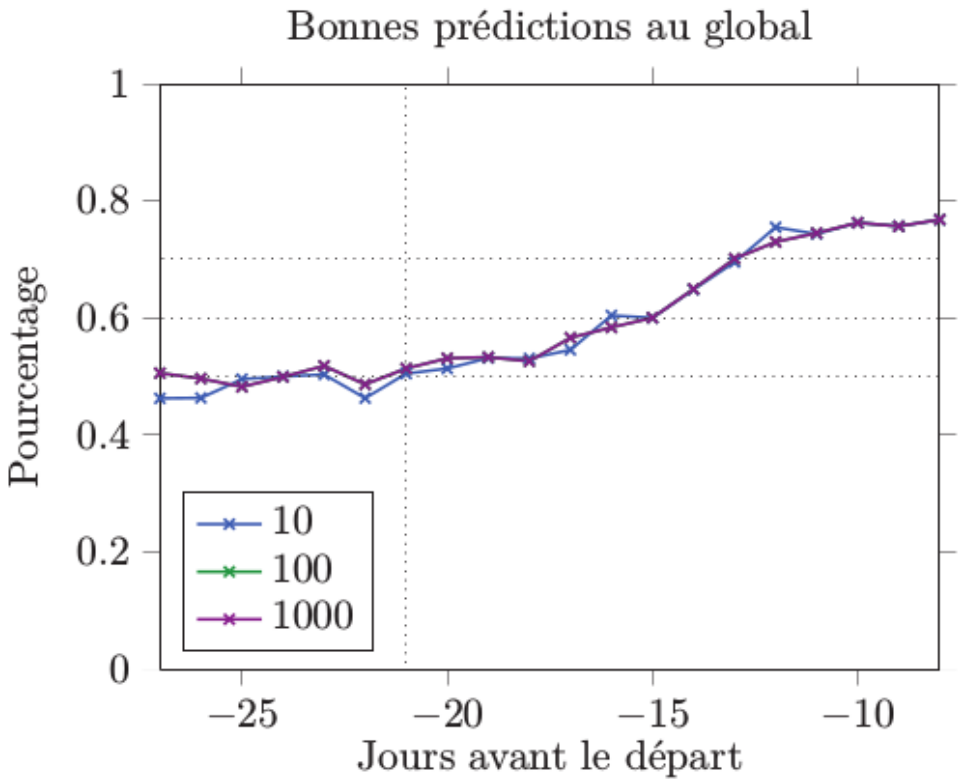


Figure: K-Means - Évolution du pourcentage de bonnes prédictions à -21 jours du départ pour différents nstart **[1]**

Résultat de segmentation

1. K-means

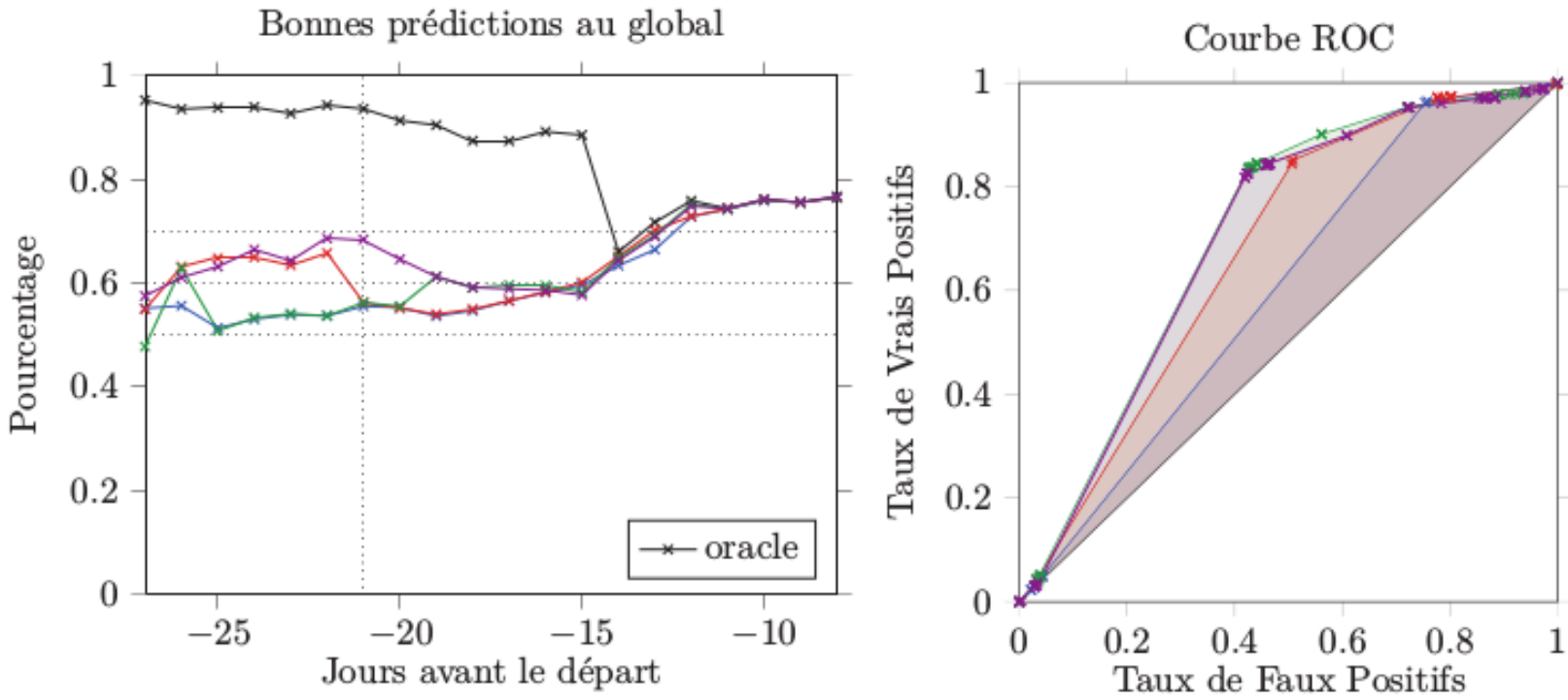


Figure: K-Means - Évolution du taux de bonne prédiction à -21 jours avant le départ pour différents nombres de clusters **[1]**

Résultat de segmentation

1. K-means

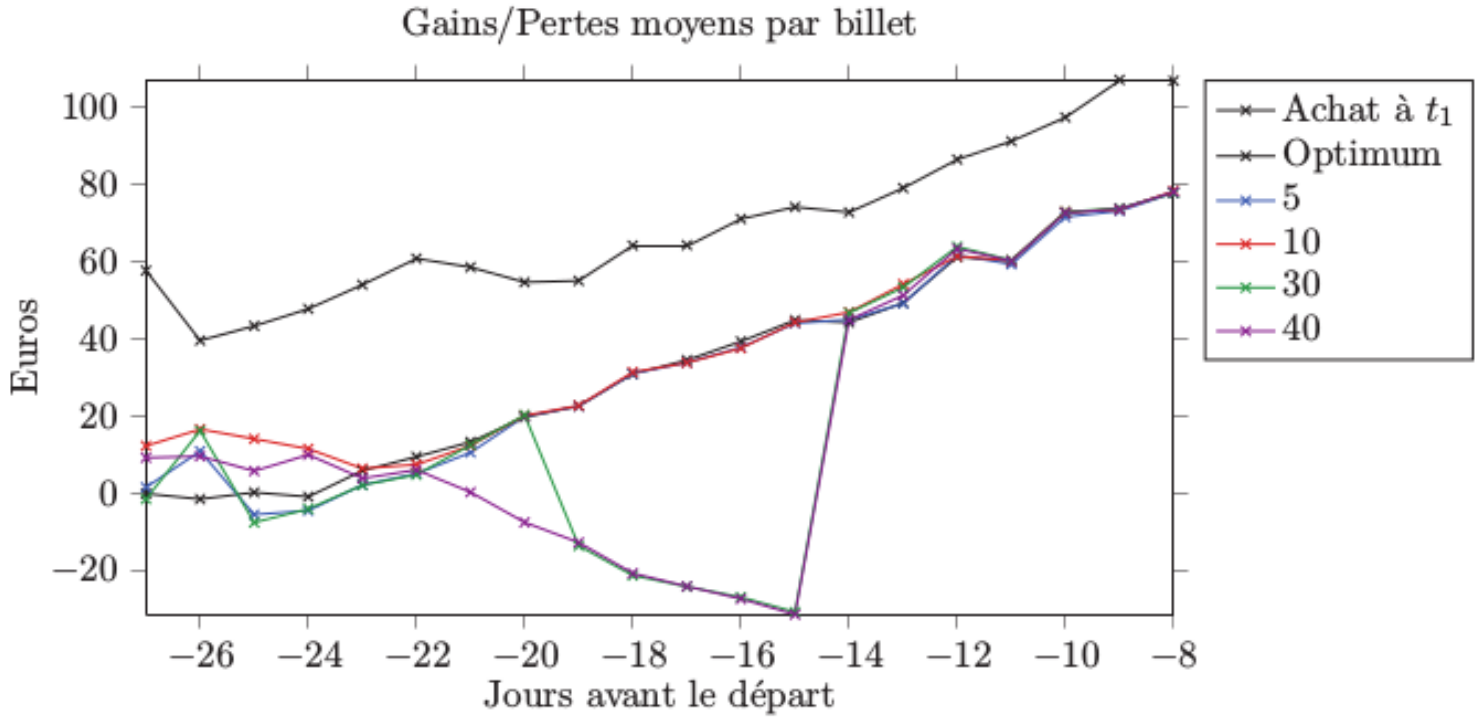


Figure: K-Means - Évolution des Gains/Pertes à -21 jours avant le départ pour différents nombres de clusters **[1]**

Résultat de segmentation

2. Bagged K-means

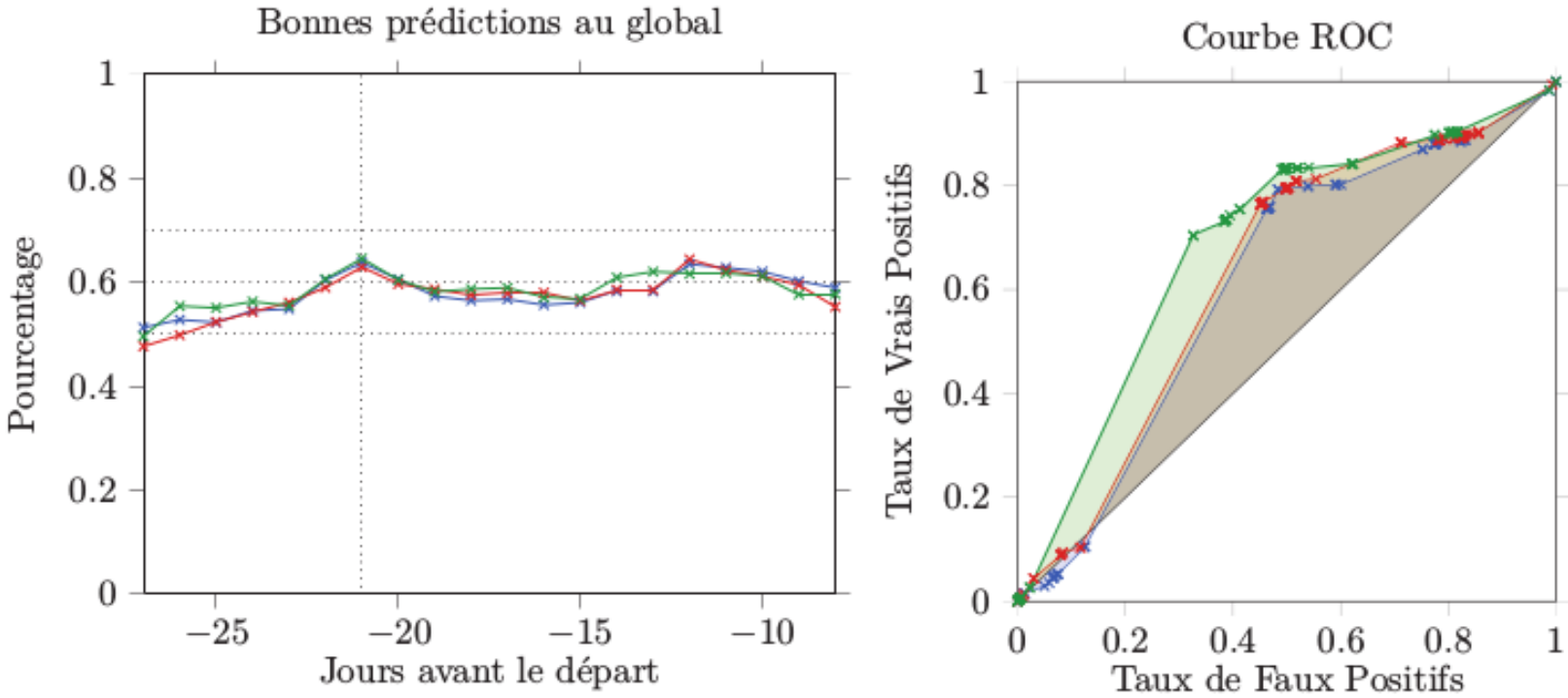


Figure: Comparaison du taux de bonne prédiction entre Bagged K-means et K-means **[1]**

Résultat de segmentation

2. Bagged K-means

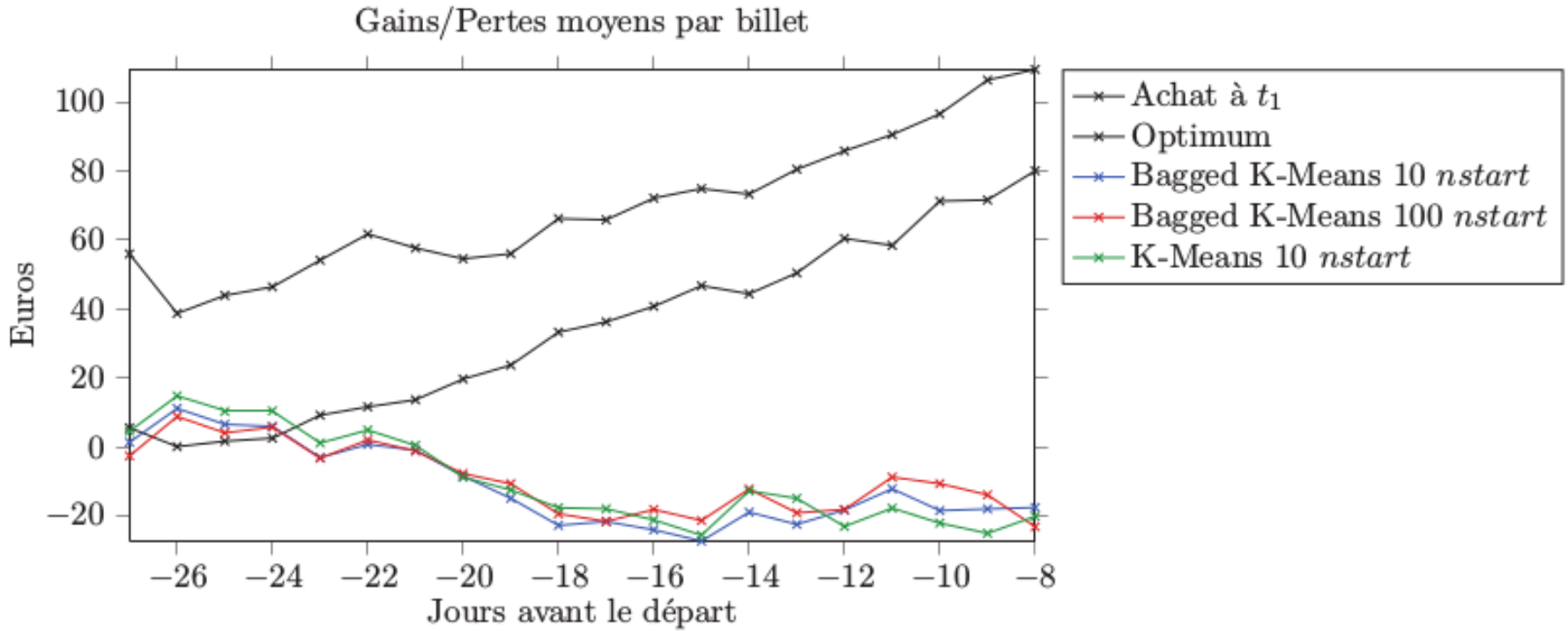


Figure: Comparaison des Gains/Pertes entre Bagged K-means et K-means **[1]**

Résultat de segmentation

❸ EM

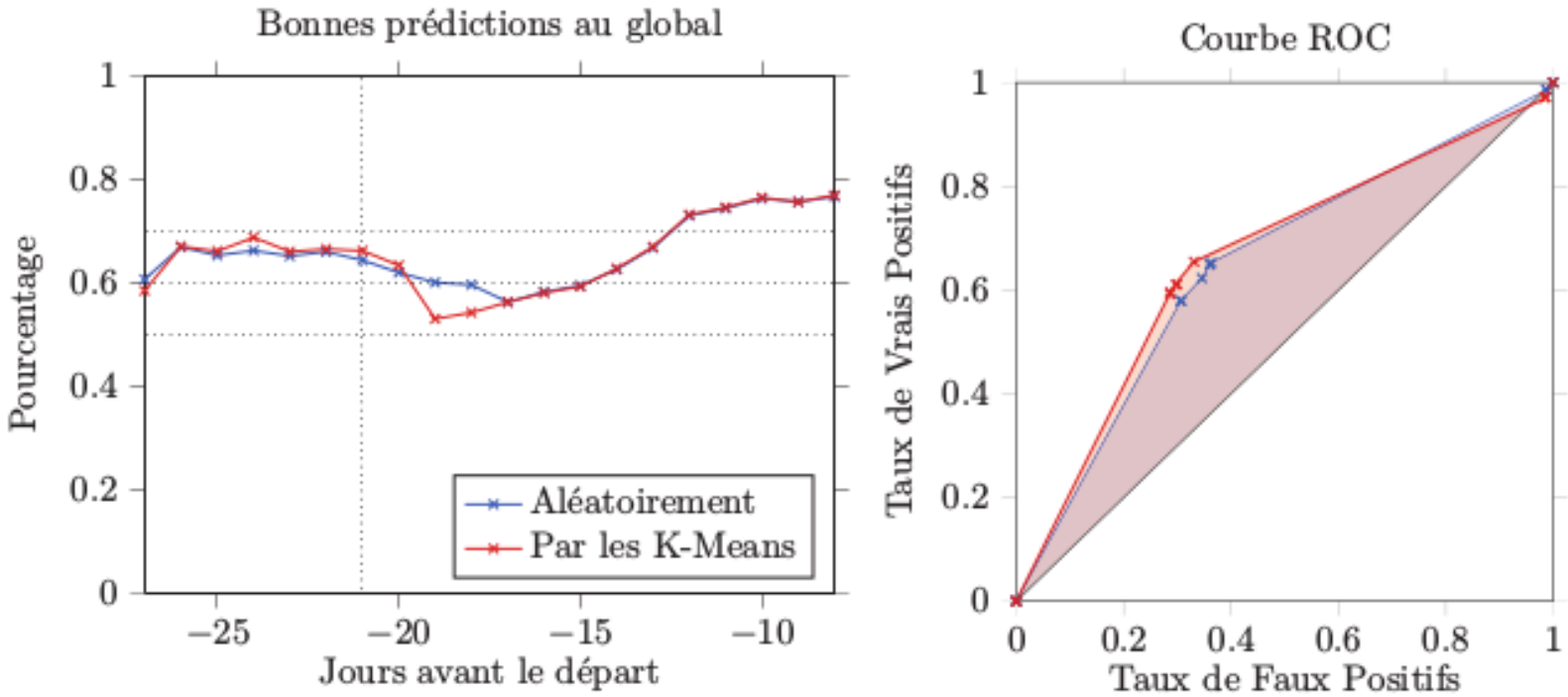


Figure: Comparaison des performances de l'algorithme EM en fonction de son type d'initialisation **[1]**

Résultat de segmentation

4. Comparatif

Méthode 1	Nb groupes	Taux BP	Economie/t_1
K-means	5	62%	-1
EM	5	66%	4
K-means	10	64%	1
EM	10	62%	-3
K-means	30	62%	-7
EM	30	64%	1
K-means	40	66%	-18
EM	40	64%	1

Figure: Comparatif **[1]**

Résultat de classification

1 CART

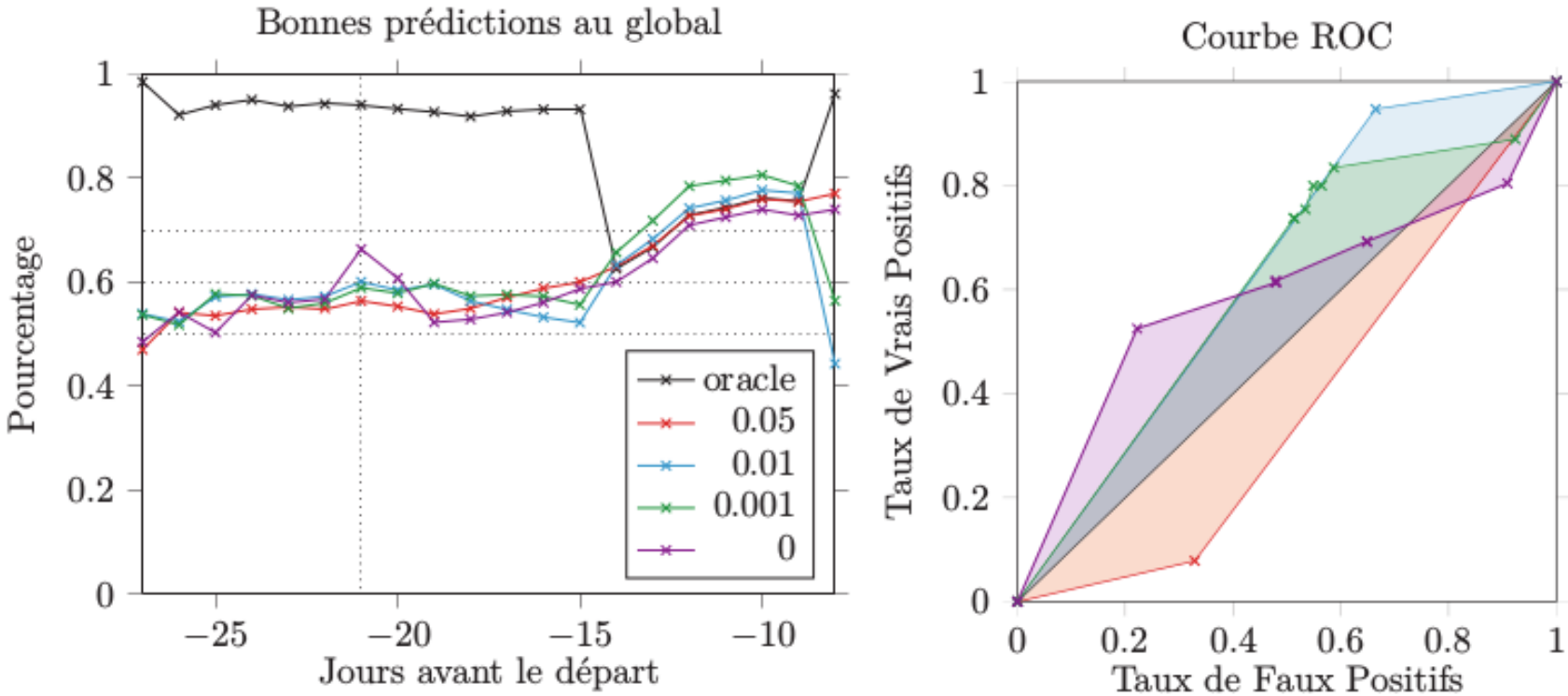


Figure: Évolution du pourcentage de bonnes prédictions à -21 jours de la date de départ pour différents élagages **[1]**

Résultat de classification

2. C4.5

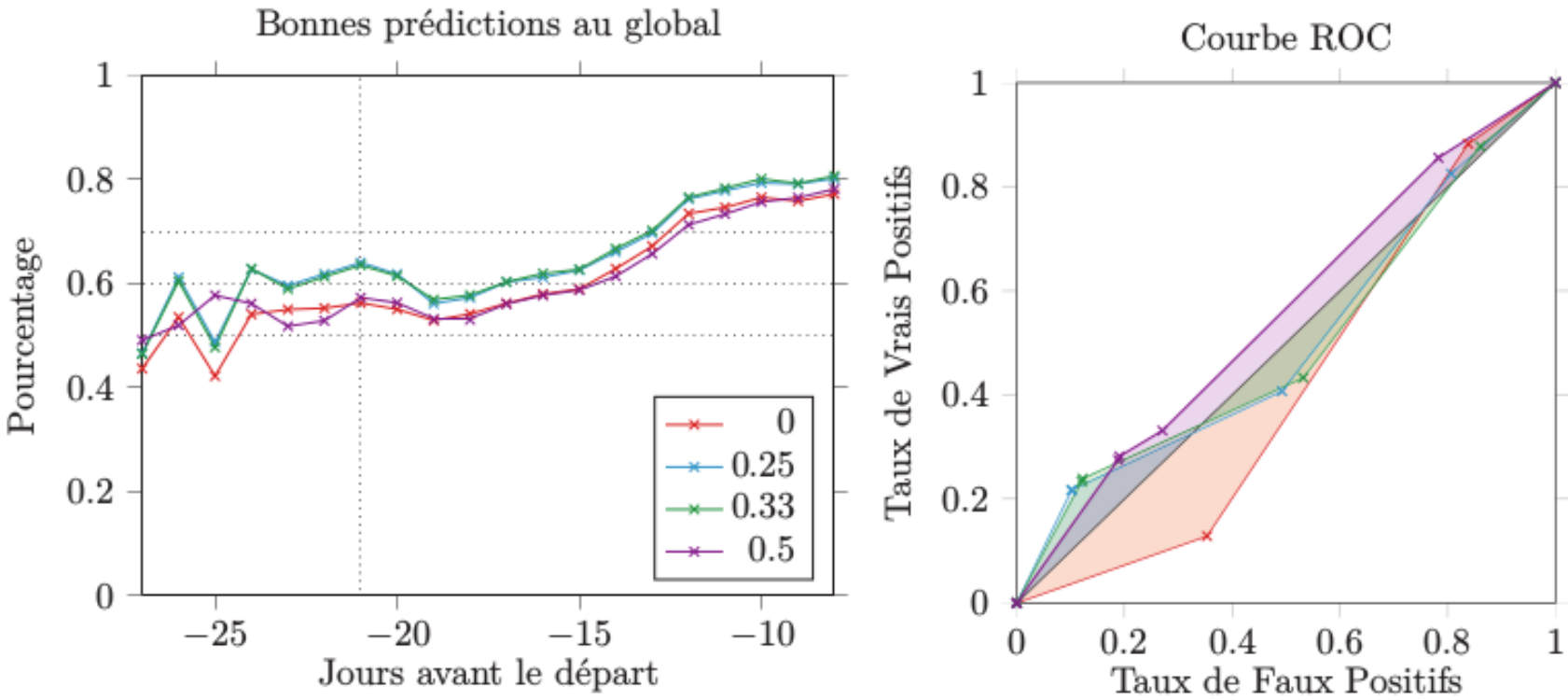


Figure: Évolution du pourcentage de bonnes prédictions à -21 jours de la date de départ pour différents élagages **[1]**

Résultat de classification

3. Forêts aléatoires

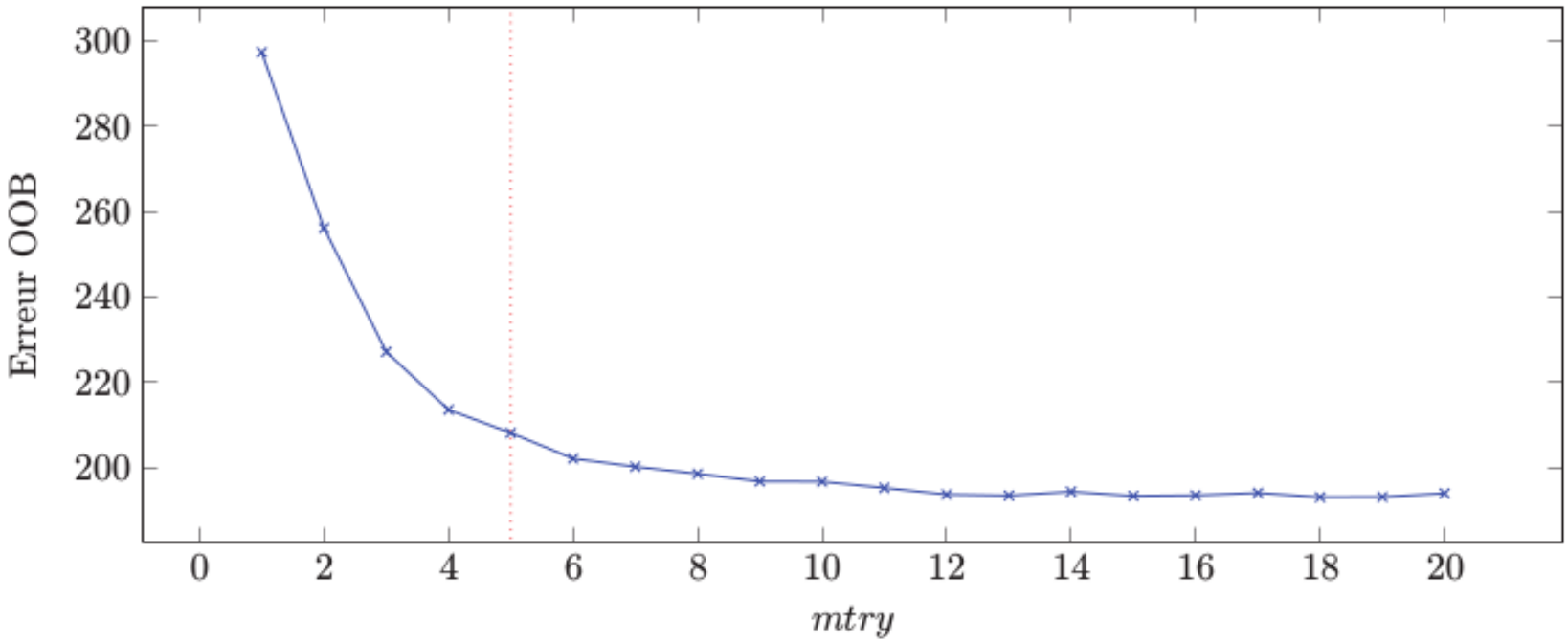


Figure: Évolution du taux d'erreur OOB en fonction du paramètre mtry **[1]**

Résultat de classification

4 Adaboost

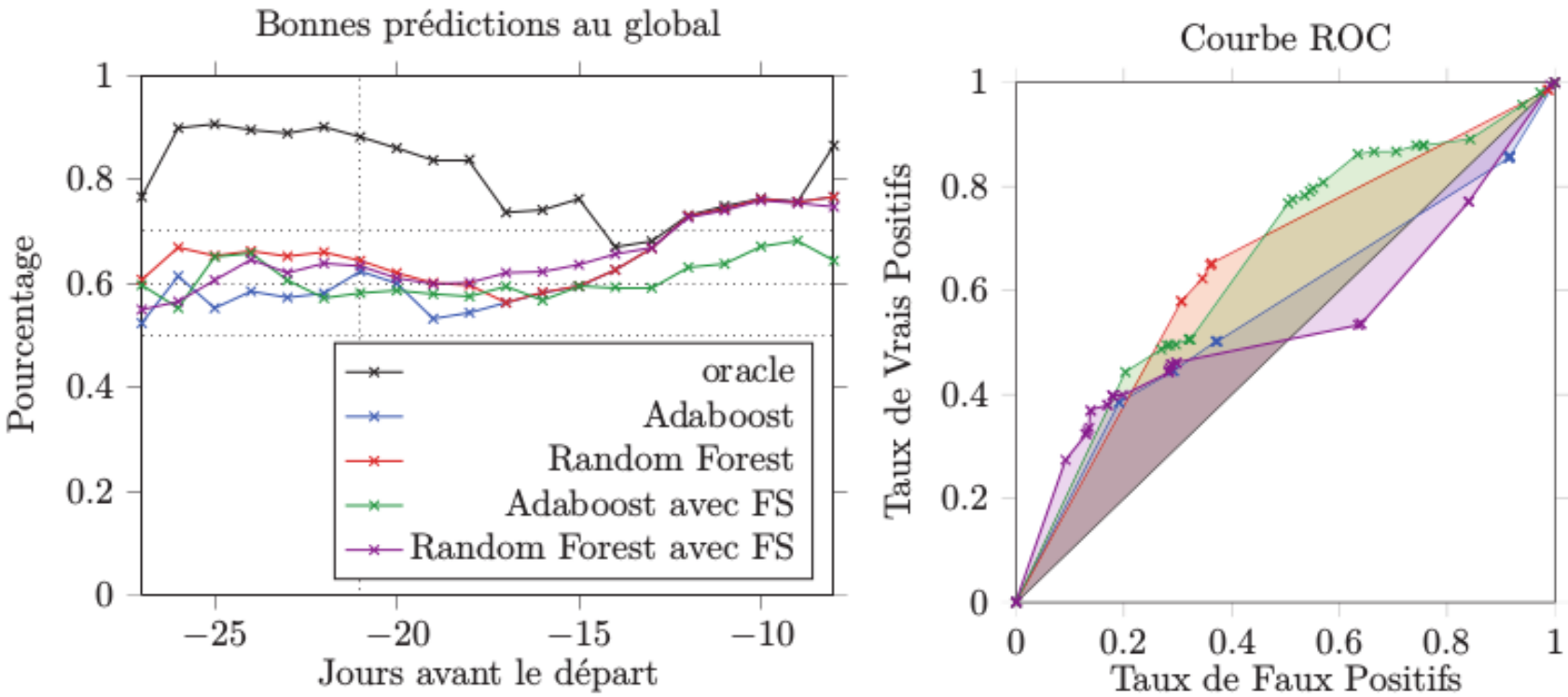


Figure: Comparaison entre Adaboost et Random Forests **[1]**

Méthodes retenues

Comparatif entre les méthodes :

Méthode 1	Nb groupes	Taux BP	Economie/t_1
RF	5	66%	4
RF	30	64%	1
Adaboost	5	63%	0
CART	5	62%	0
CART	30	62%	-3
C4.5	5	66%	-2
C4.5	30	61%	-7

Figure: Comparatif **[1]**

Méthodes retenues

1. Segmentation : EM (K-means avec 5 groupe et 100 nstar)
2. Classification : Forêts aléatoires (12 attributes et 100 arbres)

Évolution des performances dans le temps

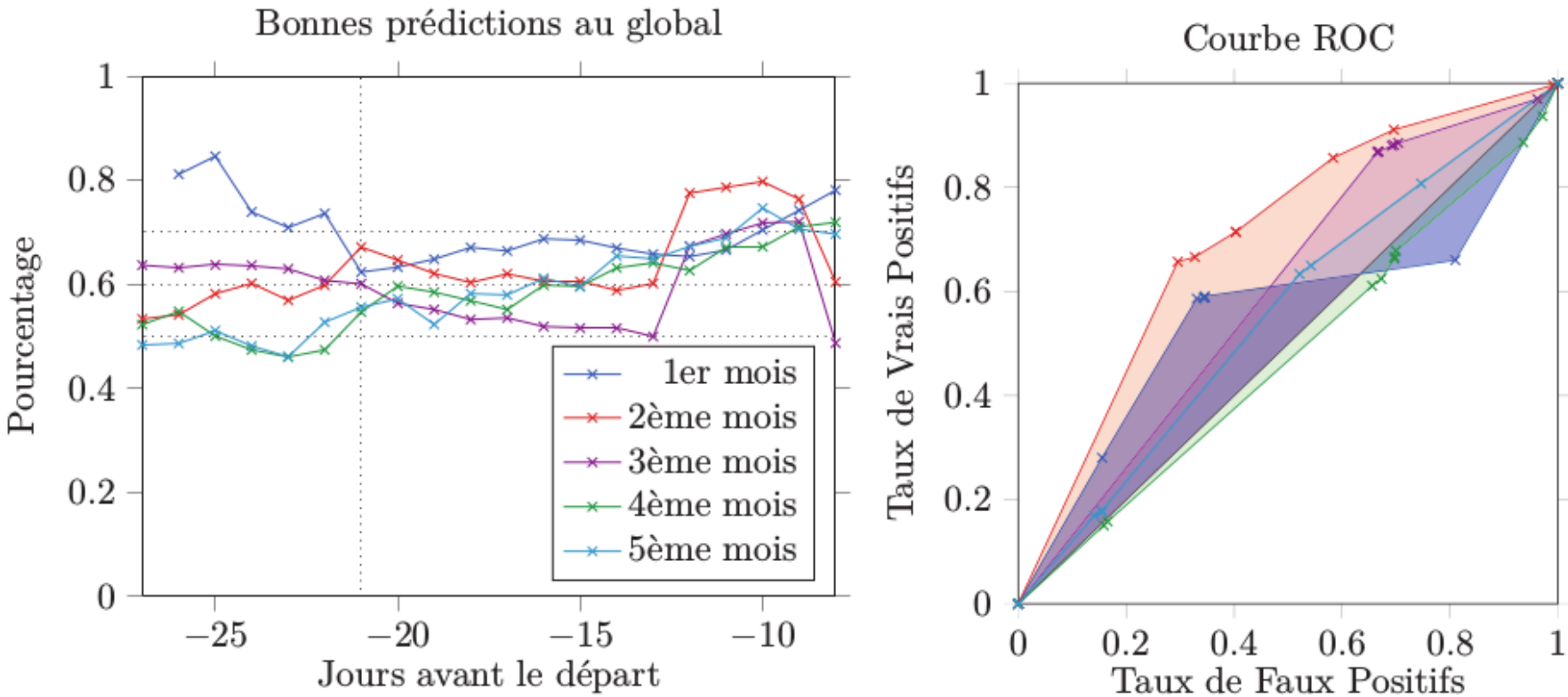
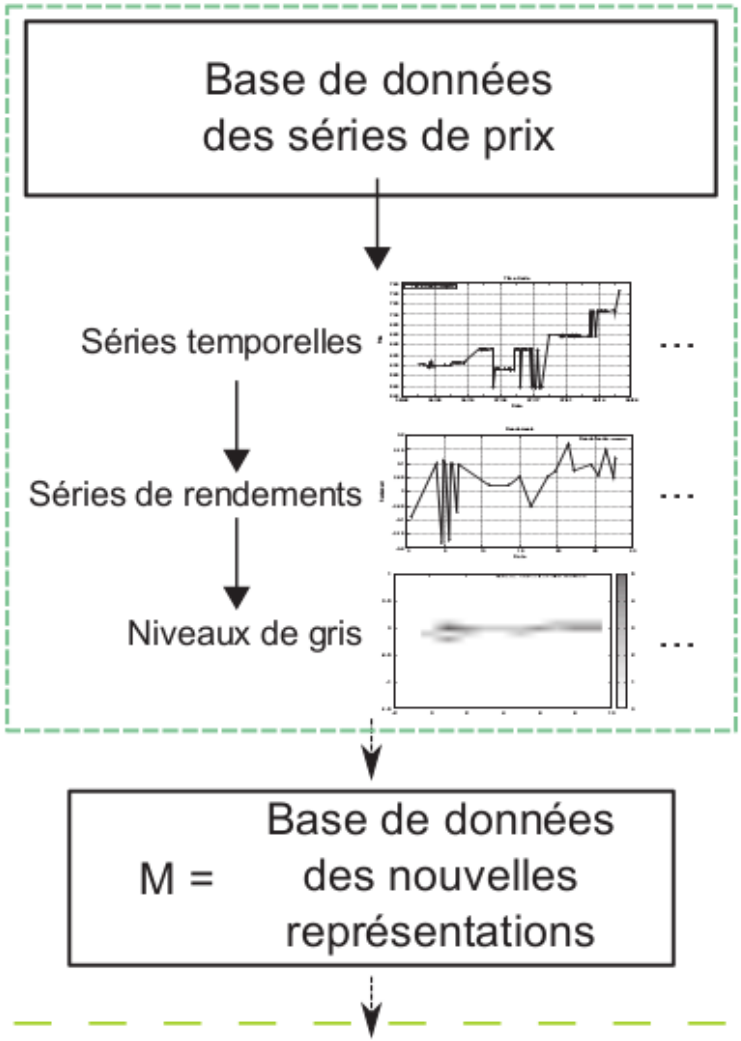
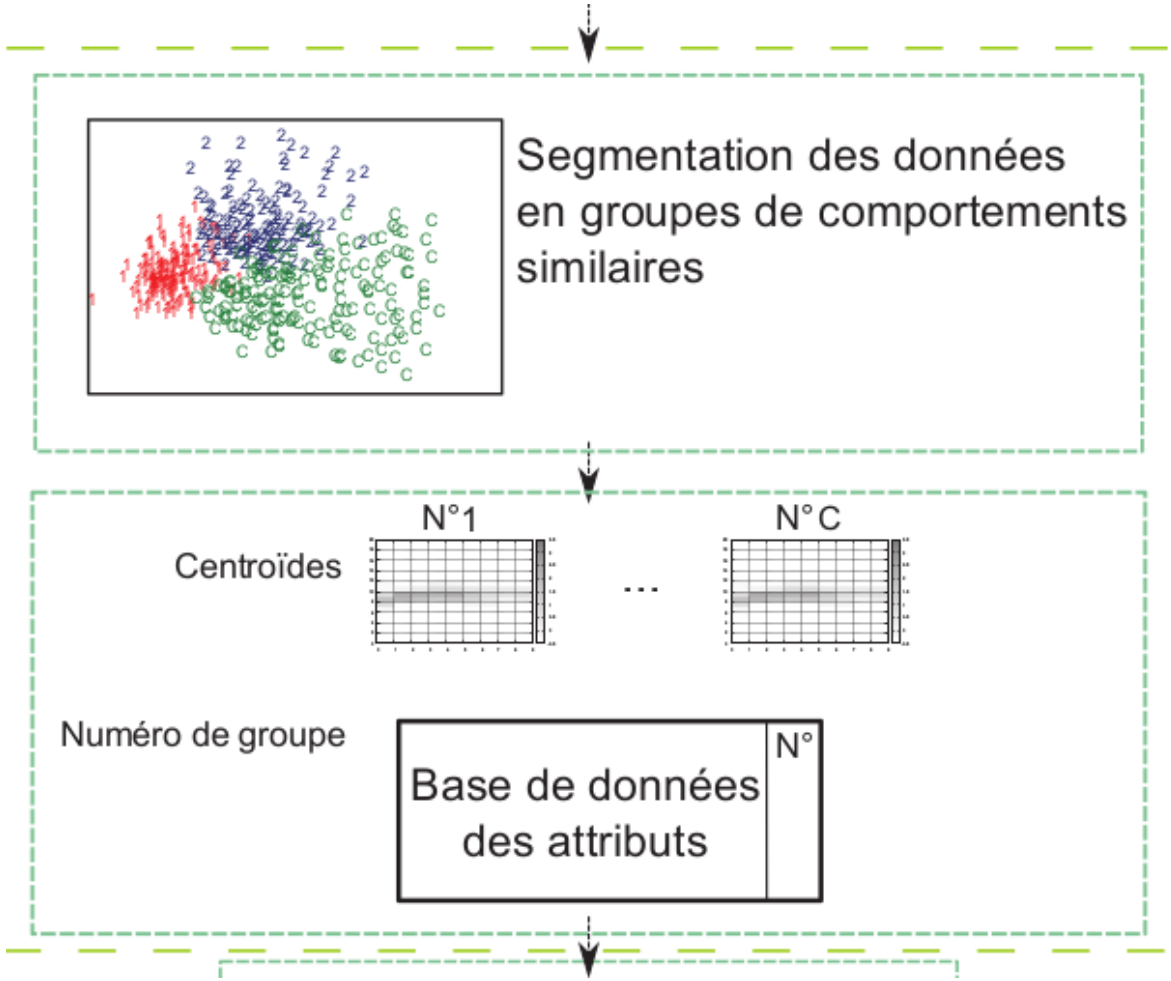


Figure: Évolution du taux de bonnes prédictions en fonction du temps **[1]**

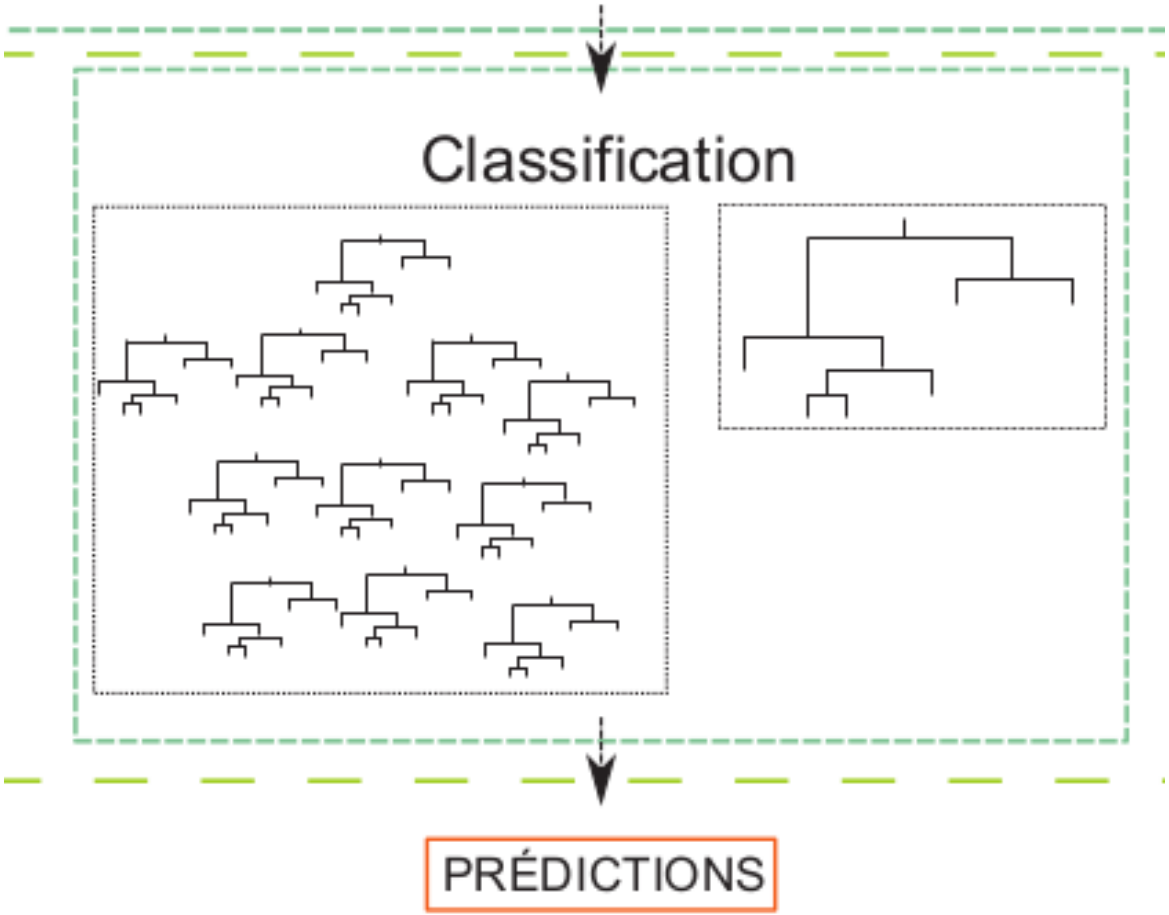
1. Préparation des données



2. Segmentation des données



3. Classification et prédiction



[1] Till WOHLFARTH.
Méthodes de fouilles de données pour la prédiction de l'évolution du prix d'un billet et application au conseil à l'achat en ligne.
2013.

Questions

QUESTIONS ?