Présentation projet NF26

LE Tran Hoang Long & BRIZARD Clément

Université de Technologie de Compiègne

27 juin 2019

Overview

Introduction

- Finlande, 2005-2014
- 477 Mo
- Une seule station jusqu'en 2010
- Beaucoup de données nulles

Question 1 : Exploitation spatiale - Conception

- Conception BD : `PRIMARY KEY ((station, longitude, latitude), year, month, day)`
- Technologies de base de données : *Spark*, *Cassandra*
- Visualisation des données : *matplotlib*, *seaborn*
- Requête selon station / (longitude, latitude)
- Requête selon intervalle de temps / année spécifique

Question 1 : Exploitation spatiale - Visualisation

- Visualiser sur un intervalle d'années

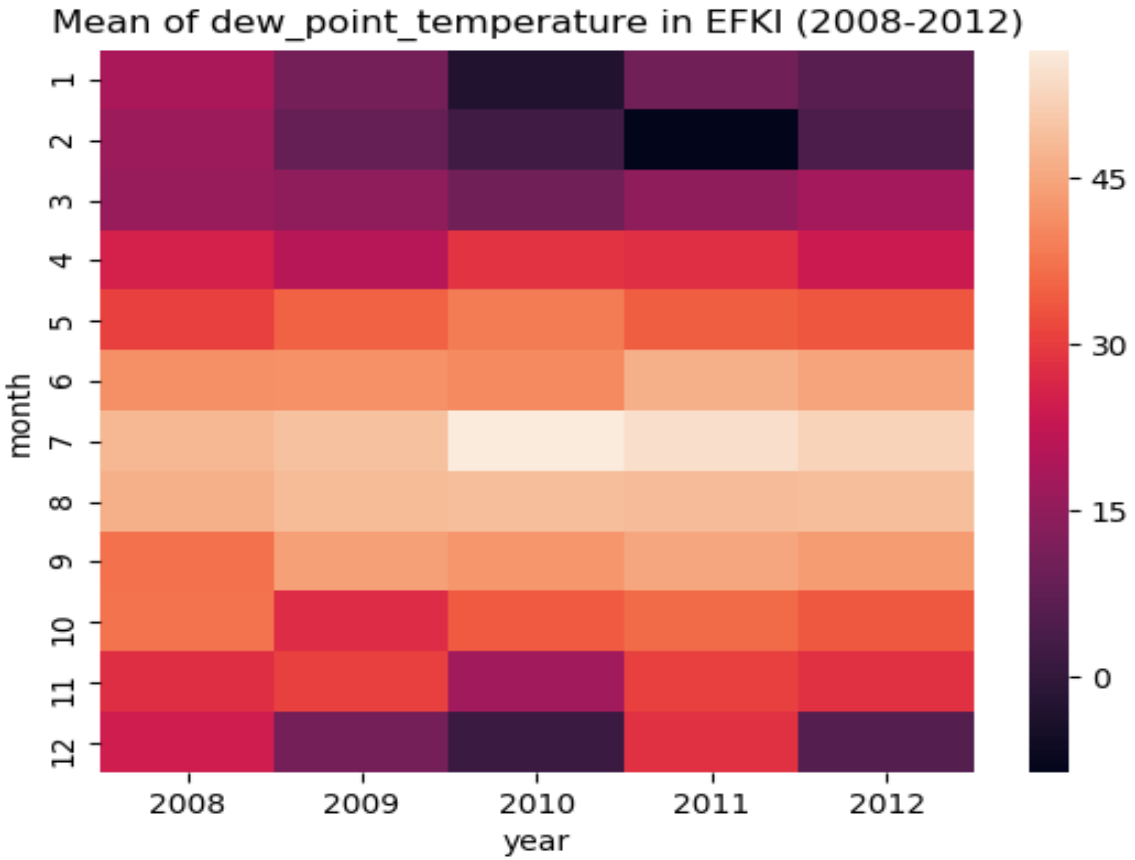

Figure – Point de rosée de la station EFKI entre 2008 et 2012

Question 1 : Exploitation spatiale - Visualisation

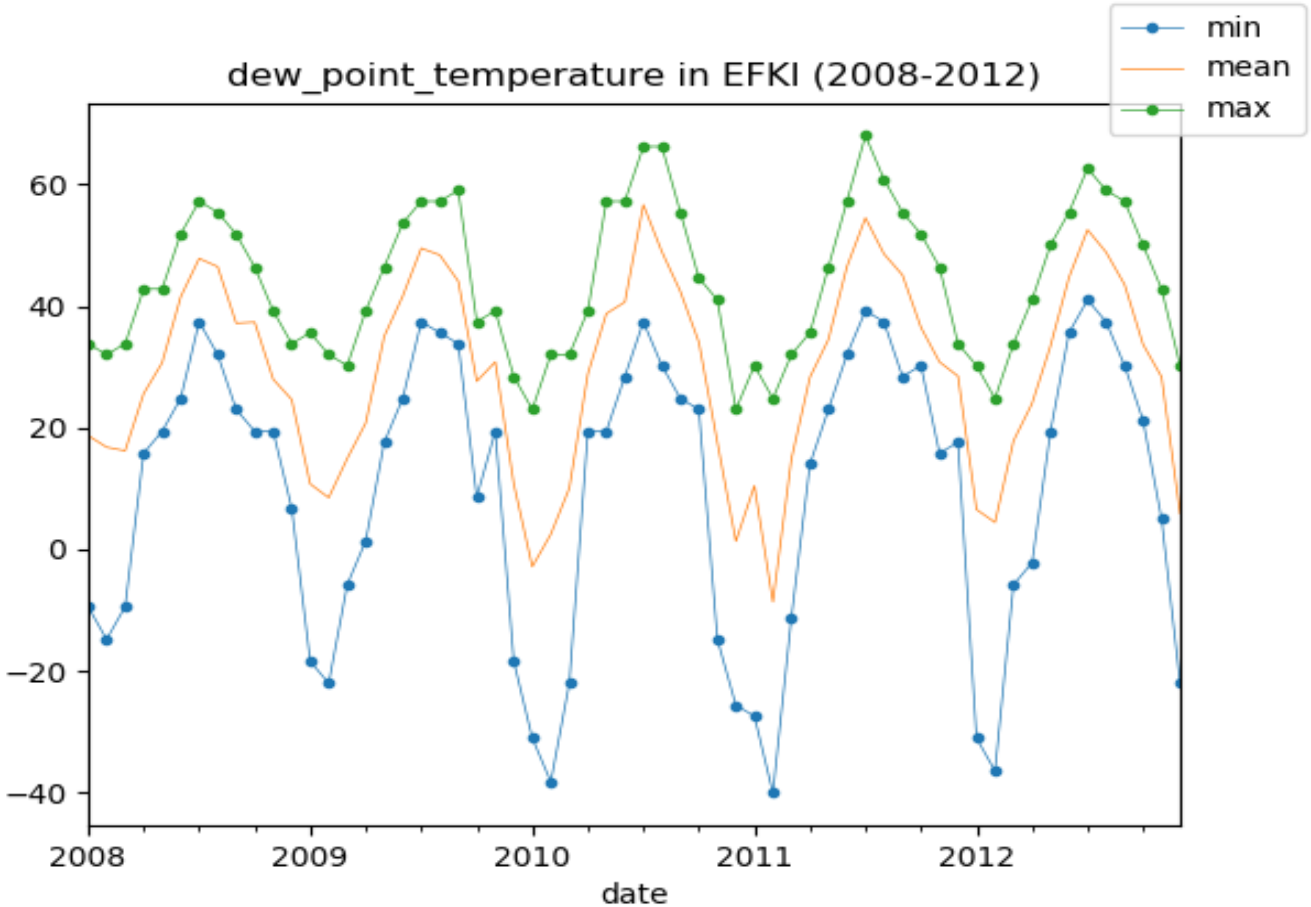

Figure – Min, max et moyenne du point de rosée de EFKI entre 2008 et 2012

Question 1 : Exploitation spatiale - Visualisation

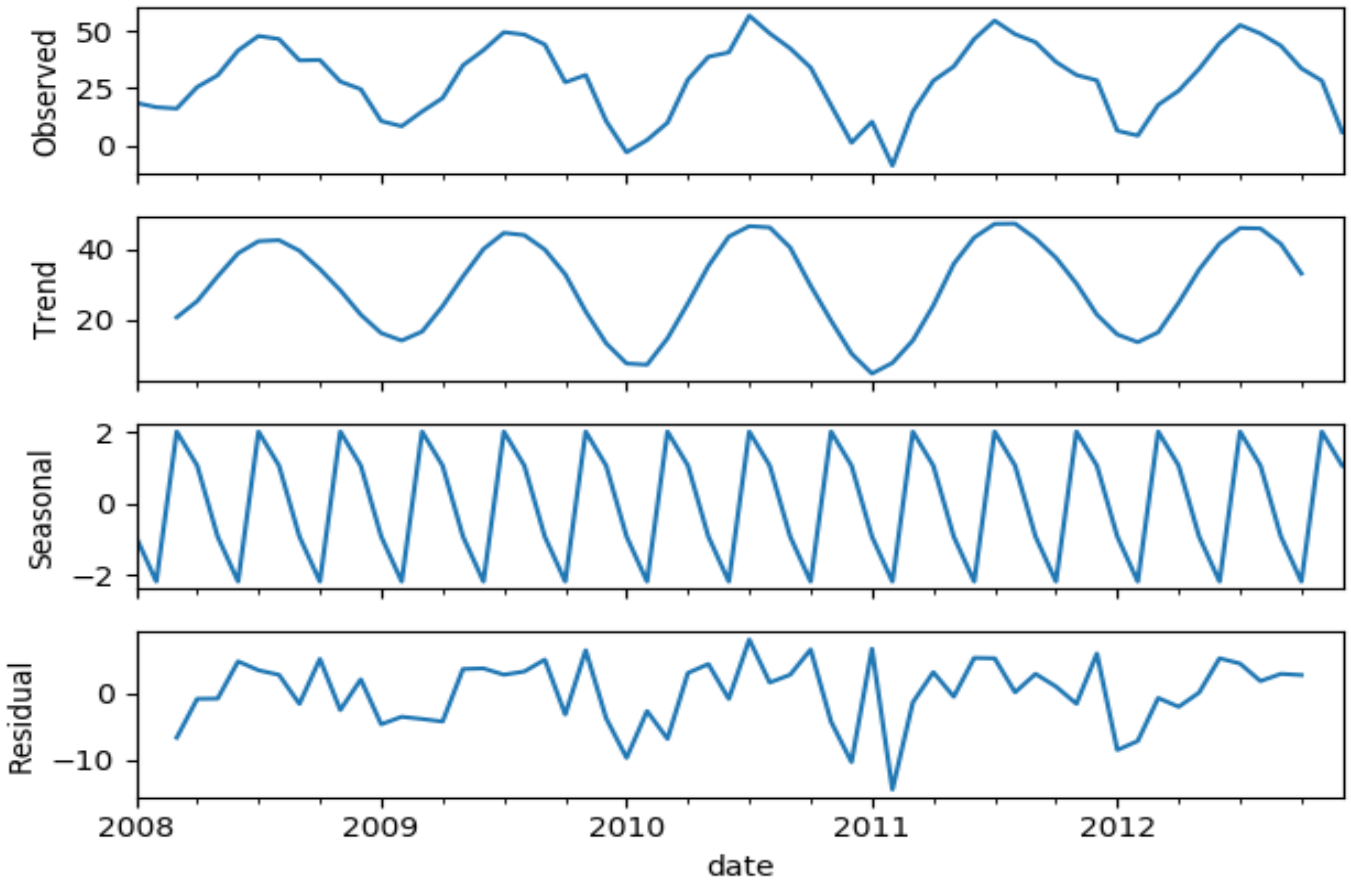

Figure – Séries temporelles

Question 1 : Exploitation spatiale - Visualisation

- Visualiser pour une année spécifique

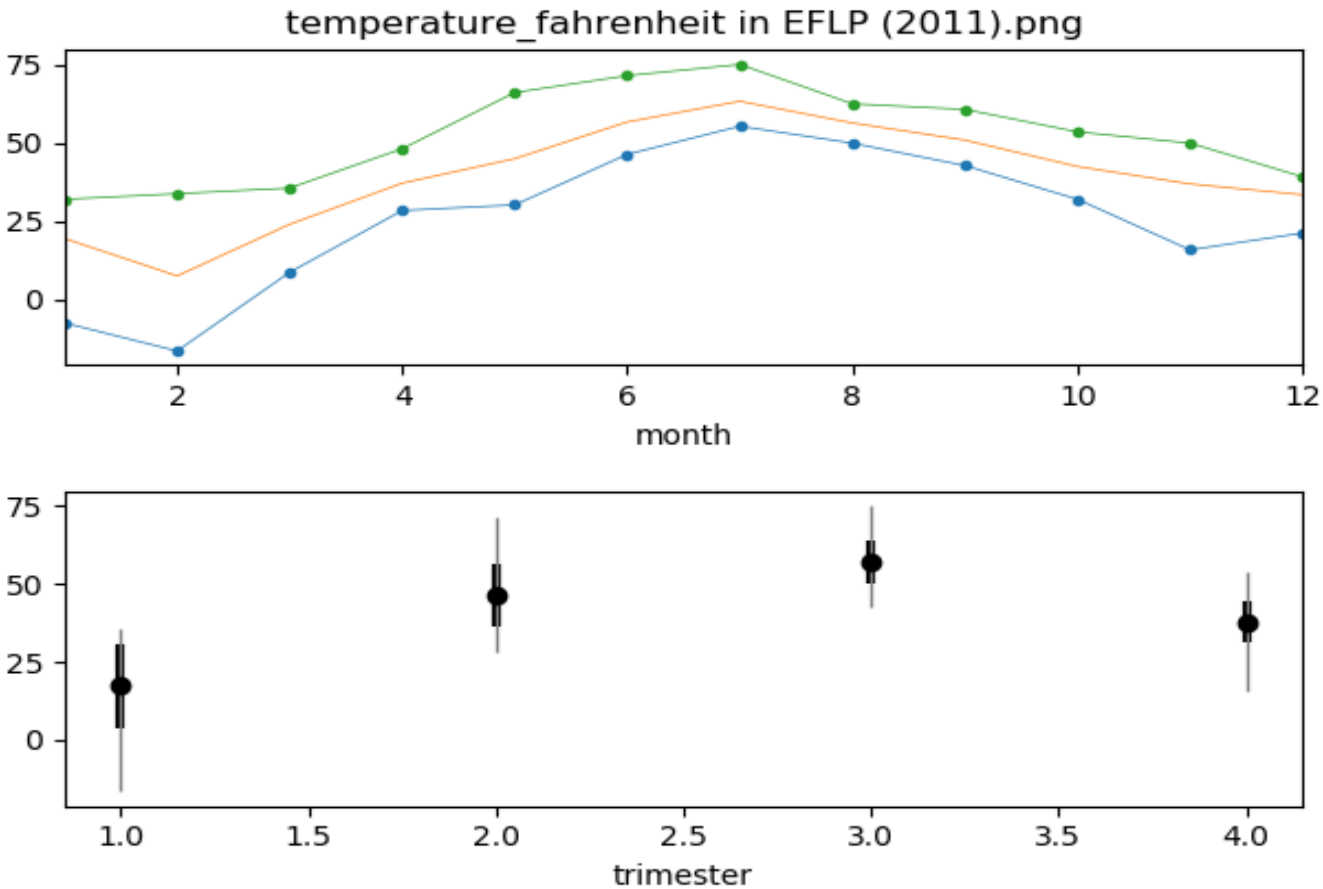

Figure – Visualisation d'une année spécifique

Question 1 : Les requêtes par longitude et latitude

- Calculer la distance avec les stations
- Trouver la station le plus proche
- Requête sur cette station
- Amélioration :
  - Ajouter d'autres types de graphiques
  - Ajouter plusieurs indicateurs en même temps

Questions 2 et 3 : Exploitation temporelle - Conception

- Visualisation : `Matplotlib Basemap Toolkit`
- Stockage :
  - `((year, month), day, hour, minute)`
  - un instant donné $\longrightarrow$ une partition. Pire des cas : parcourir toute la partition
  - une période donnée $\longrightarrow$ m partitions, m, nombre de mois
  - si `day` dans partitionnement, problème pour période donnée

Question 2 : Exploitation temporelle - Instant donné

```
$ map_by_indicator_and_time indicator year-month-day hour
```

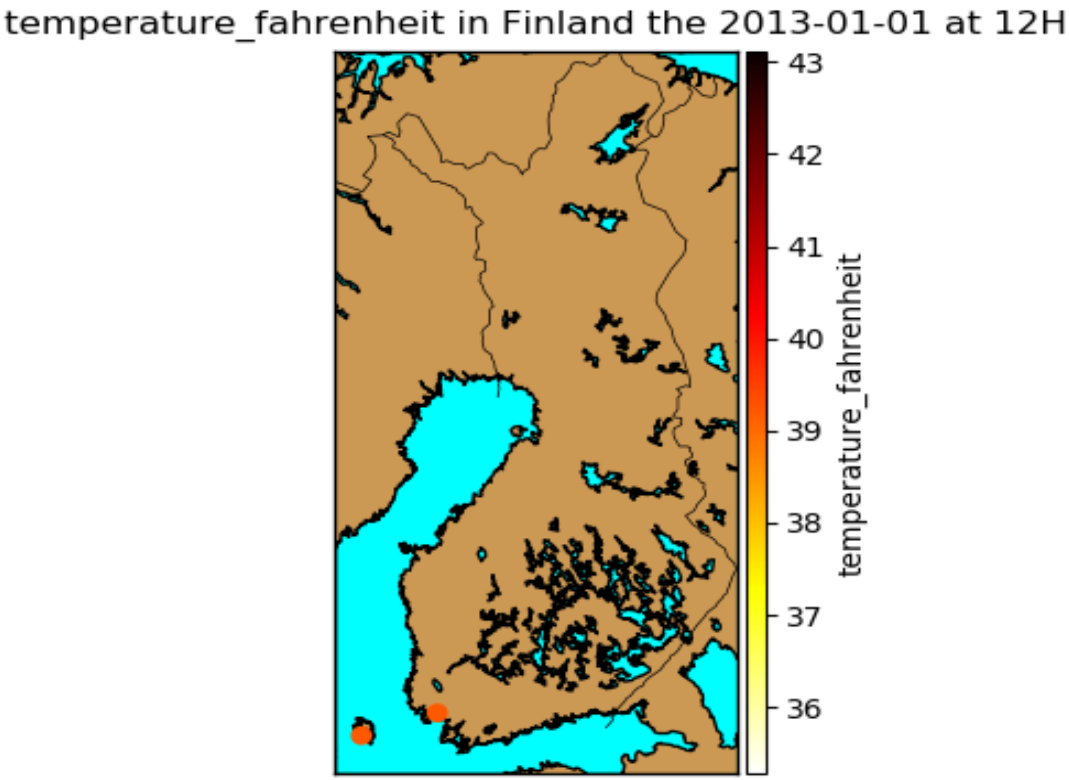

Figure – Températures le 1er janvier 2013 à midi

Question 3 : Exploitation temporelle - Période donnée

- Période = intervalle de dates $\longrightarrow$ problème
  - Ex : 2005-05-01 au 2006-06-01
  - contrainte CQL : $05 \geq \mathit{month} \leq 06$
  - Résultat : on rate 2005 après le mois de juin
- Solution possible : fixer année (pour intervalle de mois) ou année-mois (pour intervalle de jours)
- Décision : intervalle d'années

Question 3 : Exploitation temporelle - Période donnée

- Technologies : `sklearn.cluster.KMeans`
- *Clustering* sur trois variables quantitatives
- Nombre de *clusters* optimal : méthode du coude avec *package* `kneed`

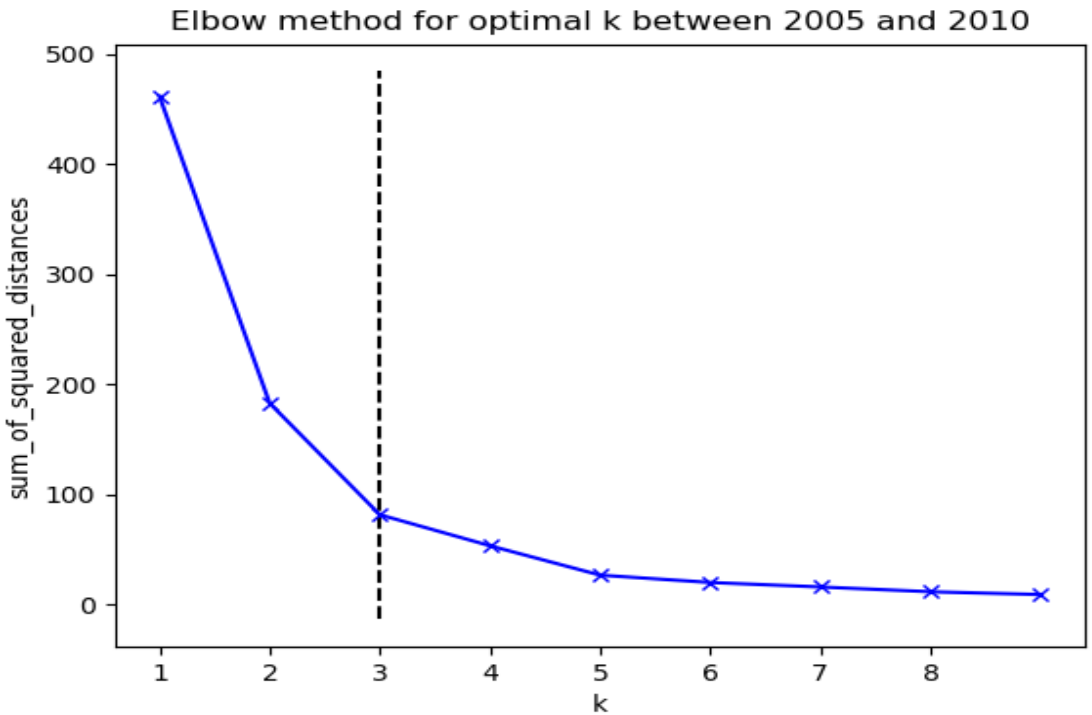

Figure – Inertie intra-classe en fonction de k entre 2005 et 2010

Question 3 : Exploitation temporelle - Période donnée

```
$ cluster_by_period start_year end_year
```

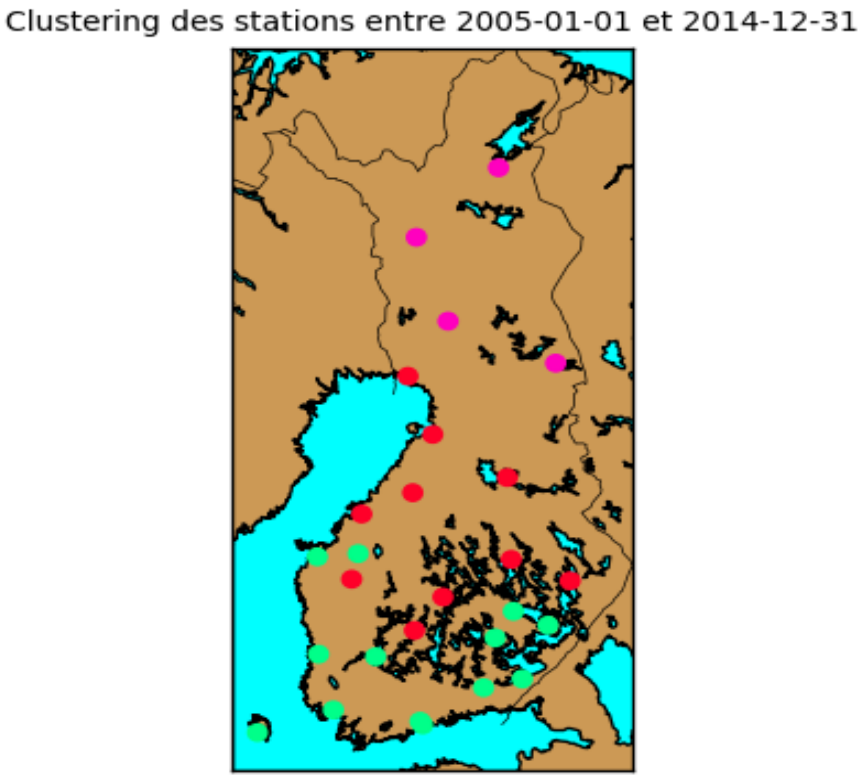

Figure – Clustering des stations sur l'ensemble de la période disponible

Conclusion

- Plus de graphes pour les stats et avec plus d'indicateurs
- Rectifier le stockage : processus itératif
- Améliorer les cartes
- Plus de variables (qualitatives) pour le *KMeans*

Merci pour votre attention !