RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

MÉTHODES DE QUALIFICATION DE DONNÉES ROUTIÈRES

JOURNÉES MOBILITÉS DU RST

Guillaume COSTESEQUE – Cerema Ouest

24/09/2021

PLAN DE LA PRÉSENTATION

1. **Présentation du stage de Baldé MAHMOUD**
   - Quelques définitions
2. **Données aberrantes**
3. **Données manquantes**
   - Synthèse bibliographique
   - Résultats du stage
4. **Synthèse**
   - Conclusion
   - Pistes

1. PRESENTATION DU STAGE

Principaux éléments

ETUDE CLASSIQUE EN INGÉNIERIE DU TRAFIC

1. Commande du **gestionnaire**

2. Réalisation du **diagnostic** de trafic
   - Comprendre la situation actuelle
   - Recueil, traitement et analyse de données

3. « Cœur » de l'étude :
   - **Evaluation a priori** d'amélioration future : comparaison entre différents scénarios d'aménagement de l'infrastructure ou de régulation du trafic
   - **Evaluation a posteriori** : comparaison de la situation avant et après-projet traitement et analyse de données

4. **Communiquer** les résultats au client :
   - Supports visuels

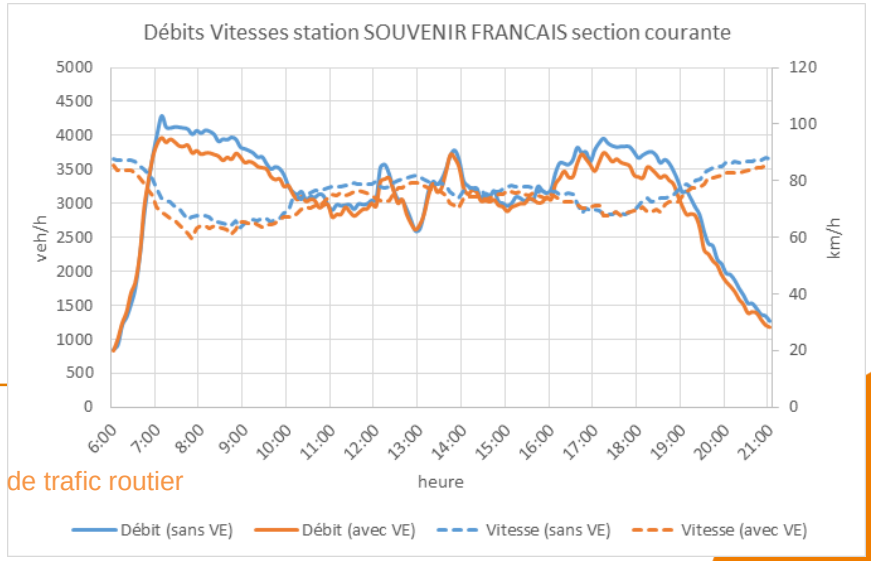
apports attendus du stage

PRÉLIMINAIRES


```
code;horodatage;QT;VT;TT
MWL44.b_2;2019-10-14T00:00:00;37.000000;83.000000;1.000000
MWL44.b_2;2019-10-14T00:06:00;39.000000;84.000000;1.000000
MWL44.b_2;2019-10-14T00:12:00;35.000000;84.000000;0.000000
MWL44.b_2;2019-10-14T00:18:00;30.000000;90.000000;0.000000
MWL44.b_2;2019-10-14T00:24:00;24.000000;89.000000;0.000000
MWL44.b_2;2019-10-14T00:30:00;32.000000;88.000000;0.000000
MWL44.b_2;2019-10-14T00:36:00;26.000000;82.000000;0.000000
MWL44.b_2;2019-10-14T00:42:00;20.000000;89.000000;0.000000
```

débit | vitesse | taux d'occupation

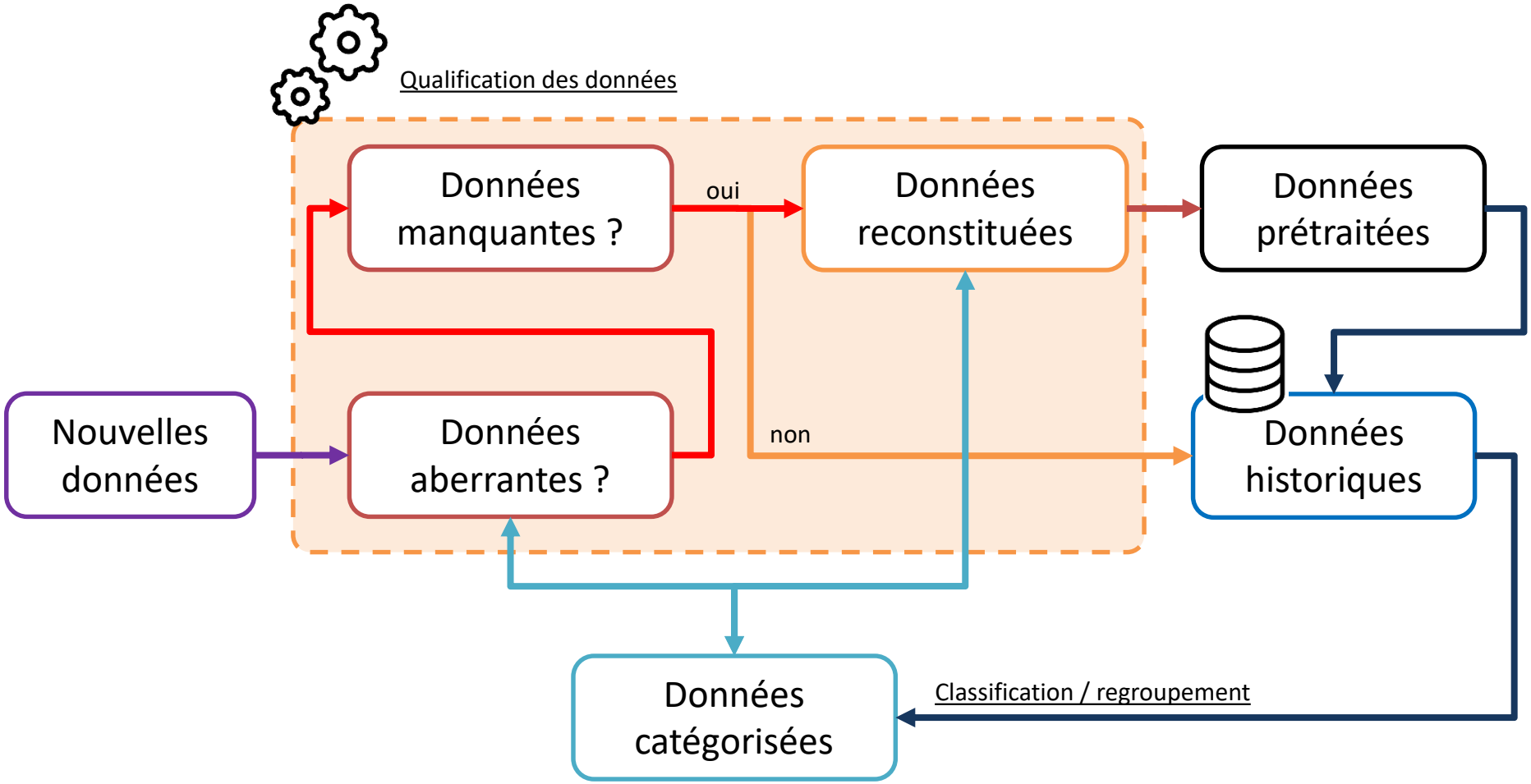
- **Données considérées**
  - 23 points de mesure (SIREDO) sur le périphérique de Nantes (DIR Ouest)
  - Débit, taux d'occupation, vitesse et horodatage → séries temporelles
  - Pas de temps = 6 minutes
  - Profondeur = année 2019 entière



OBJECTIFS DU STAGE

- **Structurer** les données (numériques)
- Générer des **profils-types**
  - apprentissage non-supervisé : *clustering* (exemple : classification ascendante hiérarchique)
  - apprentissage profond
- **Détecter** automatiquement d'éventuelles anomalies
  - les supprimer, les corriger ou les conserver pour un usage ultérieur
- **Compléter** automatiquement les éventuelles **données manquantes**
  - corrélation spatiale et/ou temporelle…
  - réseau de neurones, apprentissage profond (*deep learning*)
- Produire un **tableau de bord** (*dashboard*)
- **Approfondissements** selon le temps disponible
  - données textuelles ?

WORKFLOW CIBLE



PRÉLIMINAIRES

- **Origine des données manquantes / aberrantes :**
  - Capteur en défaut (panne, travaux, vandalisme…)
  - Réseau de transmission en défaut
  - Système informatique en défaut

- **Classification :**
  - **MCAR** : Missing Completely At Random
  - **MAR** : Missing At Random
  - **MNAR** : Missing Not At Random

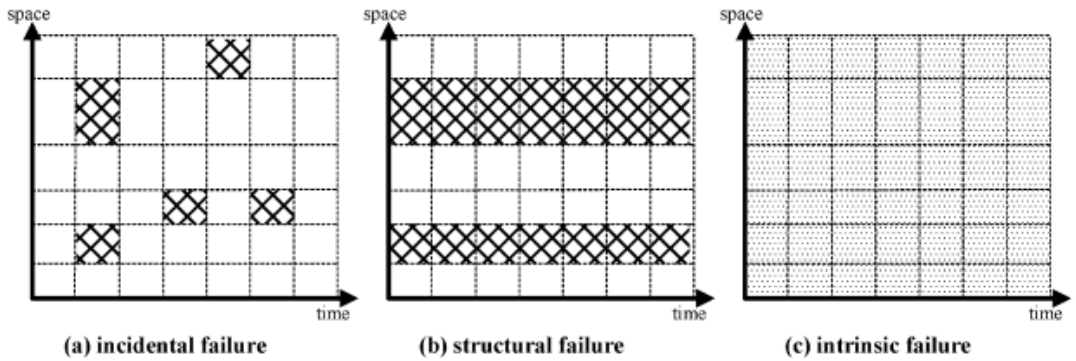


Fig. 4. Classification of possible input failure (i.e. missing or unreliable data from traffic detectors). In practice a mixture of all types of failure will occur.

SPÉCIFICITÉ DU TRAFIC ROUTIER

- **Tenir compte de la redondance de l'information :**
  - **Approche temporelle :** cyclicité du trafic
    - Jour ouvré, week-end, jour férié, veille de jour férié…
    - Vacances scolaires / Eté…
    - Heure de pointe du matin, heure de pointe du soir…
  - **Approche spatiale :** écoulement physique du trafic

2. DONNEES ABERRANTES

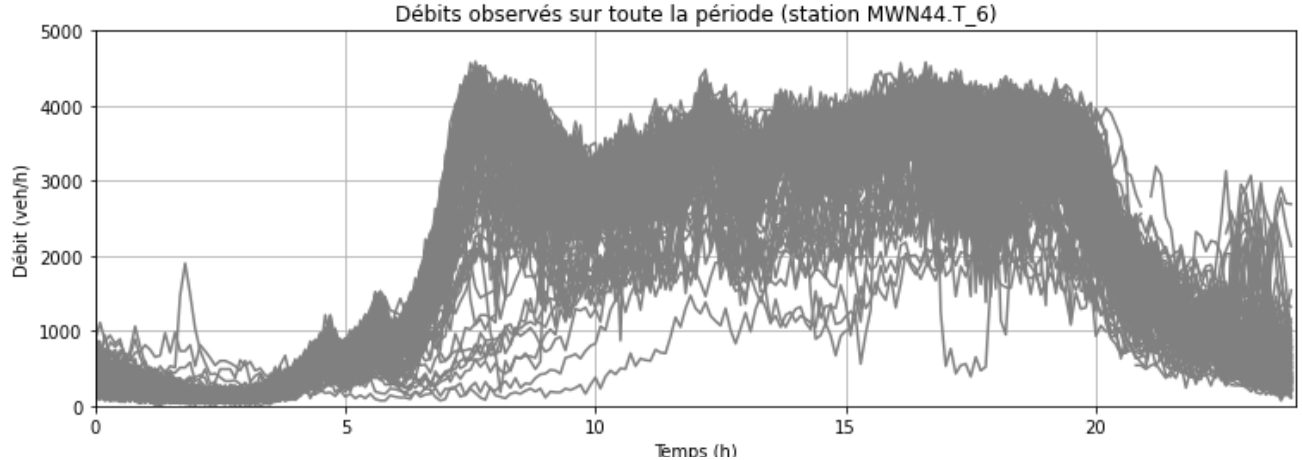
Méthodologie

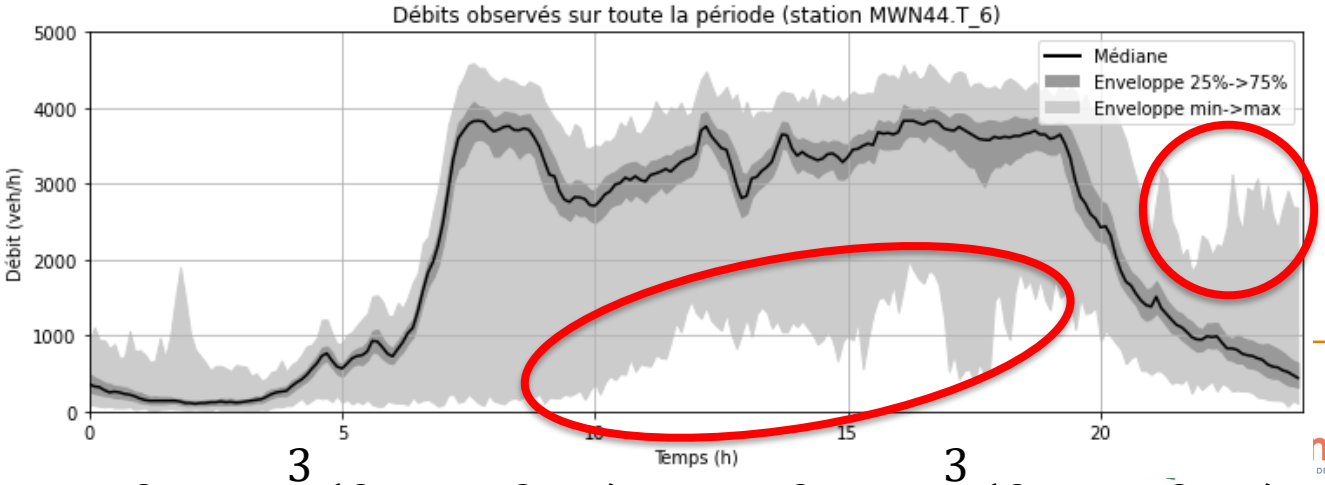
DONNÉES ABERRANTES / ANOMALIES

- **Choix pour le stage : approche naïve**
  - **Seuils statiques** sur les valeurs de débits, vitesses et taux d'occupation
    - Vitesses comprises entre 0 et 150 km/h
    - Débits compris entre 0 et 3000 véh/h et par voie
    - Taux d'occupation compris entre 0 et 100%
  - **Combinaisons impossibles**
    - Si variable 1 = 0, alors variable 2 = variable 3 = 0
  - Valeurs constantes sur un nombre successif de pas de temps

DONNÉES ABERRANTES / ANOMALIES

- **Analyse statistique (~ Box plot)**
  - Code par Aurélien CLAIRAIS

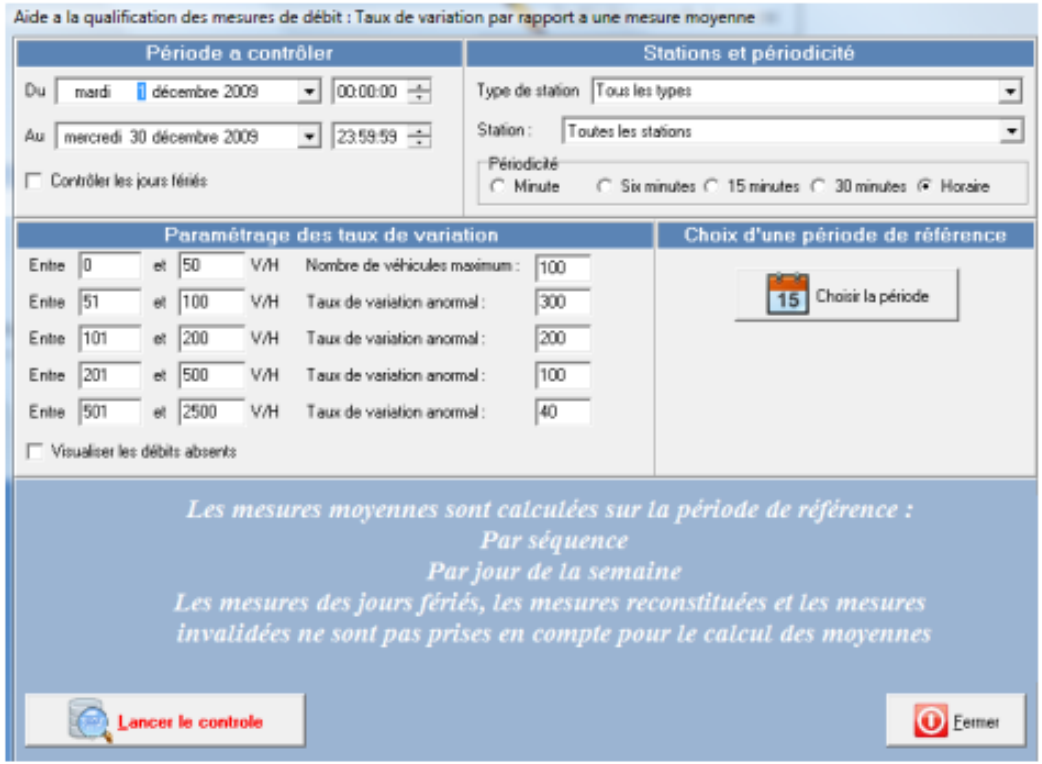




$$Q_{25\%} - \frac{3}{2}(Q_{75\%} - Q_{25\%}) \leq x \leq Q_{75\%} + \frac{3}{2}(Q_{75\%} - Q_{25\%})$$

MÉTHODE UTILISÉE PAR LES PROFESSIONNELS

- **PCADC3 (LABOCOM)**
  - **Seuils statiques** sur les taux de variation



3. DONNÉES MANQUANTES

Synthèse bibliographique

MÉTHODES USUELLES

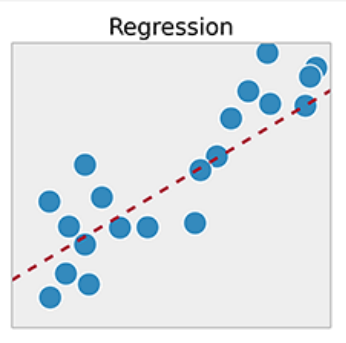
- **Approches « univariées »**
  - Ne tenir compte **que de la variable** pour laquelle on cherche à imputer les valeurs manquantes

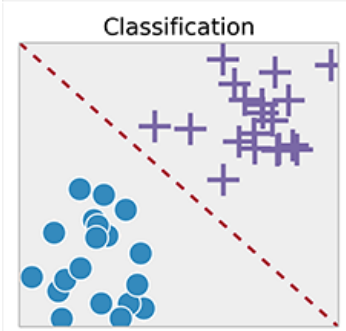
    Exemple : pour imputer des valeurs de débit manquantes, je ne considère que la série temporelle des débits (pas des vitesses et/ou des taux d'occupation)

- **Approches « multivariées »**
  - Tenir compte de **l'ensemble des variables** (non manquantes) pour imputer les valeurs manquantes d'une variable

    Exemple : pour imputer des valeurs de débit manquantes, je considère les séries temporelles des débits et des vitesses et/ou des taux d'occupation

MÉTHODES USUELLES (UNIVARIÉES)





- **Approches naïves**
  - Suppression de la série incomplète
  - Remplacement par une valeur constante : moyenne, médiane…
  - Remplacement par la valeur précédente / suivante
- **Approches paramétriques**
  - Régression linéaire simple, multilinéaire
  - Interpolation : linéaire, polynomiale, splines…
  - Inférence Bayésienne : filtre de Kalman
  - Approches autorégressives de moyennes mobiles (ARMA, ARIMA, SARIMA…)
- **Approches non paramétriques**
  - Classification : linéaire par SVM, k plus proches voisins (kNN), classification ascendante hiérarchique (CAH), décomposition en valeurs singulières (SVD)…
  - Clustering : K-means, Mean Shift, DBSCAN…

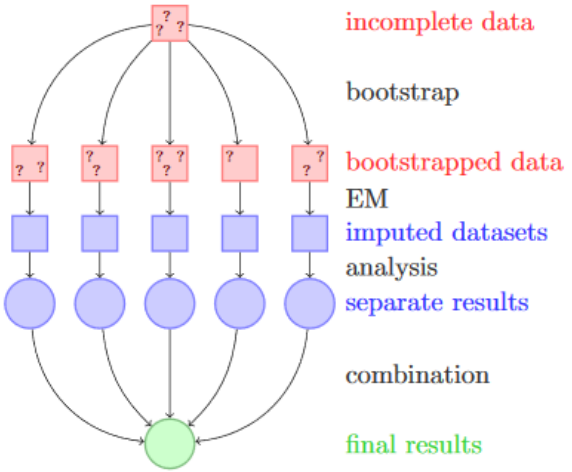
SCHÉMA CLASSIQUE : COMBINAISON DE MÉTHODES

1. Réduction de la dimensionnalité du problème

- Analyse factorielle (Analyse en Composantes Principales, analyse des correspondances…)
- Regroupement :
  - Classification (apprentissage supervisé)
  - Clustering (apprentissage non-supervisé)

2. Imputation

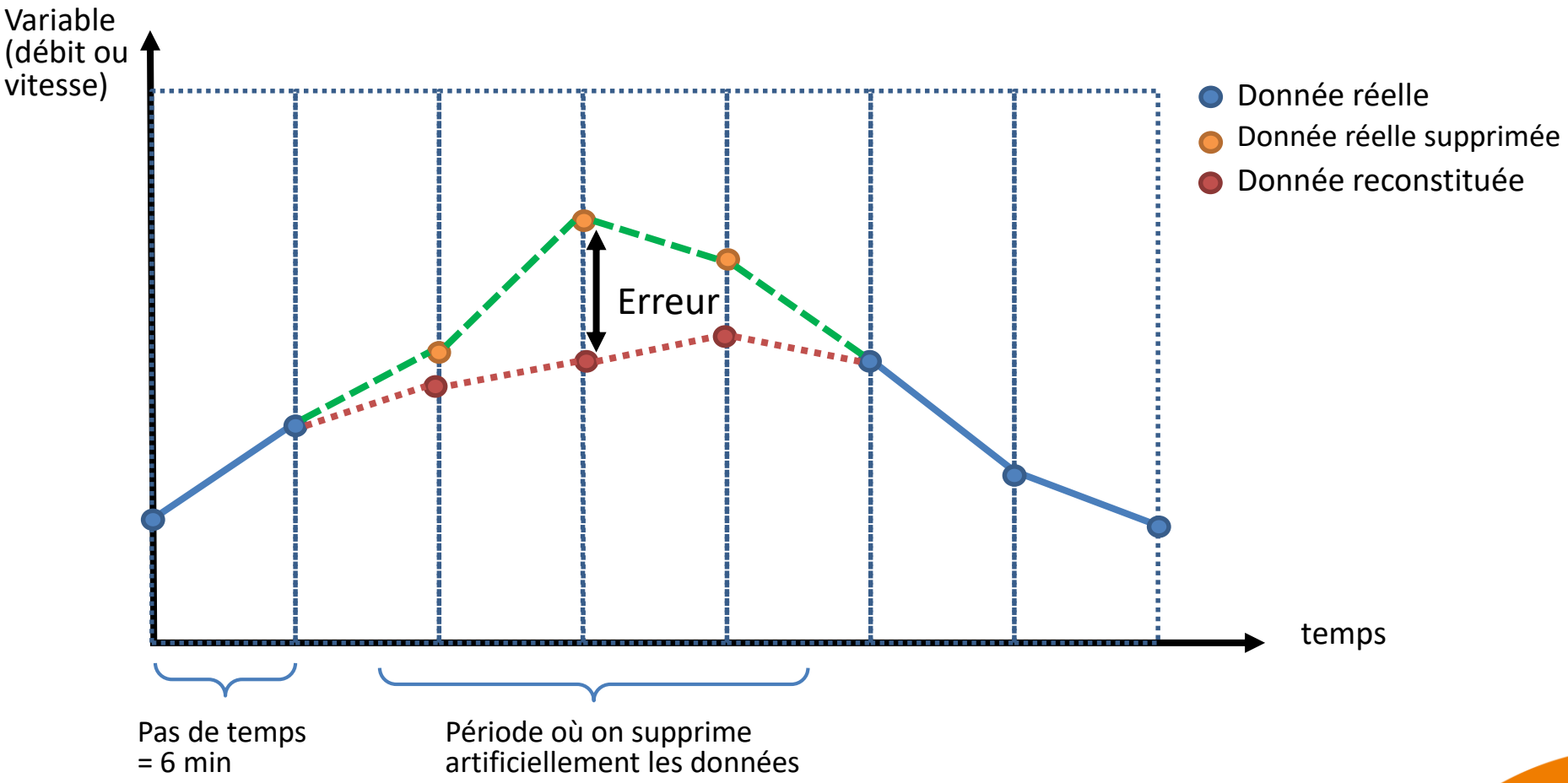
- Régressions, interpolations…
- Imputations multiples : exemple de MICE sous R
  - Régression linéaire quadratique
  - Moyenne prédictive
  - Arbres de régression et classification
  - Forêts aléatoires



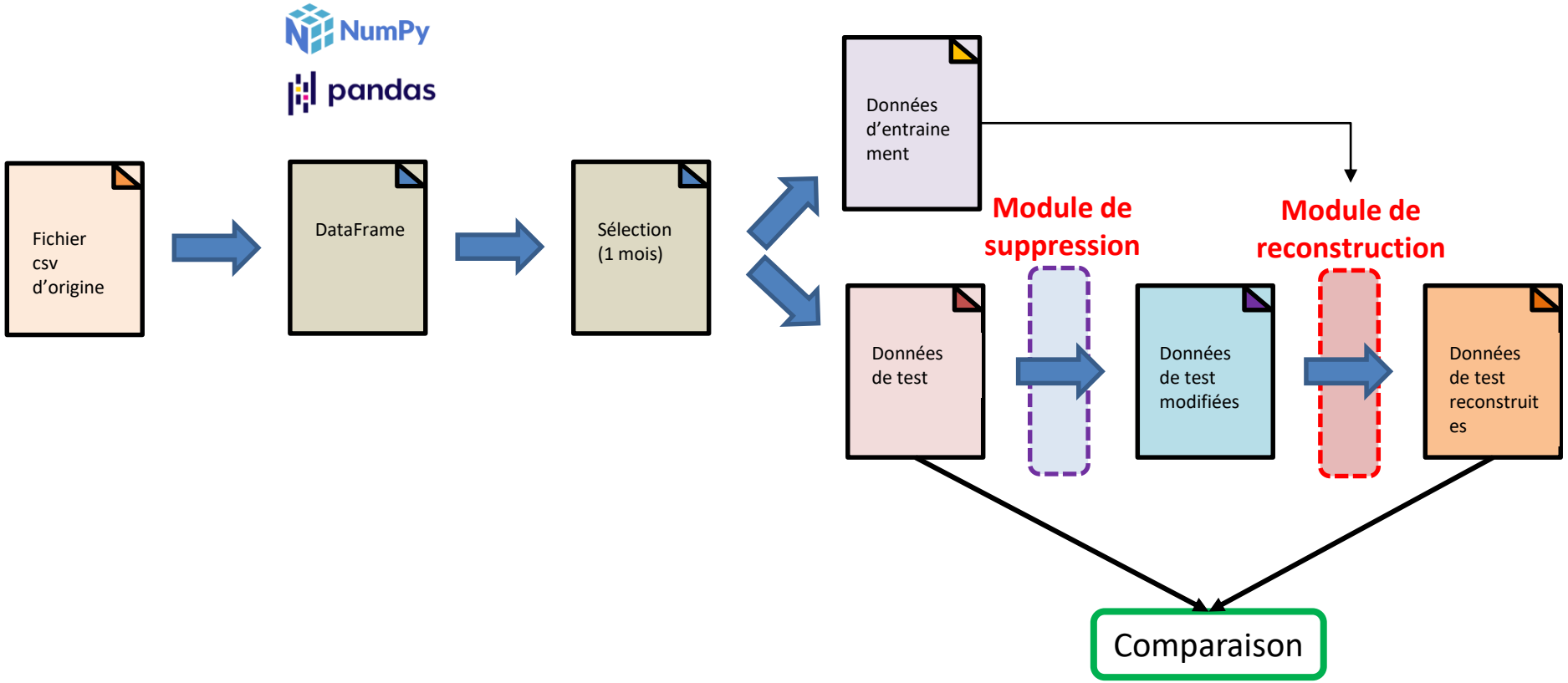
COMPARAISON DE MÉTHODES D'IMPUTATION

- **Méthodologie :**
  - Suppression de données de manière **contrôlée** dans une série temporelle « propre » selon différents critères :
    - Position de la première valeur manquante dans la série temporelle
    - Nombre de valeurs manquantes successives
  - Imputation des valeurs manquantes par **7** différents algorithmes :
    1. Imputation par la moyenne de la série
    2. Interpolation linéaire
    3. Filtre de Kalman
    4. K plus proches voisins (kNN)
    5. Forêts aléatoires (MissForest)
    6. Imputation multiple (R MICE / Python Scikit-learn IterativeImputer)
    7. Analyse en Composantes Principales (ACP)
  - Calcul de l'erreur commise (**RMSE**) et conclusion

COMPARAISON DES MÉTHODES D'IMPUTATION



SUPPRESSION DES DONNÉES



COMPARAISON DES MÉTHODES D'IMPUTATION

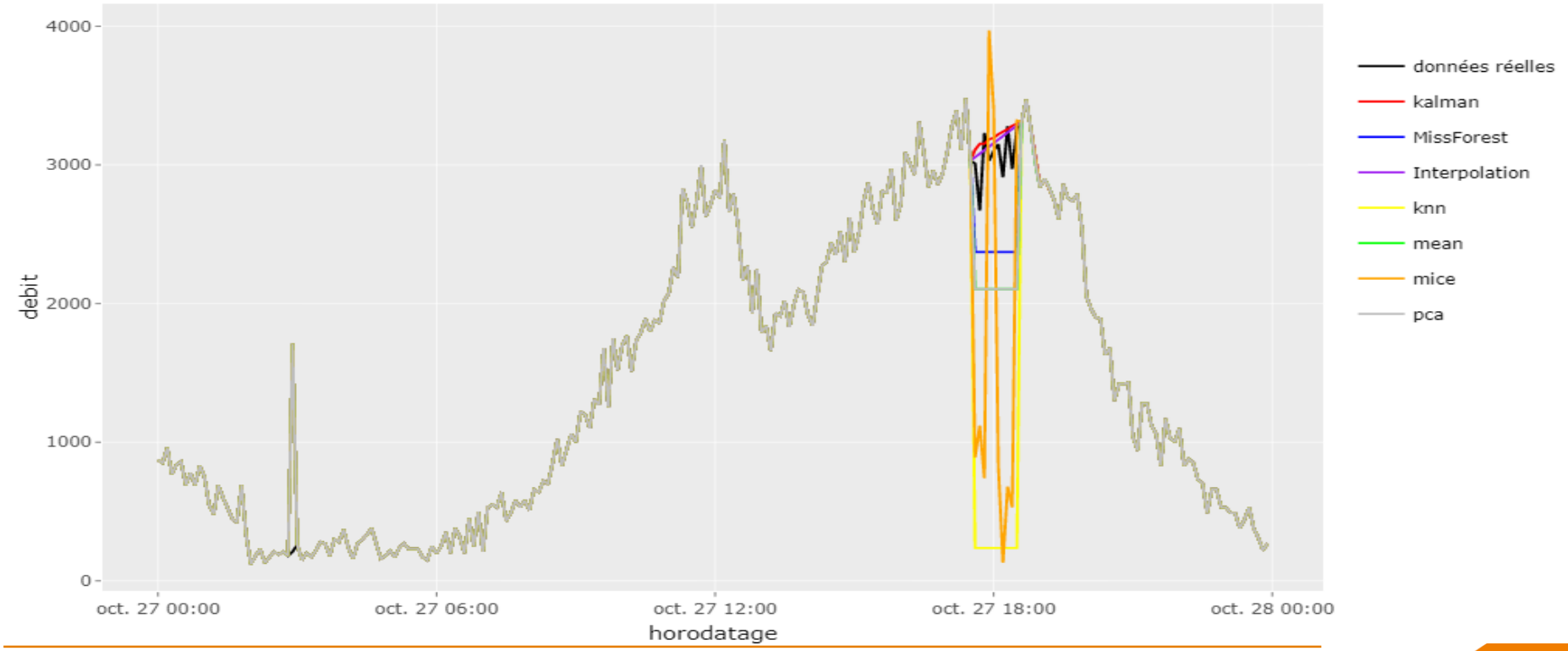
- **Résultats (pour la variable « débit ») :**
  - Fonction métrique RMSE
  - Meilleurs résultats pour interpolation linéaire et filtre de Kalman
  - Similaire pour les variables « vitesse » et « taux d'occupation »

Nbre_val_con	Horaires	knn	moy	missF	Kalman	interpo	mice1	pca1
1	Nuit	92.47643	95.14350	96.99616	91.77866	91.77815	92.29215	95.14350
1	Journée	111.60718	95.69412	94.06104	91.77769	91.78929	107.54051	95.69412
2	Nuit	94.05430	96.92197	99.87503	91.87094	91.84968	93.61490	96.92197
2	Journée	130.65586	100.74124	97.49066	91.87434	91.96652	101.64853	100.74124
5	Nuit	95.88153	106.66728	114.44817	92.23499	92.32493	103.00302	106.66728
5	Journée	166.98048	108.73710	102.41231	92.17867	92.15430	164.38453	108.73710
10	Nuit	187.71581	112.39822	107.22299	95.92731	96.00240	147.52709	112.39822
10	Journée	196.85015	109.57221	101.68017	92.78958	92.55067	153.41529	109.57221

- Interrogation sur le paramétrage des méthodes KNN et MICE

COMPARAISON DES MÉTHODES D'IMPUTATION

- **Résultats (pour la variable « débit ») :**
  - Exemple graphique pour 10 valeurs consécutives manquantes



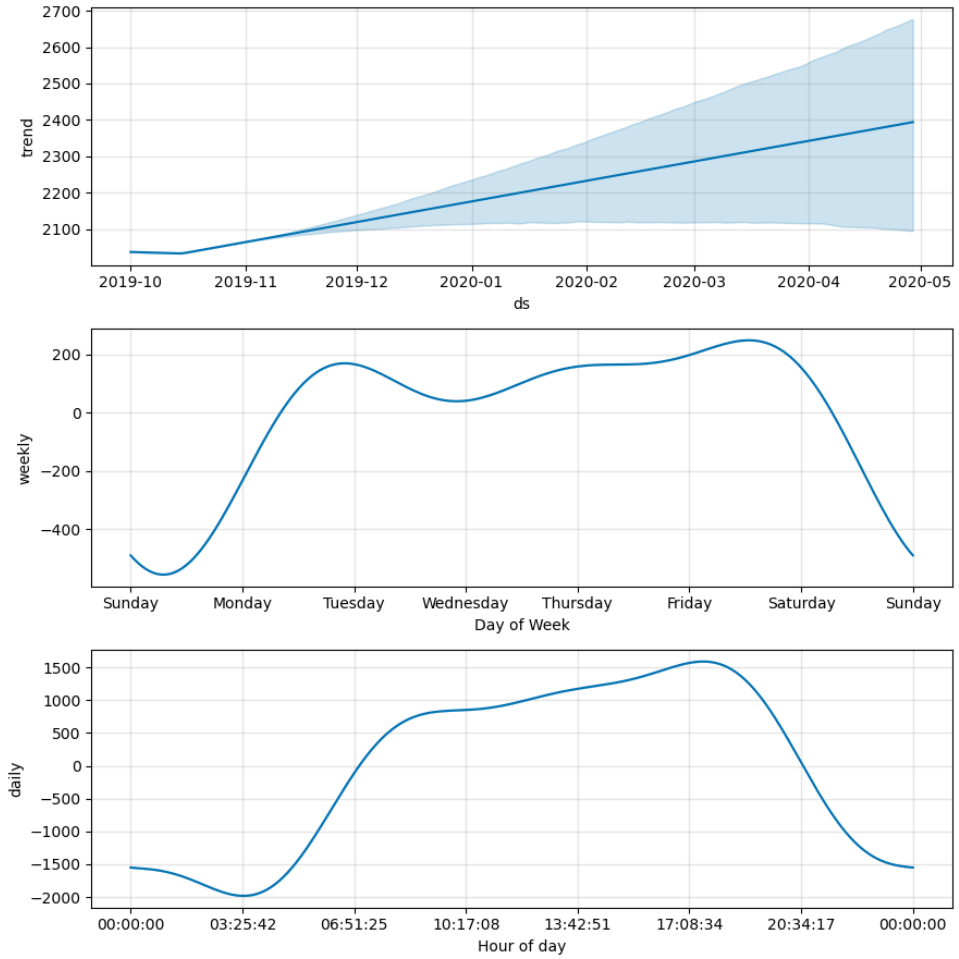
4. SYNTHÈSE

Conclusion et pistes de travail envisagées

OBJECTIFS DU STAGE

- **Structurer** les données (numériques) 
- Générer des **profils-types** 
  - apprentissage non-supervisé : *clustering* (exemple : classification ascendante hiérarchique)
  - apprentissage profond
- **Détecter** automatiquement d'éventuelles anomalies 
  - les supprimer, les corriger ou les conserver pour un usage ultérieur
- **Compléter** automatiquement les éventuelles **données manquantes** 
  - corrélation spatiale et/ou temporelle…
  - réseau de neurones, apprentissage profond (*deep learning*) 
- Produire un **tableau de bord** (*dashboard*) 
- **Approfondissements** selon le temps disponible 
  - données textuelles ?

LIBRAIRIES SPÉCIFIQUES

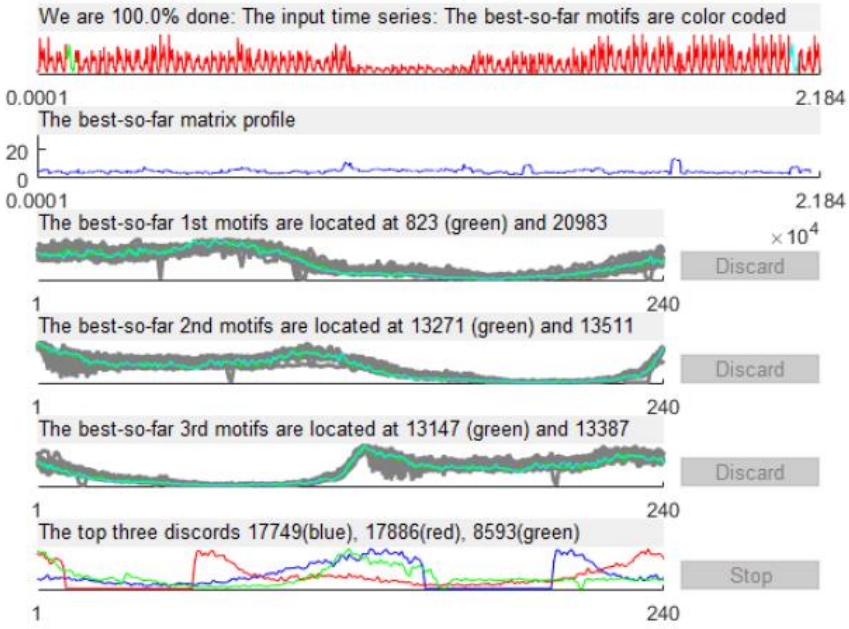


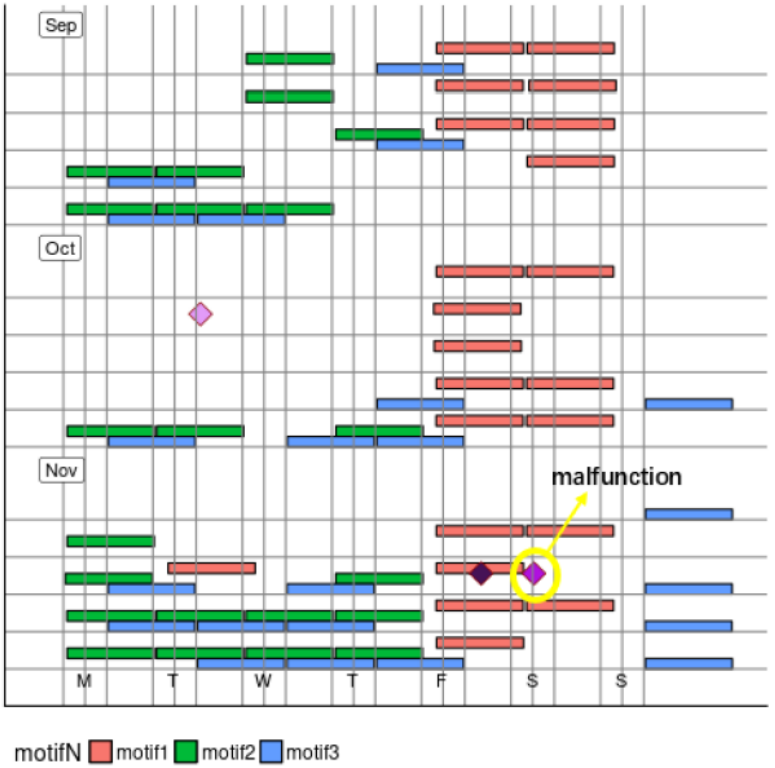
- **Prophet**
  - Développé par Facebook : https://facebook.github.io/prophet/docs/quick_start.html#python-api
  - Saisonnalités (jour / semaine / année)
  - Granularité sous l'heure ?

- **Autres librairies :**
  - statsmodels
  - sktime
  - tslearn

AUTRES MÉTHODES ?

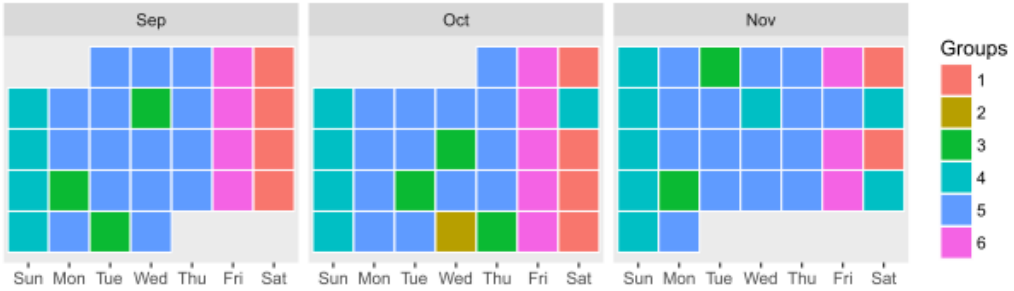
- **Découverte de motifs**
  - Cf. stage master de Y. WU (2018)

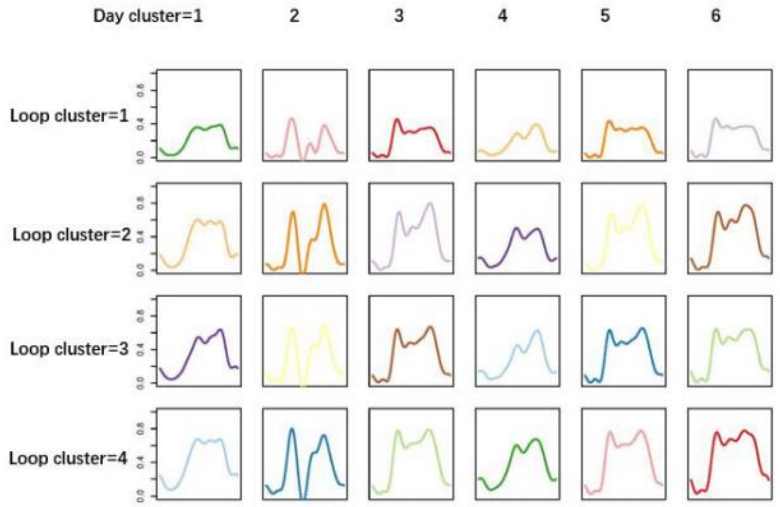




AUTRES MÉTHODES ?

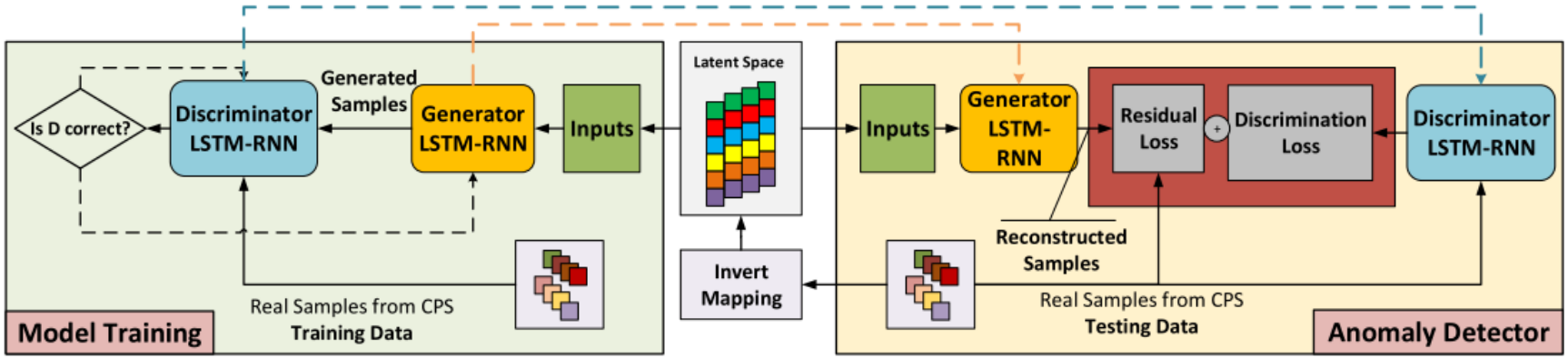
- **Apprentissage automatique :**
  - Co-clustering (cf. master Y. WU, 2018)





  - Approche multivariée / basée sur des tenseurs

  - Réseaux génératifs adverses (GAN)



RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

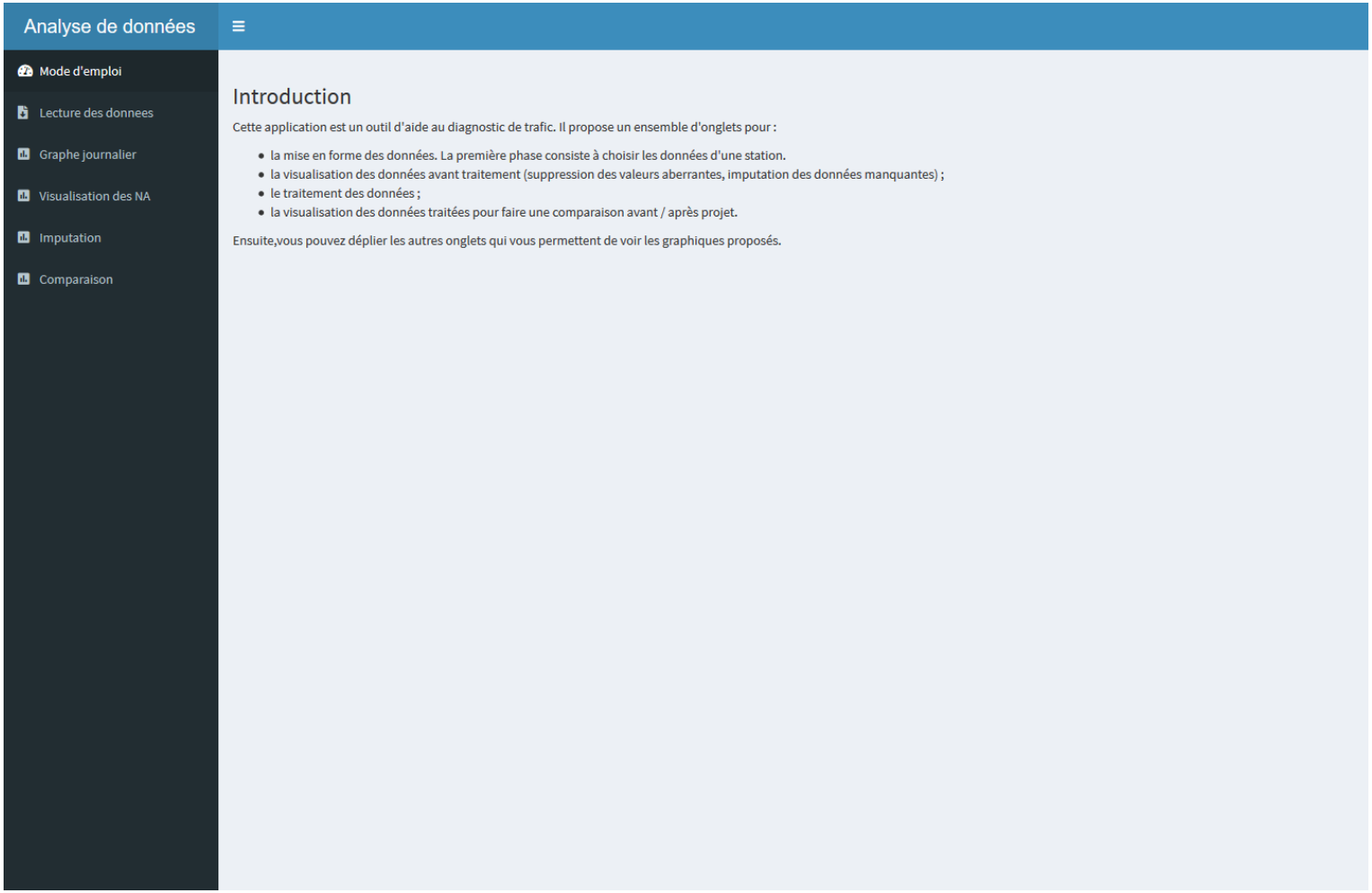
MERCI

Guillaume COSTESEQUE
Chargé d'études optimisation du trafic routier et systèmes de transport intelligents
guillaume.costeseque@cerema.fr

ANNEXES

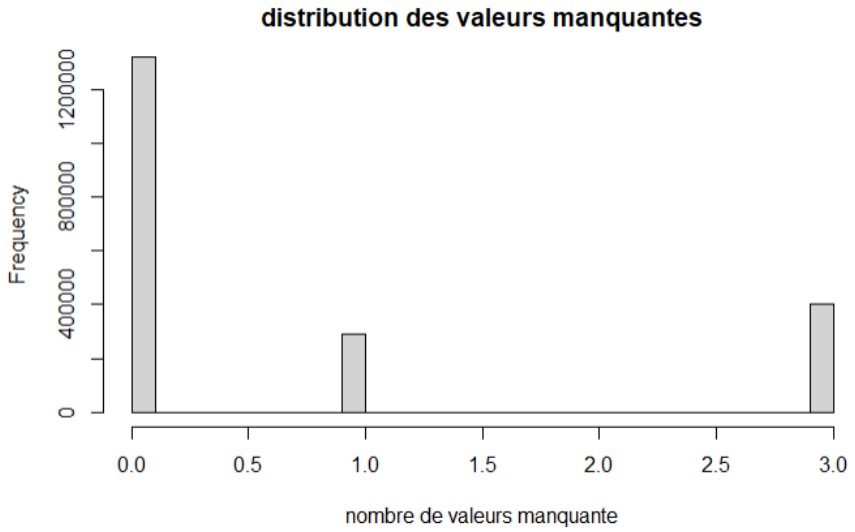
Compléments

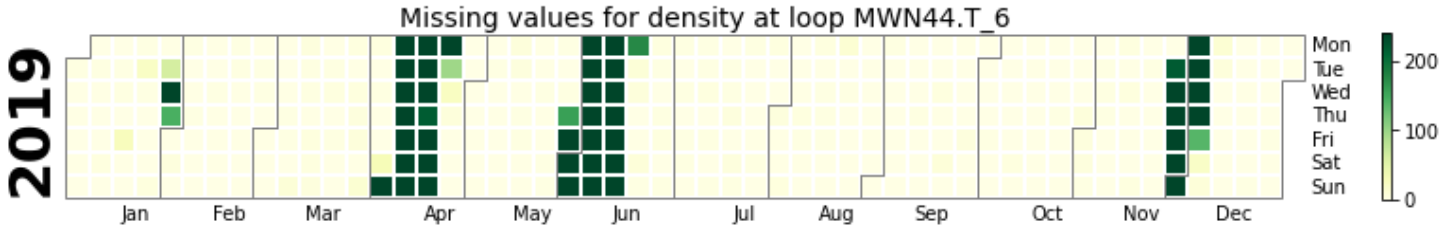
ANNEXE : APPLI R SHINY



QUELQUES DÉFINITIONS

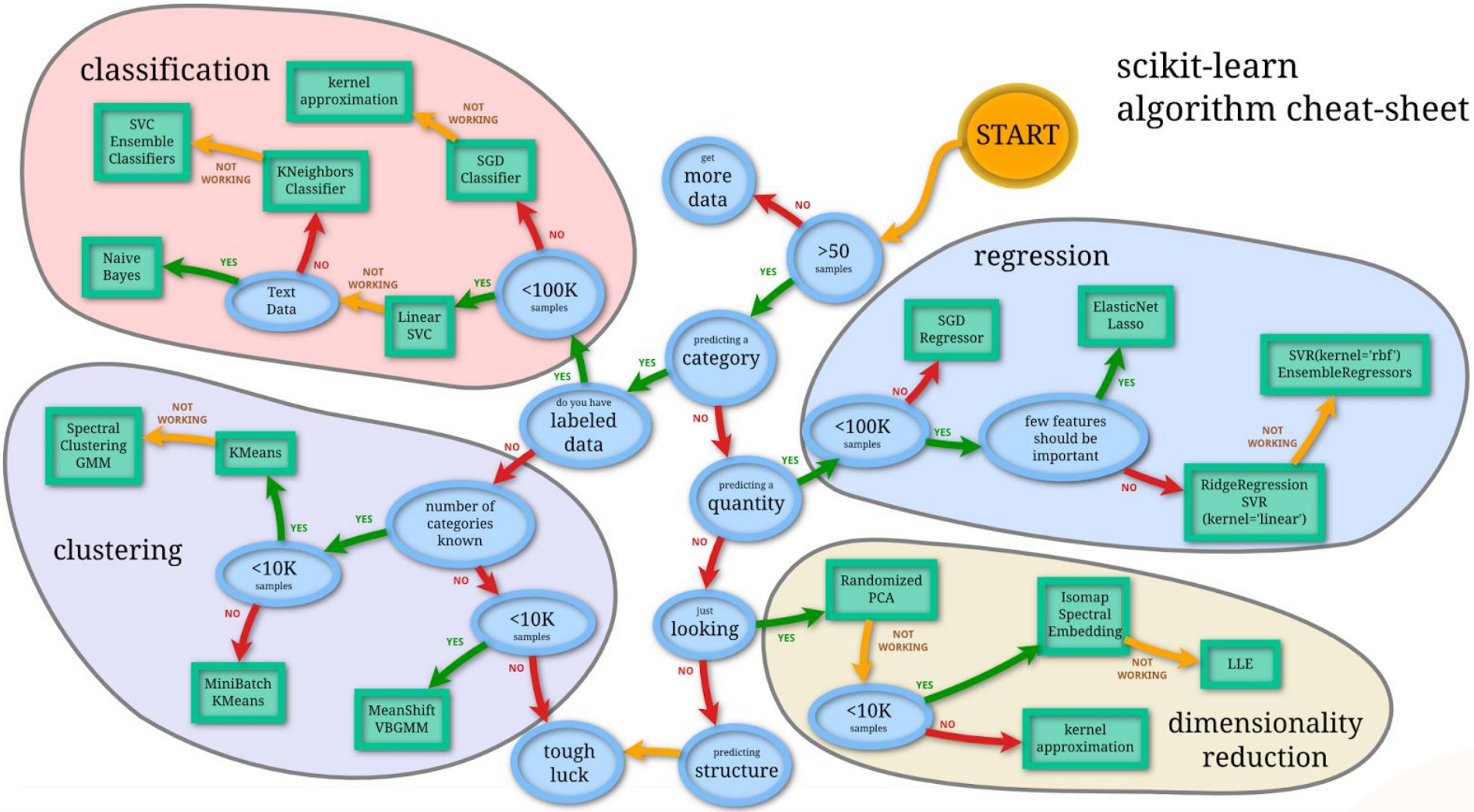
- **Configuration de données manquantes :**
  - Dans 65% des cas, aucune valeur manquante
  - Dans 15% des cas, 1 variable manquante → peut être reconstruite à partir des 2 autres
  - Dans 20% des cas, les 3 variables sont simultanément manquantes



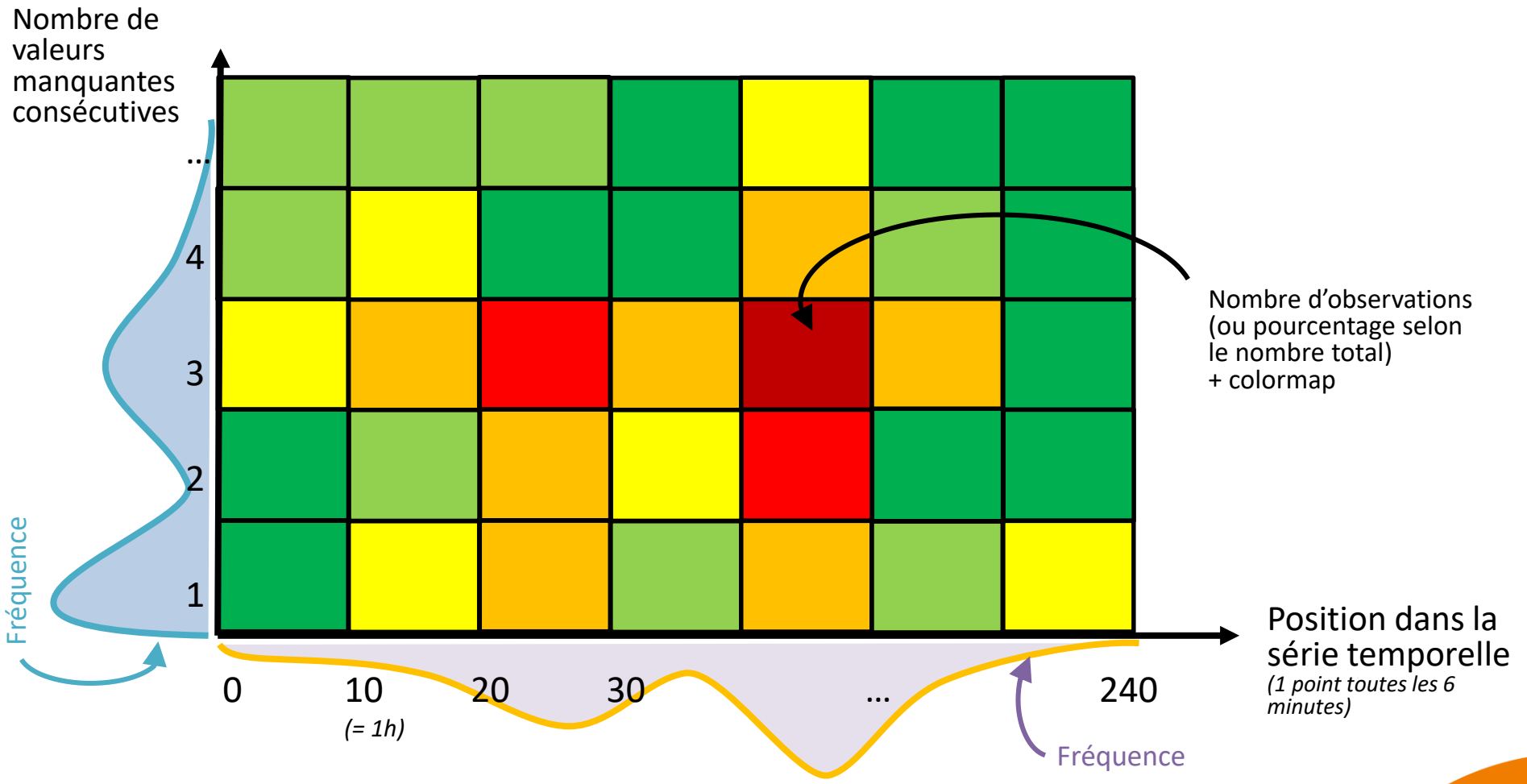


RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

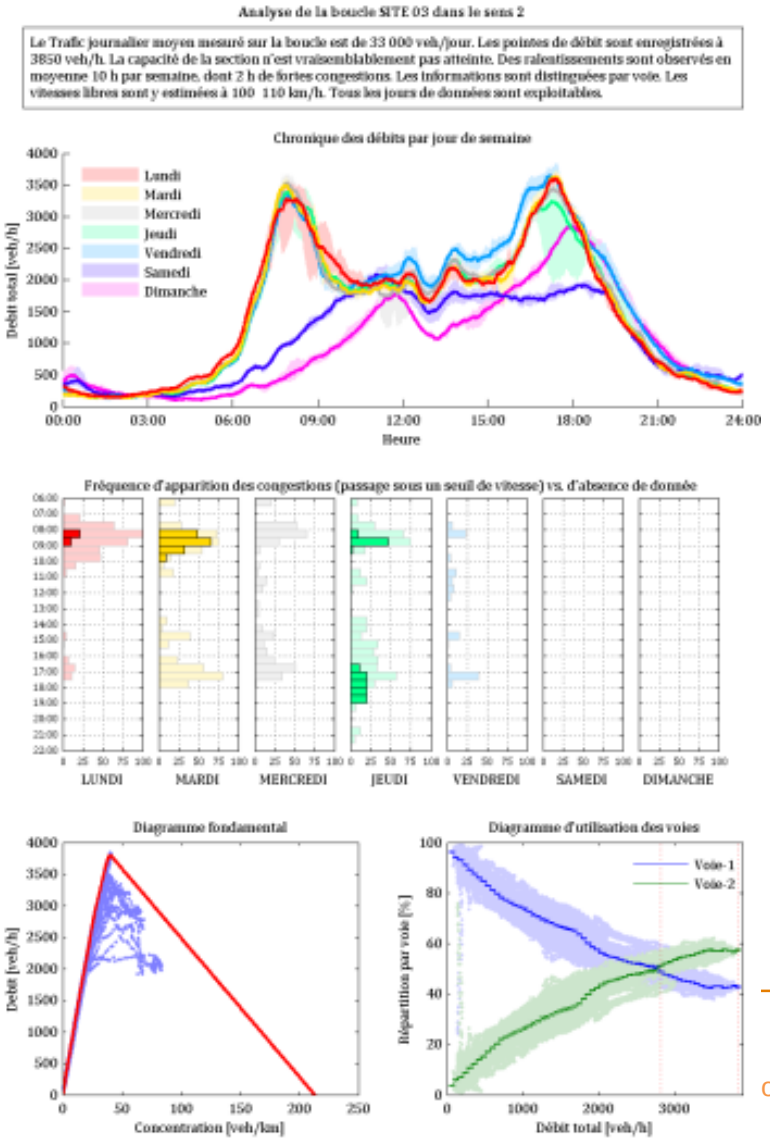
ALGORITHME DE SÉLECTION DES MÉTHODES



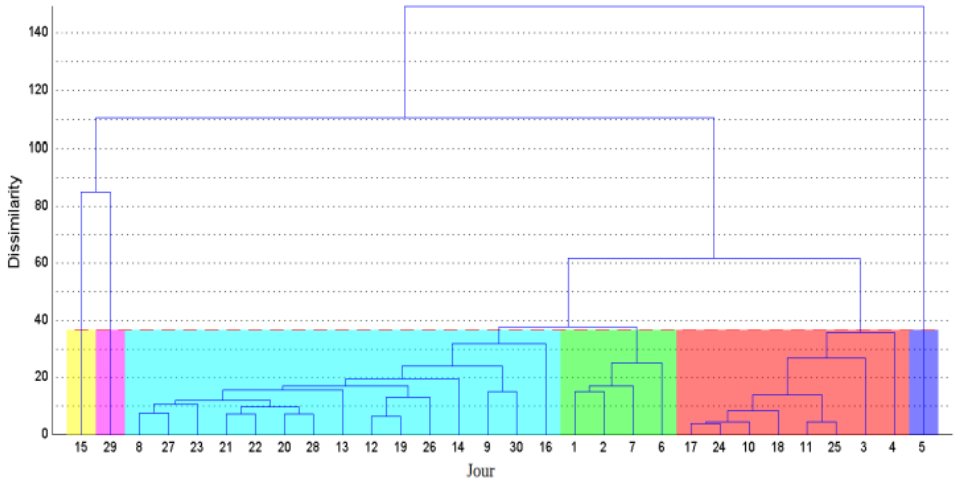
RÉPARTITION STATISTIQUE DES DONNÉES MANQUANTES



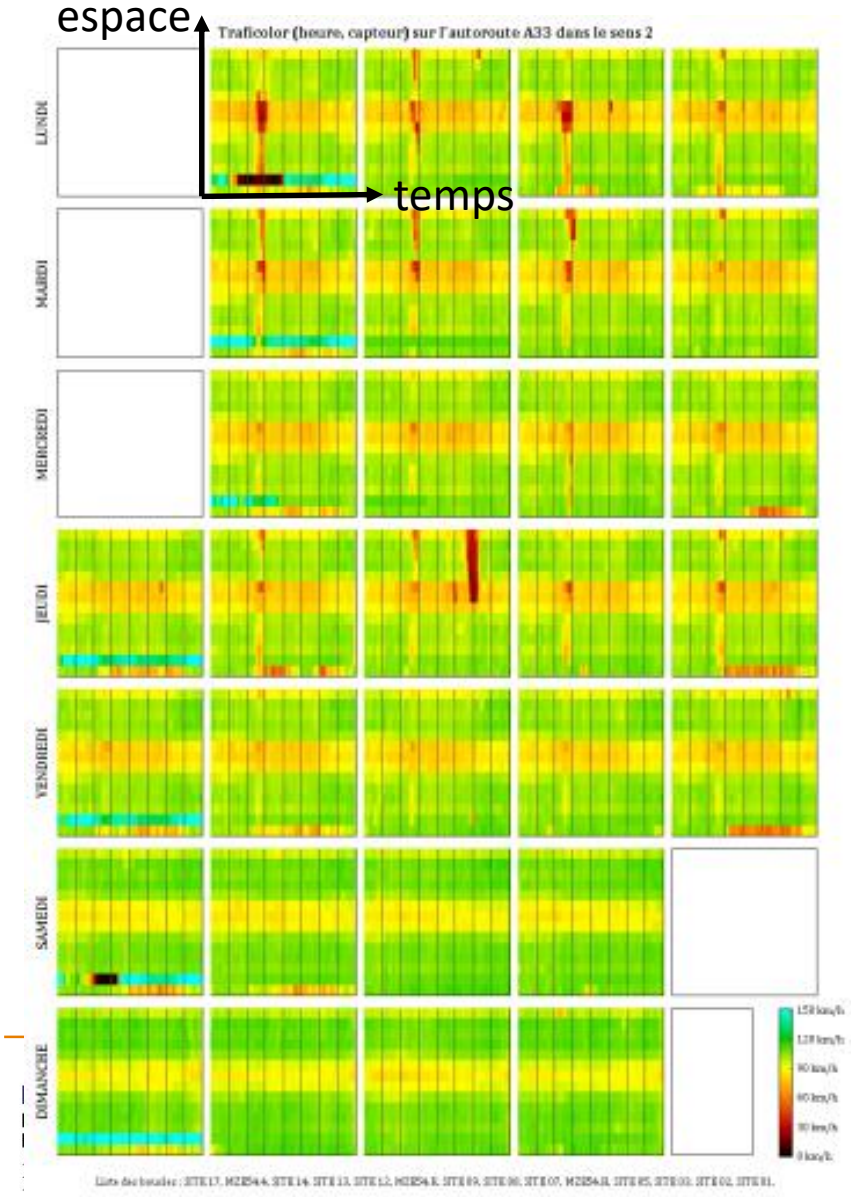
QUELQUES EXEMPLES GRAPHIQUES (Hans et al, 2018)



- Analyse des données **par capteur** :
  - Chronique de débit
  - Chronique de vitesse
  - Vue calendaire (semaine, mois, année…)
  - Diagramme fondamental
  - Débits par voie
  - …



QUELQUES EXEMPLES GRAPHIQUES (Hans et al, 2018)



- Analyse des données **sur plusieurs capteurs** :
  - Diagramme espace-temps
  - Vue calendaire (semaine, mois, année)
  - …