RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

LES RENDEZ-VOUS DU CEREMA Mobilités

neovya Mobility by Technology

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

PROJET « AVATAR »

Rendez-vous mobilités

13 décembre 2022

QUALIFICATION DE DONNÉES ROUTIÈRES

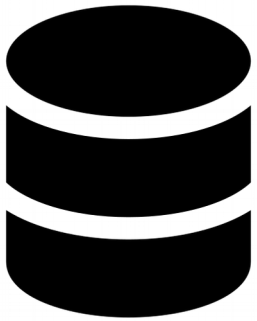
Observations :

- Données **aberrantes / erronées** :
  - Valeurs inconsistantes et/ou répétées
    - Exemple : - vitesse moyenne supérieure à 170 km/h
      - - débit égal à 1500 veh/h pendant 6 heures successives.
  - Combinaisons impossibles
    - Exemple : débit nul mais vitesse non nulle
- Données **anormales** :
  - Evènements très exceptionnels (évènements sociaux…)
- Données **manquantes** :
  - Détecteur en panne
  - Réseau informatique en panne
  - Dérive du capteur…

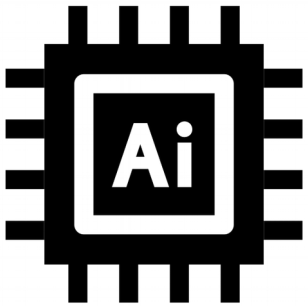
ARCHITECTURE

R

brutes
des capteurs et retourne les
mesures **qualifiées**



L'interface requête la BdD
pour faire les **affichages**



Base de données

Module IA

Tableau de bord

MODULE IA : METHODOLOGIE

1) DÉTECTION DE VALEURS ABERRANTES

En prétraitement (avant complétion) → méthodes naïves

- Suppression des valeurs :
  - Supérieures au double du 95e percentile (par variable et par capteur)
  - Constantes sur une plage de temps :
    - Valeurs non nulles : max. 5 observations successives
    - Valeurs nulles : max. 10 observations successives

2) MÉTHODOLOGIE DE COMPLÉTION

Traitement hors ligne

Etape 1 :

- Recueil des données **historiques**
  - 3 à 5 ans
- Filtrage des données aberrantes
- Formatage des données (dates)

Etape 2 :

- Entrainement du modèle
- Complétion des données manquantes
- Mise en base

Traitement en ligne

Etape 3 :

- Chargement du modèle entrainé
- Recueil et prétraitement des données **temps réel**

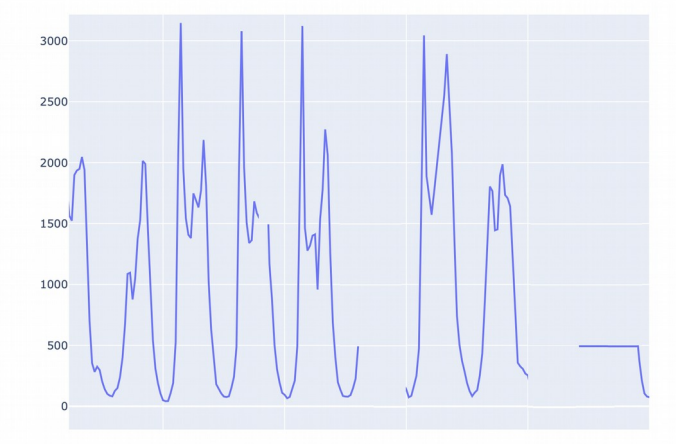
Etape 4 :

- Complétion des données manquantes
- Mise en base

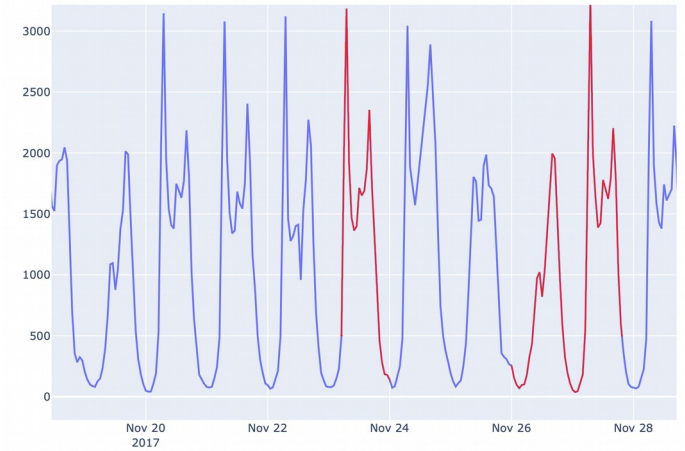
Réentrainement ponctuel du modèle

2) MÉTHODOLOGIE DE COMPLÉTION

Etape 1



Etape 2

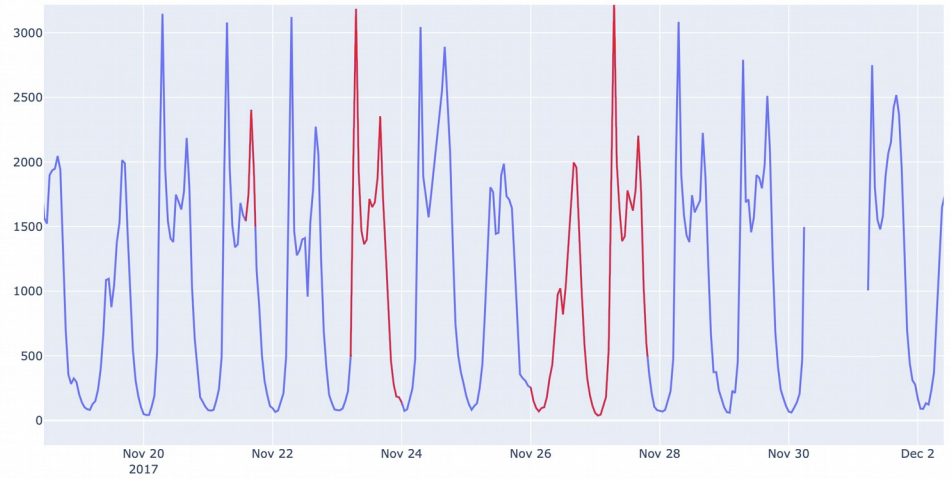


Données reçues

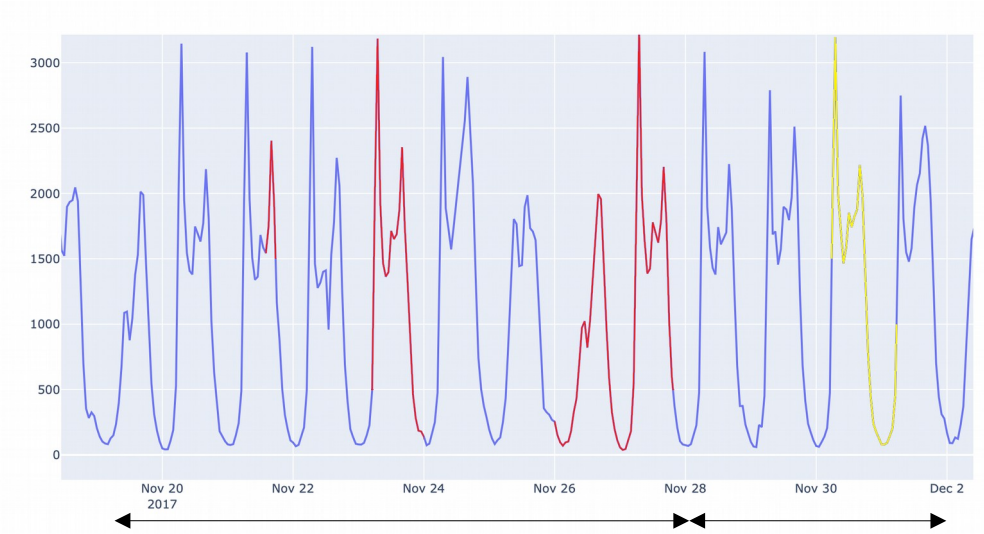
Données complétées

Données complétées en temps réel

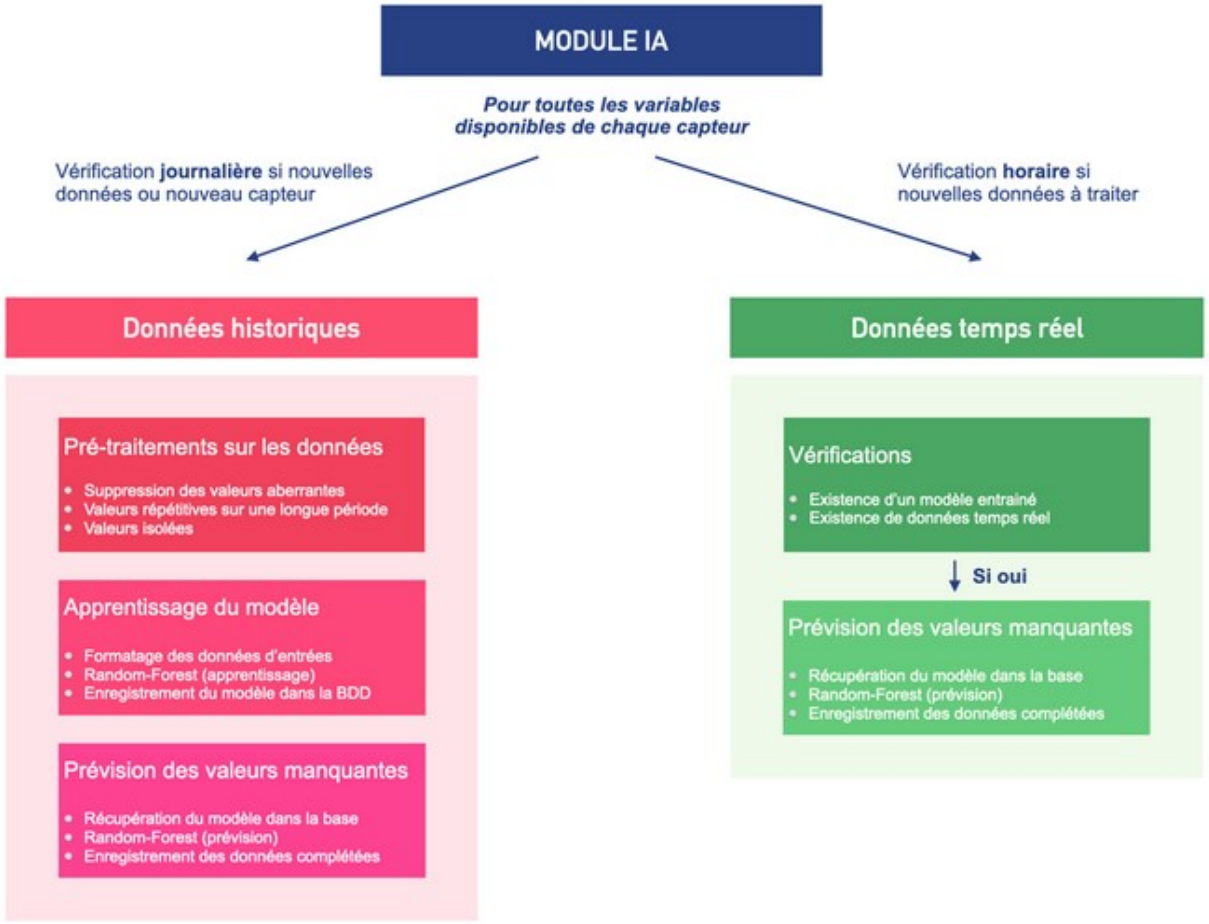
Etape 3



Etape 4



MÉTHODOLOGIE GÉNÉRALE



Méthodologie et synthèse du fonctionnement du module IA

Vérifications

- Journalière : données historiques
  - Nouveaux capteurs ?
- Horaire : données temps réel
  - Nouvelles données ?

3) DÉTECTION DE VALEURS ABERRANTES

En post-traitement (après complétion)

- Indicateur de qualité (de A à G) par mesure de chaque variable

Pour les données initialement présentes :

A : écart **inférieur à 5%** de Dmax

B : écart **entre 5% et 10%** de Dmax

C : écart **entre 10% et 25%** de Dmax

D : écart **supérieur à 25%** de Dmax (évènement exceptionnel suspecté)

E : donnée trivialement **aberrante** supprimée en pré-processing

Pour les données inconnues ou nulles

F : donnée **absente initialement** niveau de confiance non estimable

G : donnée **initialement nulle** supprimée en pré-processing

- Dmax = valeur maximale rencontrée dans l'échantillon

INTELLIGENCE ARTIFICIELLE

RECONSTITUTION DE VALEURS MANQUANTES

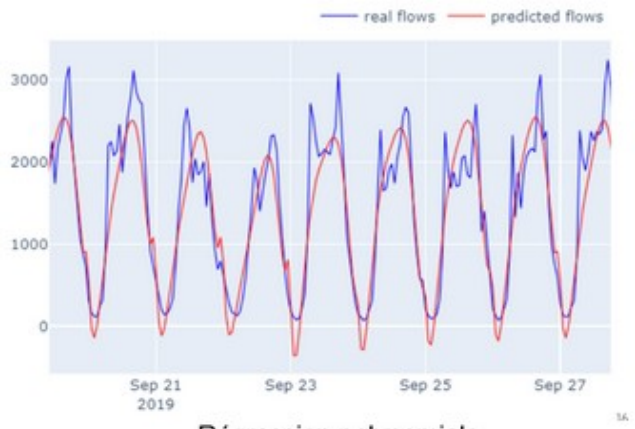
Problème de **régression** :

- Variable cible = variable **numérique** (débit / taux / vitesse)

Méthodes considérées :

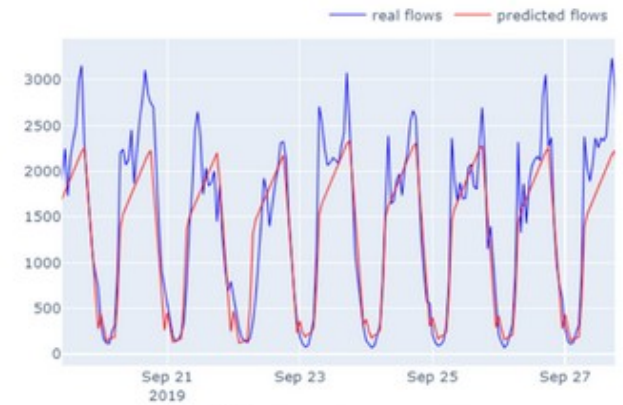
- Méthodes naïves
  - Reproduction selon la temporalité (semaine ou année précédente)
- Machines à vecteur de support (SVR)
- Régression polynomiale
- Réseaux de neurones (MLP)
- Forêts aléatoires (Random Forest)
- Gradient Boosting (CatBoost)
- Méthodes de séries temporelles : ARIMA

RECONSTITUTION DE VALEURS MANQUANTES



Régression polynomiale

Vecteurs supports



Réseaux de neurones

Forêts aléatoires

Analyses comparatives de plusieurs méthodes de machine learning

Quelques observations :

- Régression polynomiale → prédiction de valeurs négatives
- SVR → pointes mal reproduites
- Réseau de neurones (1 seule couche cachée) → pointes du matin mal reproduites

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cerema
CLIMAT & TERRITOIRES DE DEMAIN

CONCLUSIONS ET PISTES

CONCLUSIONS

Méthode de détection et de complétion des données :

- Comparaison de différentes méthodes plus avancées que les méthodes « traditionnelles »
- Algorithme de forêt aléatoire (Random Forest) retenu
- Meilleur compromis entre temps de traitement et résultats

Pistes envisagées :

- Réapprentissage des modèles pour tenir compte des données temps réel archivées au fil du temps
- Amélioration de la robustesse de l'architecture du module IA (règles métiers, couplage de méthodes, tests de sensibilité…)
- Alertes auprès des gestionnaires sur les valeurs aberrantes

CONTACTS

Cyril VEVE

cyril.veve@neovya.fr

Chef de Projet - Data Scientist

NEOVYA Mobility by Technology

Guillaume COSTESEQUE

guillaume.costeseque@cerema.fr

Chargé d'études en mobilités

Cerema Ouest, Nantes

Merci de votre attention

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Cerema
CLIMAT & TERRITOIRES DE DEMAIN