Quantmetry
Part of Capgemini Invent

Reconstruction d'historiques d'affluence pour la prévision des réseaux ferrés urbains

16/11/2023

Julien ROUSSEL
Expert Data Scientist
jroussel@quantmetry.com

Hông-Lan BOTTERMAN
Data Scientist
hlbotterman@quantmetry.com

Nous sommes
The State of the Art AI company

Le cabinet de conseil de référence en Intelligence Artificielle, pure player, créé il y a plus de 10 ans

UN CABINET DE CONSEIL **Français**

150
Collaborateurs et consultants-chercheurs

+500
Missions IA et Data réalisées

+50
articles par an

+15
prix innovation et recherche

l'institut Quantmetry

yotta AI ACADEMY

Animé par 4 valeurs : Excellence | Exceptionnalité | Accomplissement | Esprit d'équipe

Agenda du jour

1. **Présentation du projet Aifluence**
2. **Gestion de valeurs manquantes**
   Aspects théoriques
3. **Affluence et valeurs manquantes**
   Reconstruction d'historique d'affluence pour la prédiction
4. **Présentation du package open-source Qolmat**
   Imputation de données manquantes

Les enjeux clés traités pendant la réalisation du projet

Préparation de données

- Détection d'anomalie
- Imputation de charges manquantes
- Reconstruction de la charge à bord à partir des validations

Forecast de confiance

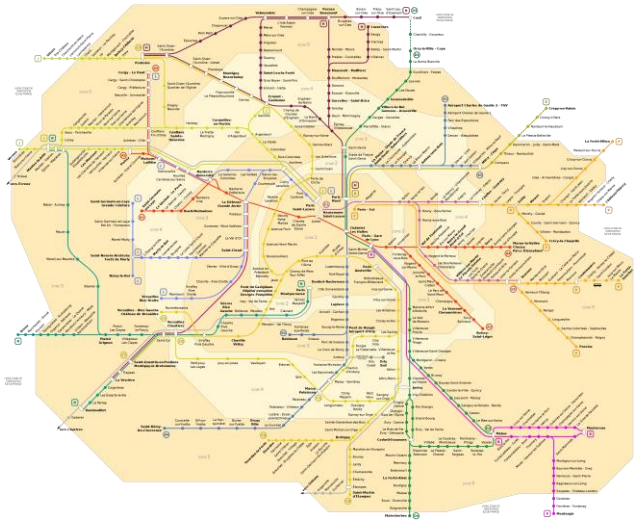
- Estimation de l'incertitude pour les séries temporelles
- Explicabilité pour la prévision des séries temporelles

Application

- Prévision court et moyen terme de la charge à bord
- Prévision moyen terme de l'affluence dans les gares

Alfluence : Prédictions d'affluence sur le réseau Transilien

Un patrimoine de données volumineux



Préparation des données

- **Validations des pass en gare**
- **Montées / descentes à bord**
- Plan de transport théorique

6 cas d'usage

- **Prédiction** de l'affluence en gare*
- **Prédiction** de la charge à bord*
- Reconstruction de données manquantes

X2 * à H+24 ou jusqu'à J+7

Quantmetry
Part of Capgemini Invent

Parties prenantes du projet

5 entités impliquées avec chacune leurs rôles et leurs responsabilités

Quantmetry

- Pilote le développement du projet
- Développe les algorithmes prédictifs

ENS Paris-Saclay

- Apporte son expertise scientifique
- Valide la démarche scientifique

SNCF Recherche & Innovation

- Apporte son expertise métier au développement des algorithmes
- Joue les intermédiaires dans l'obtention des données

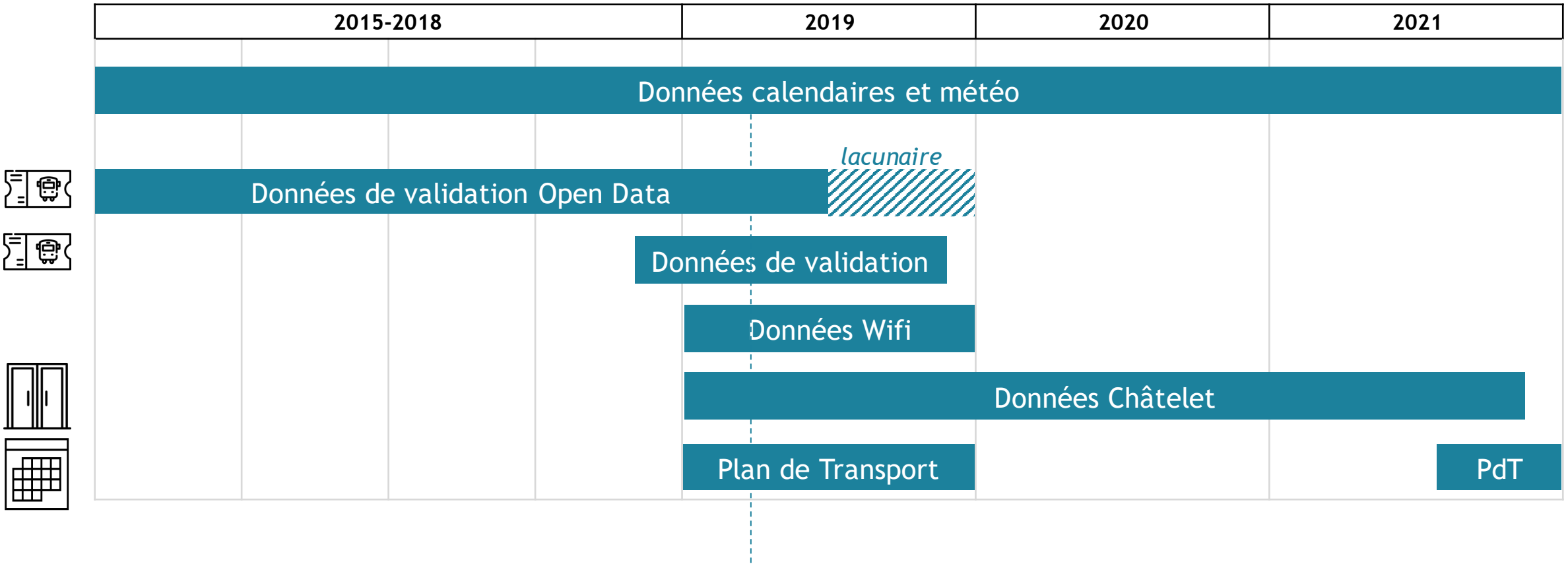
Région Ile-de-France

- Organisateur du concours
- Subventionne la moitié du projet

SNCF Gares & Connexions Transilien

- Fourni les données en gare (WIFI public)
- Fourni les données à bord des trains

Des données disponibles sur différentes périodes



Des données partiellement indisponibles

Deux types de données d'affluence

Validation en gare

Intérêt

- Données permettant une mesure de flux en entrée de gare
- Maille : 15min / gare

Limitations

- **Anomalies** à corriger sur certaines périodes
- Périodes manquantes

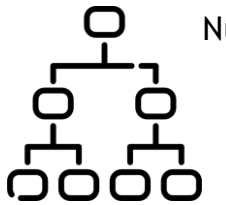
Montées / Descentes

Opportunités

- Données permettant de mesurer la charge à bord
- Maille : arrêt en gare

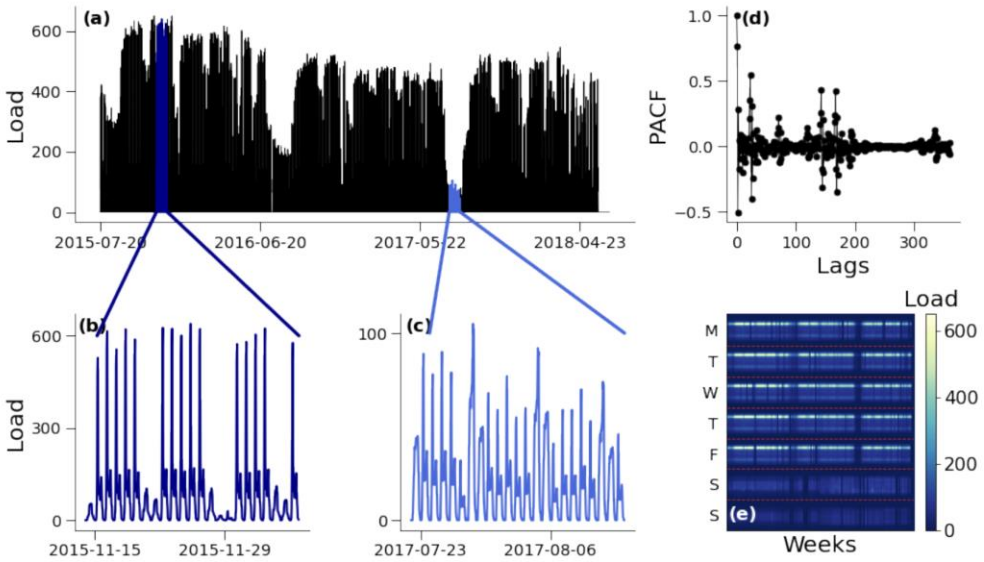
Limitations

- Plusieurs motifs de valeurs manquantes



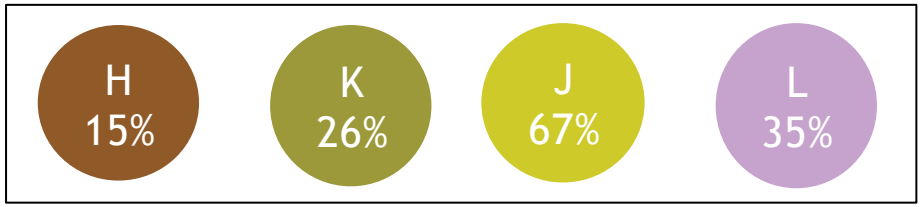
Distribution des données manquantes sur les différentes lignes

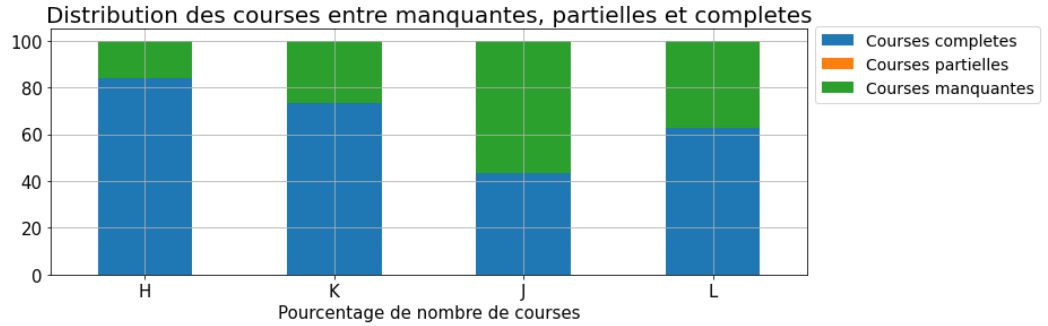
Validations en gare

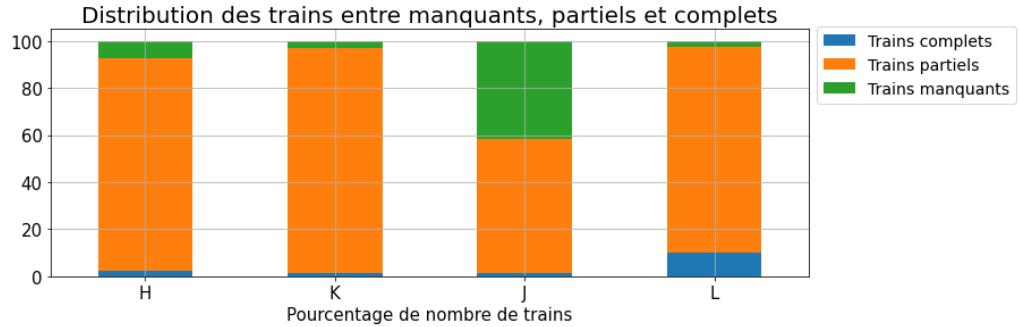


Montées / Descentes

Pourcentage des valeurs manquantes par ligne







Quantmetry
Part of Capgemini Invent

Agenda du jour

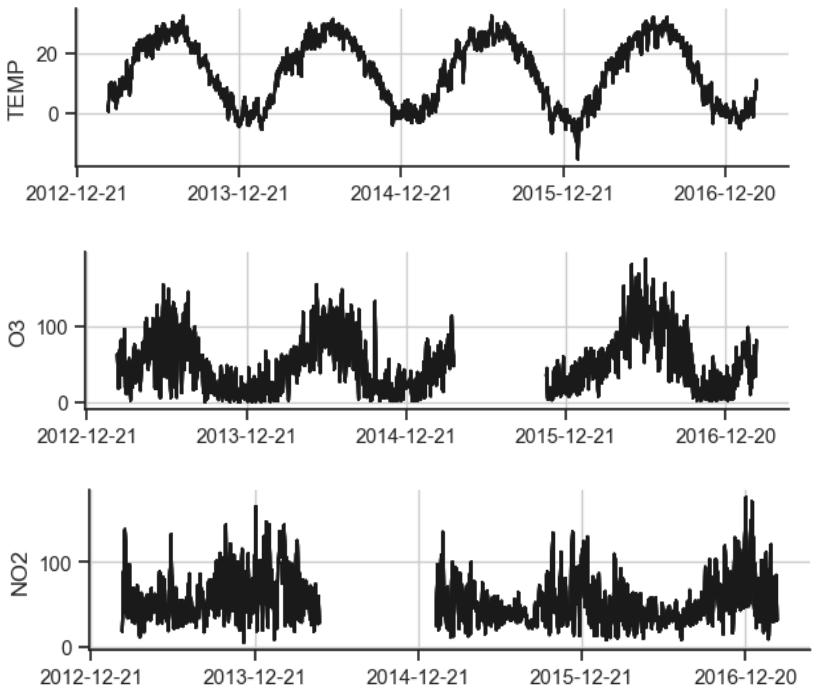
1. **Présentation du projet Aifluence**
2. **Gestion de valeurs manquantes**
   Aspects théoriques
3. **Affluence et valeurs manquantes**
   Reconstruction d'historique d'affluence pour la prédiction
4. **Présentation du package open-source Qolmat**
   Imputation de données manquantes

Les données manquantes sont incontournables

Données tabulaires

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

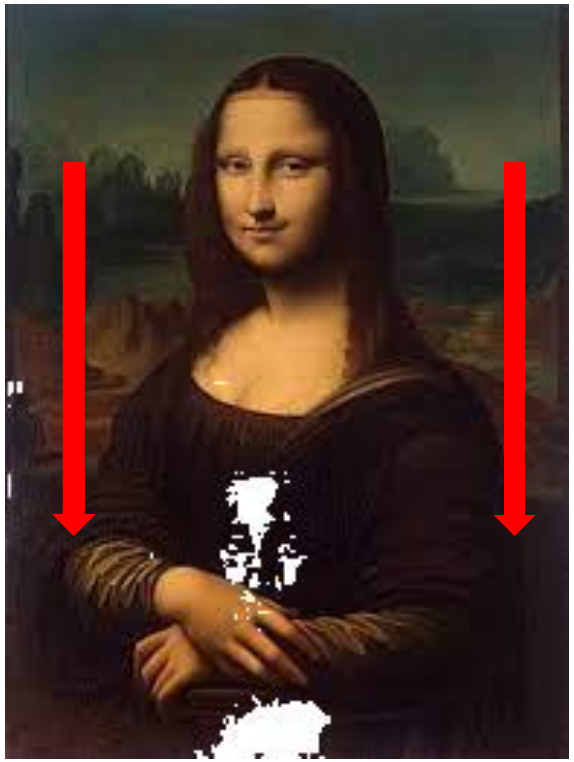
Données temporelles



Trois types de données manquantes

MCAR
Missing Completely At Random

Les trous sont indépendants des données observables
(rectangles tirés selon une loi uniforme)



MAR
Missing At Random

Les trous ne sont pas indépendants des données observables
(les zones les plus rouges rendent ceux placés en-dessous plus susceptibles d'être manquants)

MNAR
Missing Not At Random

Les trous ne sont ni des MAR ni MCAR
(les pixels les moins rouges sont plus susceptibles d'être manquants)

Trois types de données manquantes

MCAR
Missing Completely At Random

Une donnée ou un ensemble de données ont disparu pour des raisons indépendantes de l'équipement :

- défaillance d'un capteur
- diagnostic annulé pour des raisons organisationnelles
- donnée perdue

Pas de corrélation avec les données précédemment acquises.

$$p(M|Y) = p(M)$$

MAR
Missing At Random

Des données sont manquantes pour une raison dépendante à la situation de l'équipement :

- arrêt du suivi
- examen annulé pour des raisons liées à son état

Corrélation avec les données précédemment acquises mais pas avec la valeur manquante.

$$p(M|Y) = p(M|Y_{obs})$$

MNAR
Missing Not At Random

Des données sont manquantes du fait de leur valeur :

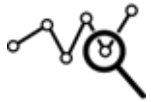
- mesure hors des seuils de détection
- données aberrantes

Corrélation avec les données précédemment acquises et avec la valeur manquante.

M est la matrice d'indication des valeurs manquantes. $Y=\{Y_{obs}, Y_{miss}\}$ est la matrice des données.

Rubin, Donald B. Inference and missing data. Biometrika 63.3 (1976): 581-592.

Caractérisation du problème d'imputation



Types de valeurs manquantes

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

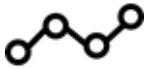
Types de lacunes

- **Unit non-respond** - ligne entière
- **Item non-respond** - une variable

Quel est le contexte dans le cas temporel ?

Dimensionalité

- **Univariée** - une variable unique
- **Multivariée** - plusieurs variables observées



Aspects temporels de la série

- **Saisonnalité** - différents cycles
- **Tendance** - présence ou non

Méthodes d'imputation

- **Forward** - valeurs ultérieures inconnues
- **Backward** - valeurs ultérieures connues

Gestion des valeurs manquantes : typologie

Fiabilité des process d'acquisition de données

- gouvernance de la donnée
- standardisation des process de collecte
- exclure les variables non pertinentes
- conception des questionnaires

Structuration de la donnée

- série temporelle régulière (ex : une ligne par patient et par mois)
- série temporelle non-régulière
- redéfinition des variables catégorielles

Suppression des données manquantes

- suppression de lignes (exemples) contenant de nombreuses valeurs manquantes
- suppression de colonnes (variables) contenant de nombreuses valeurs manquantes

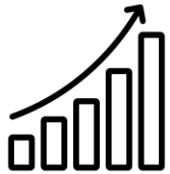
Imputation des données manquantes

- méthodes rudimentaires
- méthodes avancées
- méthodes spécifiques aux séries temporelles

Suppression ou imputation ?

Visualiser la donnée avant et après tout traitement des valeurs manquantes :

- certaines variables (sensibles) sont-elles plus touchées ?
- les trous sont-ils distribués uniformément ?



Doit-on supprimer les lignes avec des valeurs manquantes ?

Si on prend des données avec n lignes et m colonnes présentant 1% de valeurs manquantes.

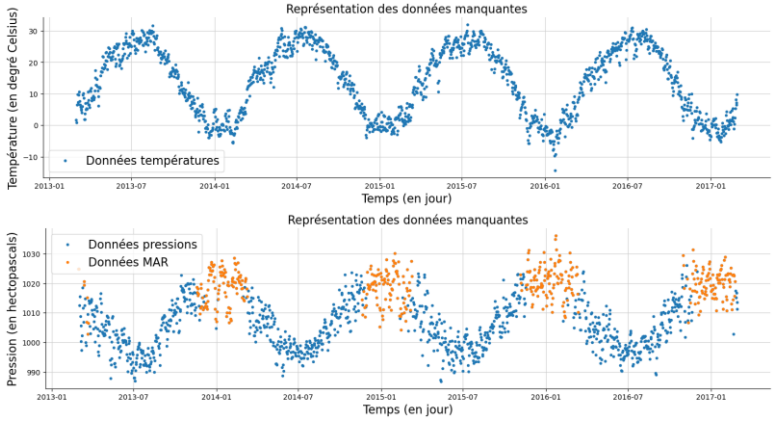
- Si m=5 alors environ 95% des lignes seront gardées ;
- Si m=300 alors environ 5% des lignes seront gardées.

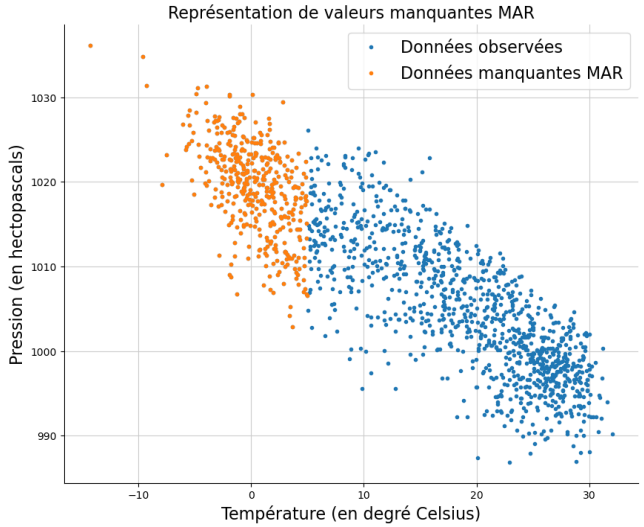
Et dans le cas de série temporelle, la structure régulière est détruite !

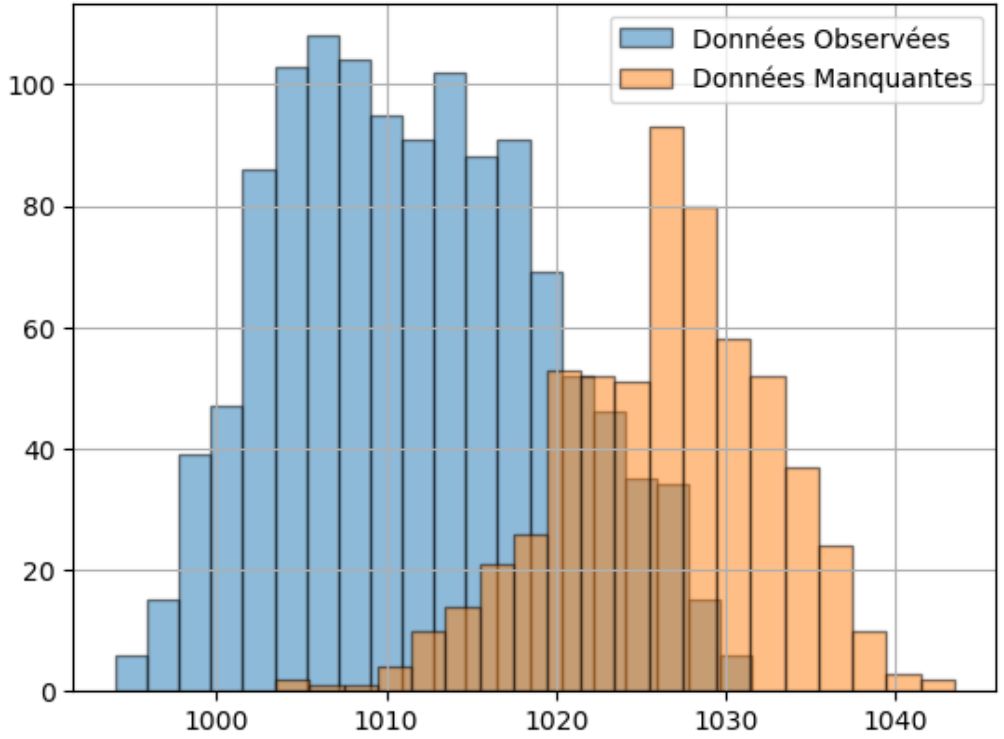
En pratique, l'imputation donne souvent de meilleurs résultats que la suppression/omission

Exemple sur des données réelles

On extrait des données de **températures et pressions de 2013-2017** à Gucheng en Chine échantillonnées jour par jour. On oublie les valeurs de la **température < 5°C** (MAR)







Comment imputer les valeurs de pressions manquantes sans introduire de biais ?
Comment préserver la distribution statistique des données ?

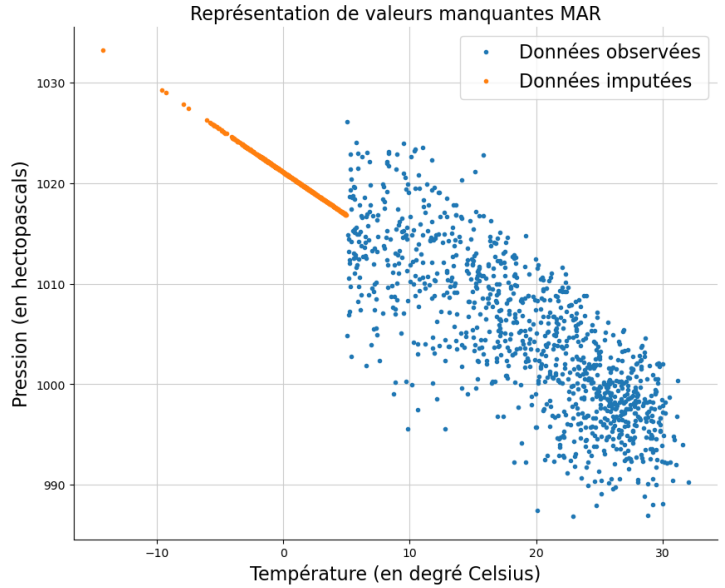
Quantmetry
Part of Capgemini Invent

Imputation par approche linéaire

Approches linéaires

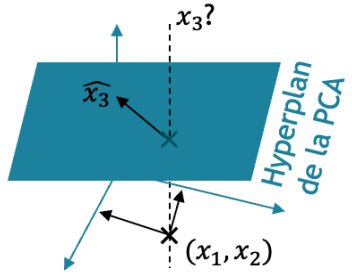
- Régression linéaire
  pour chaque variable manquante on entraîne un modèle de régression linéaire et on la prédit à partir de une ou plusieurs variables auxiliaires sans trous

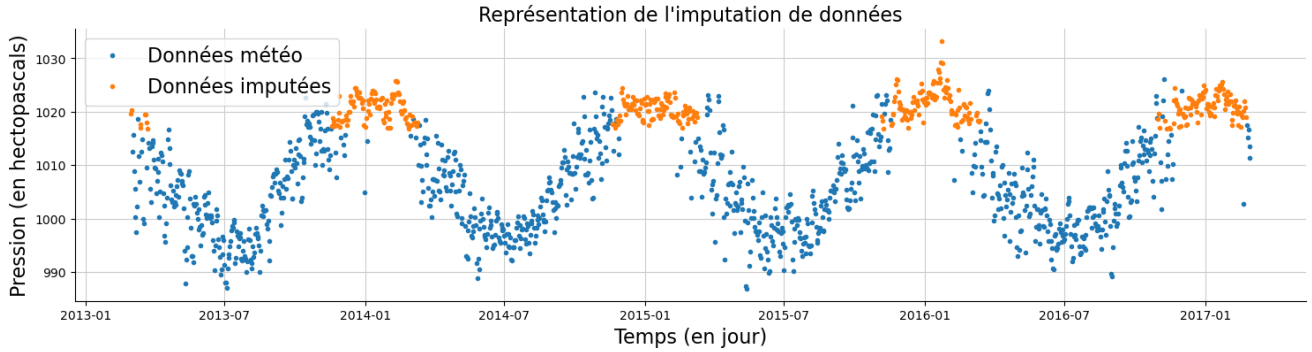
- PCA (Principal Component Analysis)
  on impute toutes les variables en même temps, grâce à un unique modèle de réduction de dimension



… et leurs limites

- Non robuste aux valeurs extrêmes
- Réduit la variance de l'échantillon
- Nécessite d'avoir des covariables sans trous
- Pas de prise en compte des tendances temporelles



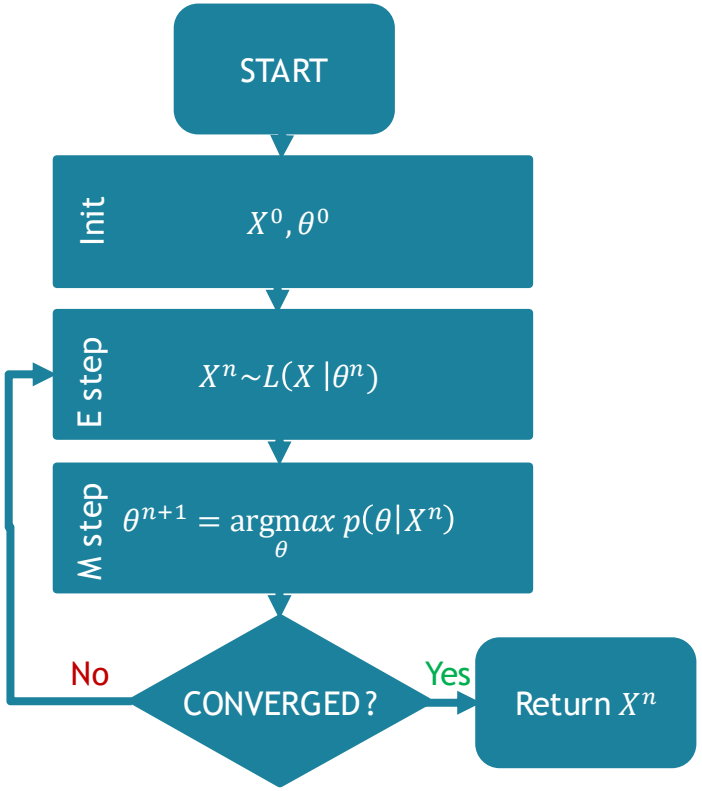


Quantmetry
Part of Capgemini Invent

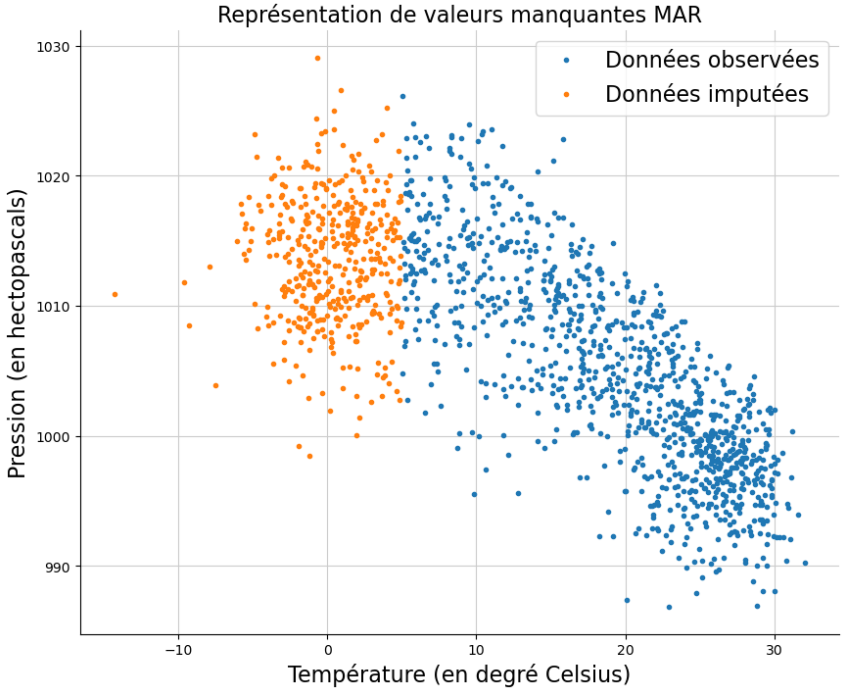
Echantillonnage par Expectation-Maximisation

Méthode d'échantillonnage Bayésien

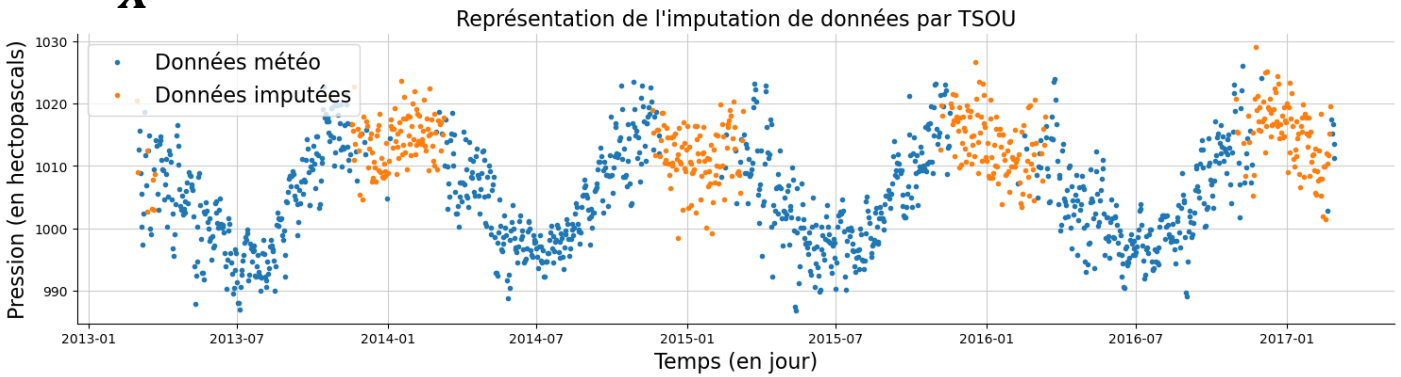
Il s'agit d'un algorithme d'optimisation qui **maximise la log vraisemblance des données complètes**, par des moyens itératifs sous la distribution conditionnelle des données non observées.



Conservation des corrélations entre variables



X^n



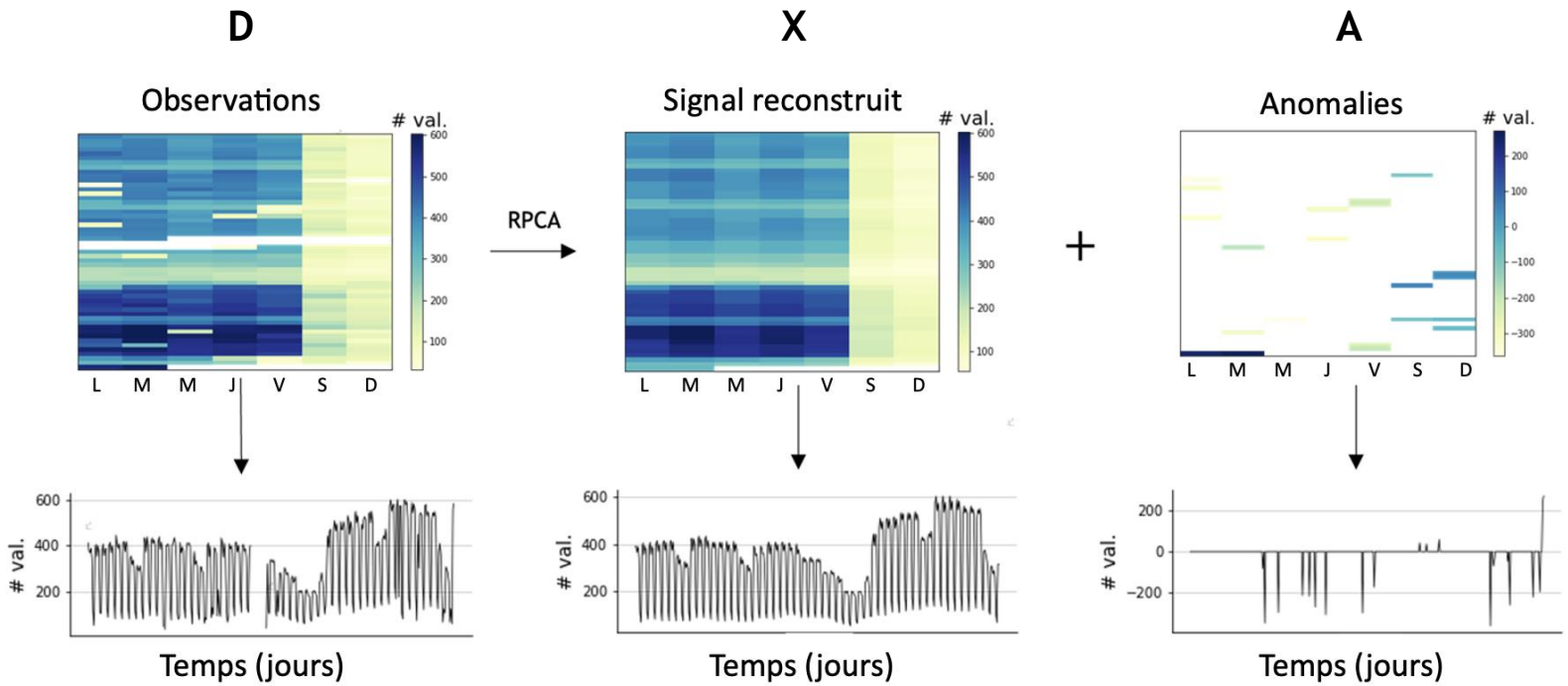
RPCA

Méthode statistique robuste

Dans le cas de signaux **saisonnier** ou de données **multivariées** l'algorithme RPCA permet de retirer les **anomalies** et d'**imputer** les trous.

$$\min_{\substack{X,A \\ D=X+A}} \|X\|_* + \lambda\|A\|_1$$

- **Robustifie** les traitements (entraînement de modèles, visualisation, statistiques, …)
- S'adapte au contexte, grâce à plusieurs variantes de l'algorithme
- Adapté pour les motifs de données manquante MAR

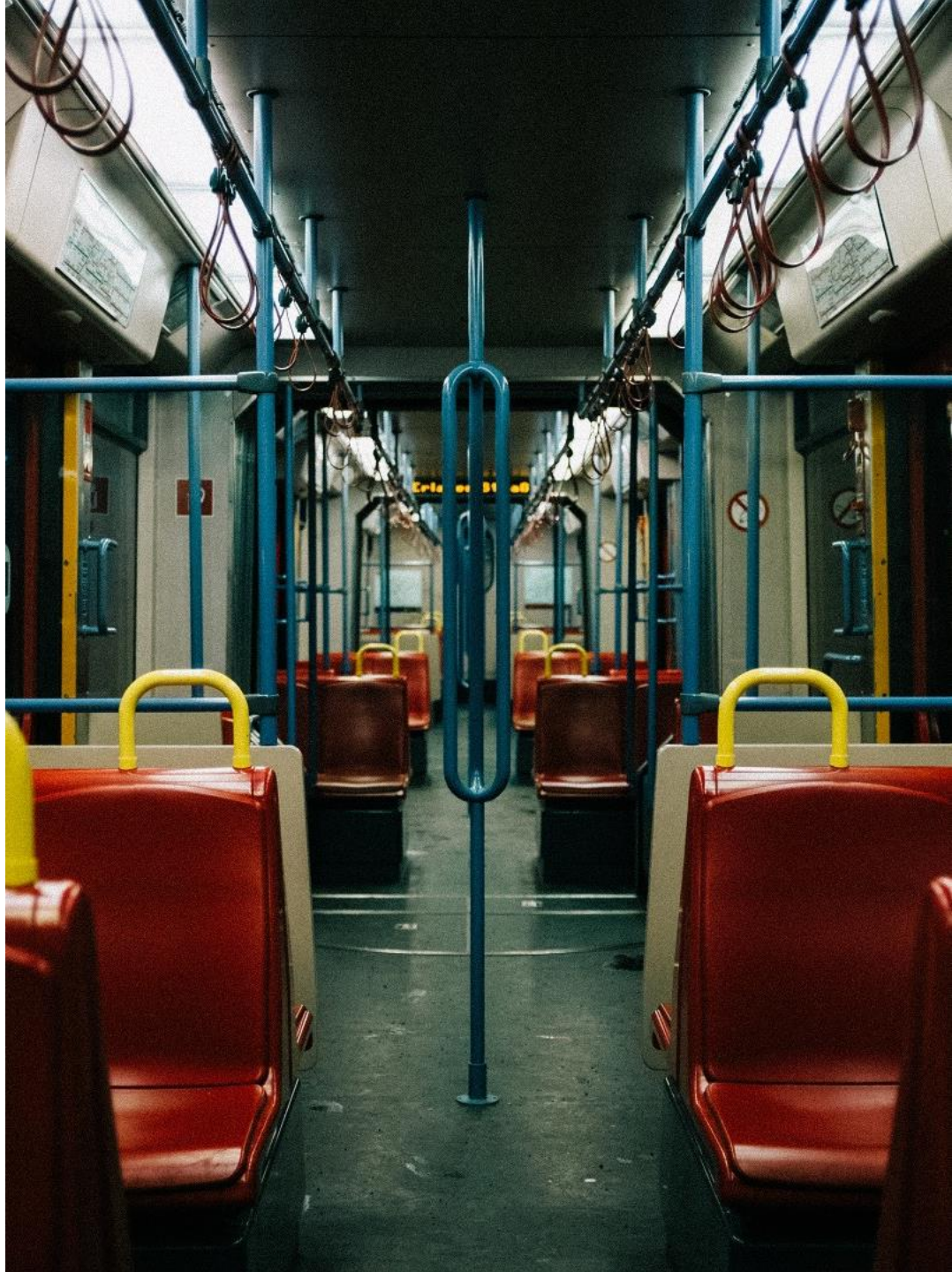


Quantmetry
Part of Capgemini Invent

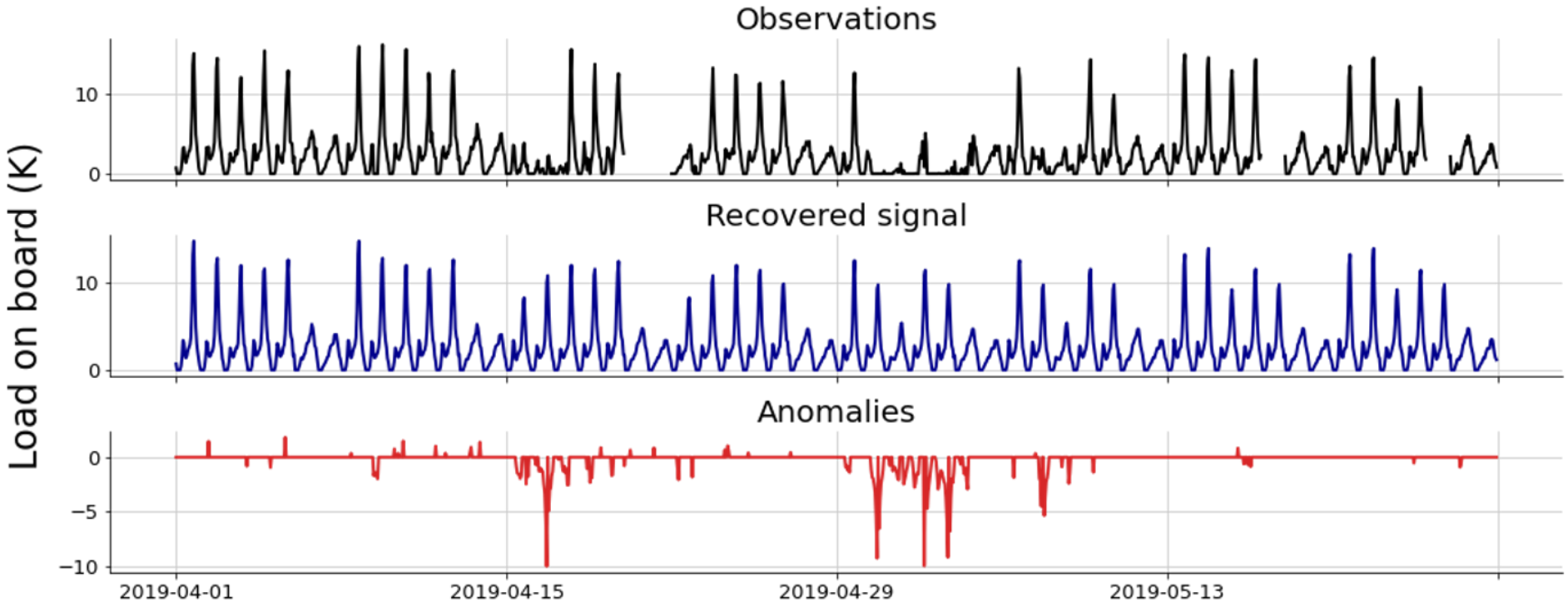
Botterman, H. L., Roussel, J., Morzadec, T., Jabbari, A., & Brunel, N. (2022). Robust PCA for Anomaly Detection and Data Imputation in Seasonal Time Series. arXiv preprint arXiv:2208.01998.

Agenda du jour

1. **Présentation du projet Aifluence**
2. **Gestion de valeurs manquantes**
   Aspects théoriques
3. **Affluence et valeurs manquantes**
   Reconstruction d'historique d'affluence pour la prédiction
4. **Présentation du package open-source Qolmat**
   Imputation de données manquantes

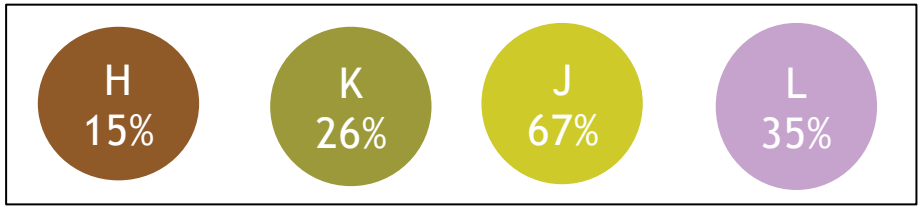


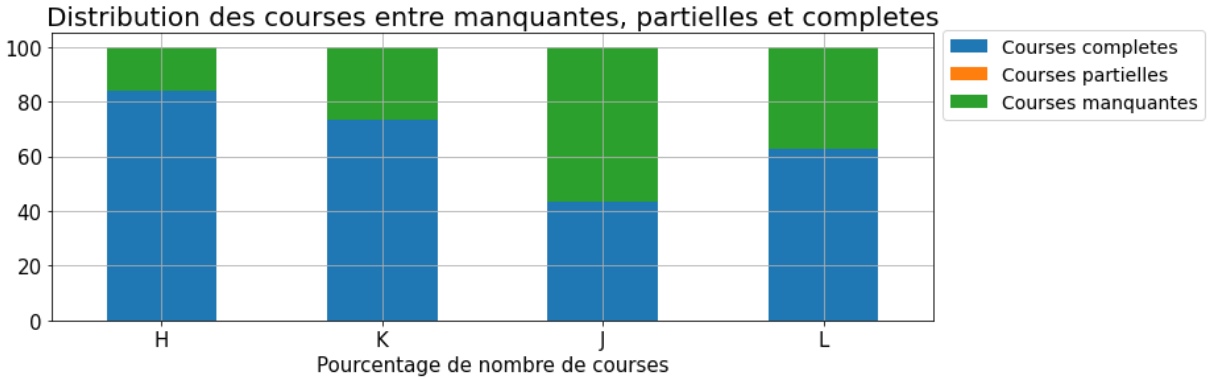
Imputation des données de validation avec la RPCA

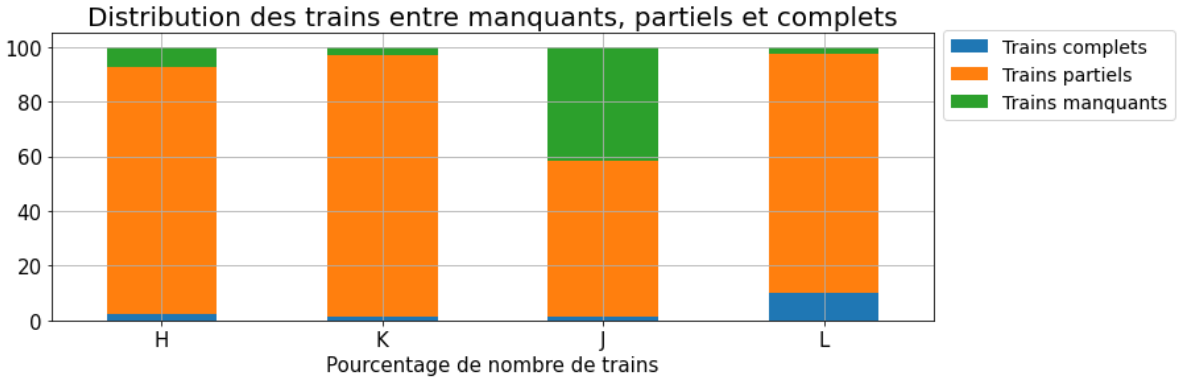


Comment imputer les données de montées / descentes ?

Pourcentage des valeurs manquantes par ligne







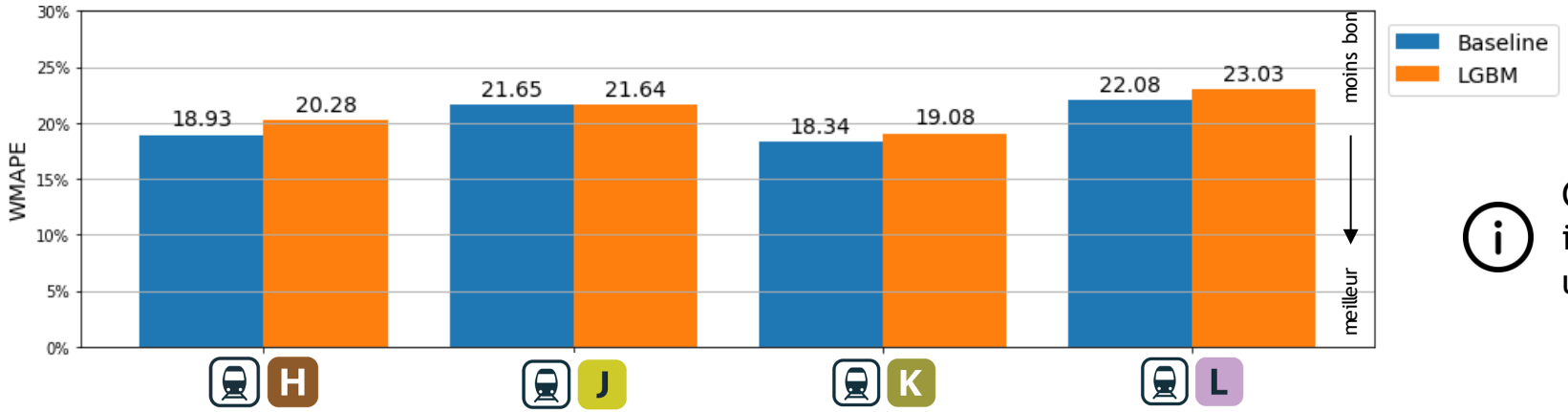
➢ Distribution des arrêts manquants non uniformes

Hypothèse

Les courses et les trains manquants sont répartis de façon
- MCAR
- uniforme

Quantmetry
Part of Capgemini Invent

Reconstruction à la maille course



Chaque ligne est traitée indépendamment : un modèle par ligne

Modèle « moyenne glissante »

Inspiré par le modèle prédictif de la SNCF basé sur la moyenne des charges disponible pour le même type de jour, station et heure.

Modèle LGBM

Modèle ensembliste (Gradient Boosting) utilisant notamment plusieurs variables exogènes (charge à bord du même train les autres jours, autres historiques de charge, destinations, mission, données calendaires, données de validation, …)

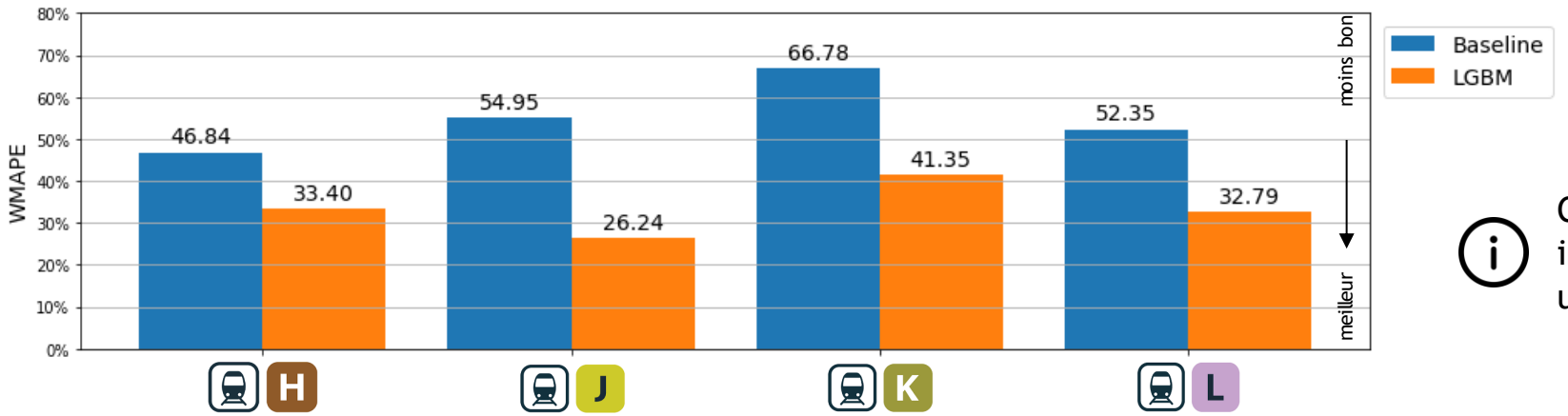
Méthode

Pour chaque ligne on sélectionne 50% des courses : on masque toutes les charges correspondantes puis on reconstruit toutes les données manquantes de la ligne. Cela est répété et moyenné 5 fois.

- L'utilisation de ce modèle LGBM apporte un faible gain en performance dans ce cas.

Reconstruction à la maille train

Approche ligne par ligne



Chaque ligne est traitée indépendamment :
un modèle par ligne

Modèle « moyenne glissante »

Inspiré par le modèle prédictif de la SNCF basé sur la moyenne des charges disponible pour le même type de jour, station et heure.

Modèle LGBM

Modèle ensembliste (Gradient Boosting) utilisant notamment plusieurs variables exogènes (charge à bord des autres trains, destinations, mission, données calendaires ainsi que les données de validation)

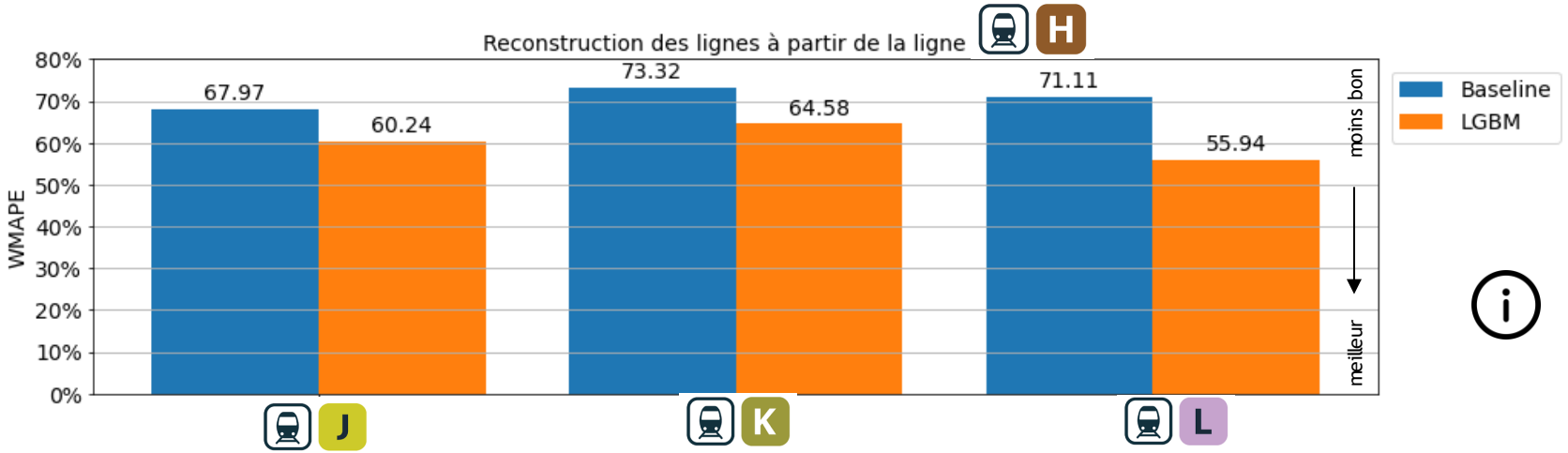
- L'utilisation de ce modèle LGBM apporte un gain très significatif

Méthode

Pour chaque ligne on sélectionne 50% des trains : on masque toutes leurs mesures puis on reconstruit toutes les données manquantes de la ligne. Cela est répété et moyenné 5 fois.

Reconstruction à la maille ligne

Apprentissage par transfert



Les deux modèles Baseline et LGBM sont entrainés sur la ligne H et la reconstruction est faite pour chacune des lignes

Modèle « moyenne glissante »

Moyenne des charges sur la ligne H par heure, jour de la semaine et type de jour.

Modèle LGBM

Modèle ensembliste (Gradient Boosting) utilisant notamment plusieurs variables exogènes (charge à bord de la ligne H, données calendaires, ainsi que les données de validation). **Pas de données de charge à bord historique utilisées**

- L'utilisation de ce modèle LGBM apporte un gain significatif, bien que l'erreur reste élevée

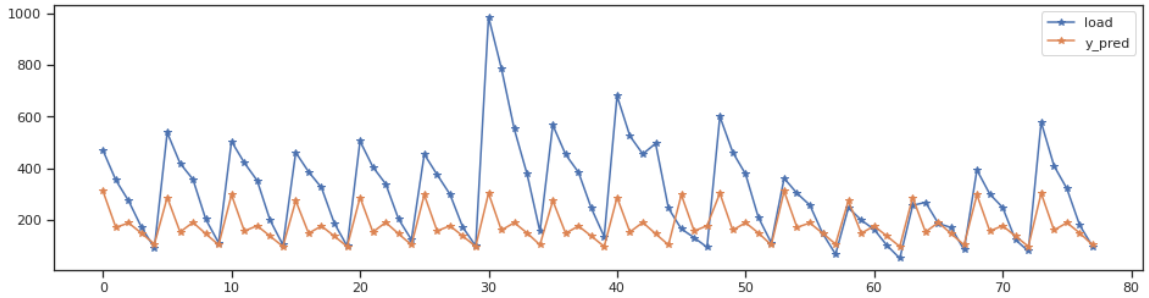
Méthode

Pour chaque ligne reconstruite : on masque toutes les charges correspondantes puis on reconstruit toutes les données manquantes de la ligne en se basant sur un modèle qui a appris à partir de la ligne H

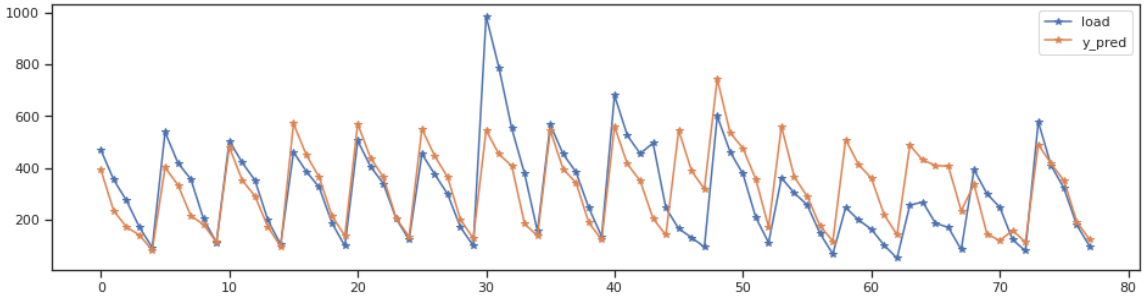
Exemple de reconstruction de deux trains de la ligne J à partir de la ligne H

Train: 136725

Baseline

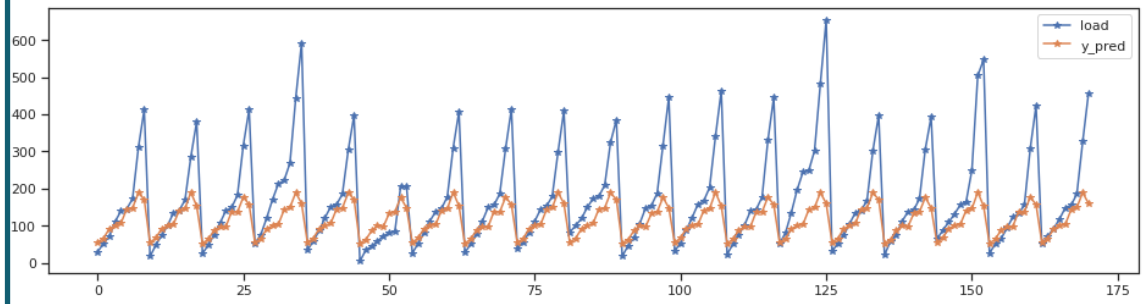


LGBM

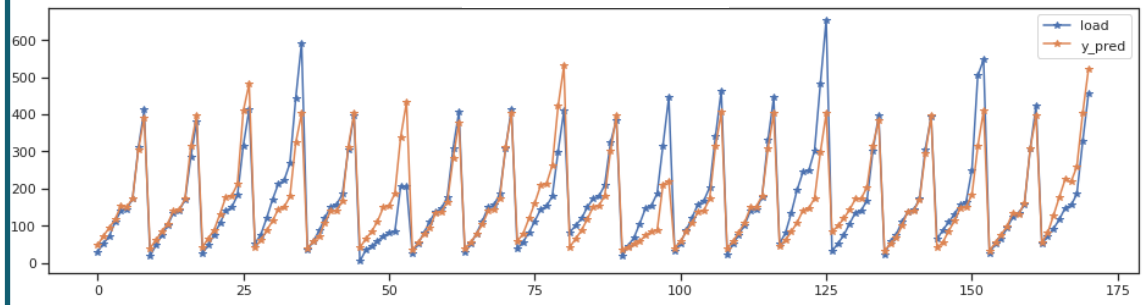


Train: 136725

Baseline



LGBM



Baseline

Quantmetry
Part of Capgemini Invent

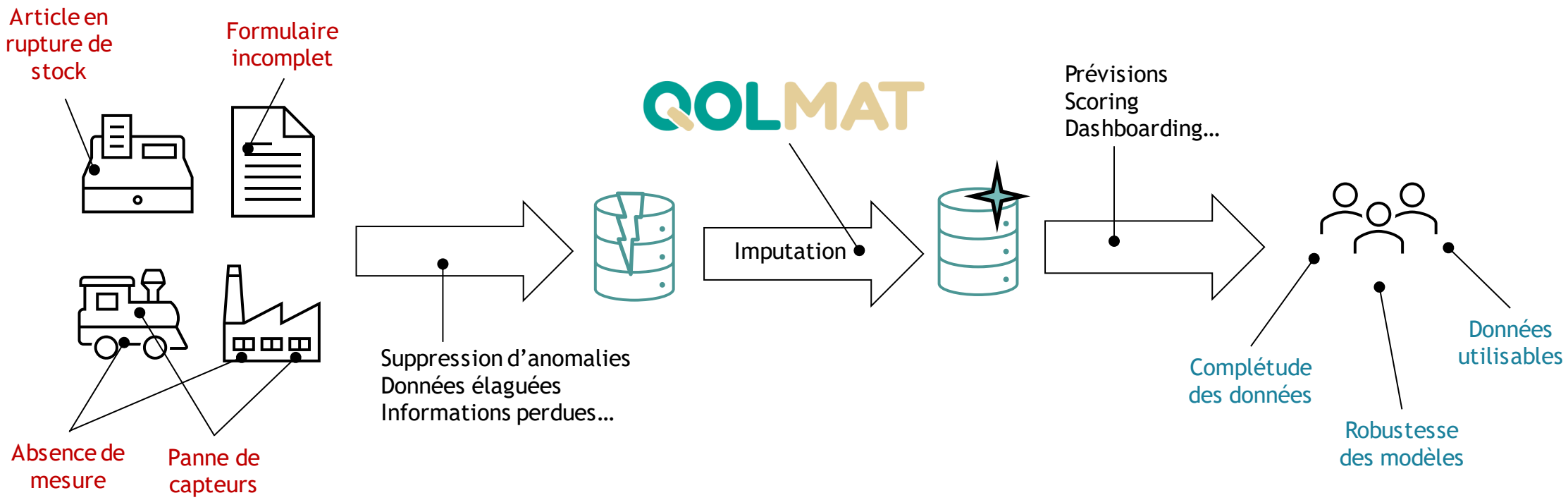
Agenda du jour

1. **Présentation du projet Aifluence**
2. **Gestion de valeurs manquantes**
   Aspects théoriques
3. **Affluence et valeurs manquantes**
   Reconstruction d'historique d'affluence pour la prédiction
4. **Présentation du package open-source Qolmat**
   Imputation de données manquantes

La qualité des données est un frein à leur valorisation

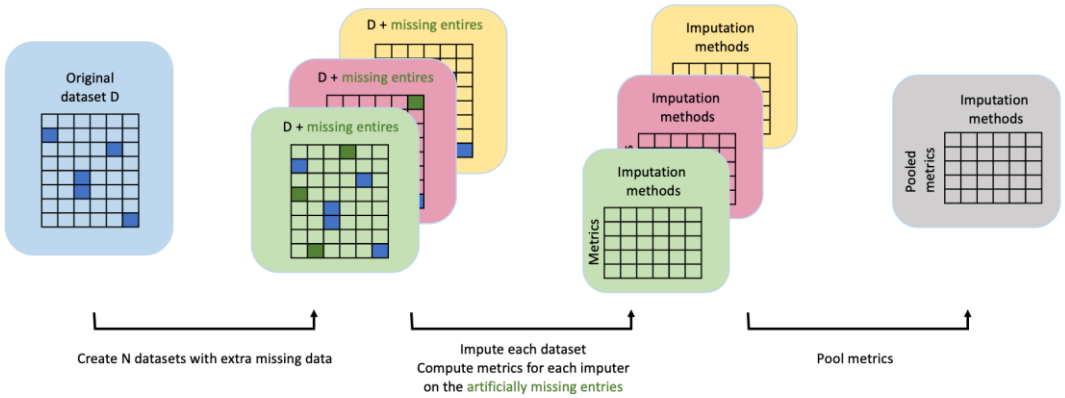
Une **mauvaise collecte** engendre des données incomplètes
mais des méthodes à l'état de l'art peuvent **restaurer la confiance des utilisateurs** !

Problèmes	Conséquences	Solution	Résultats



Benchmarking

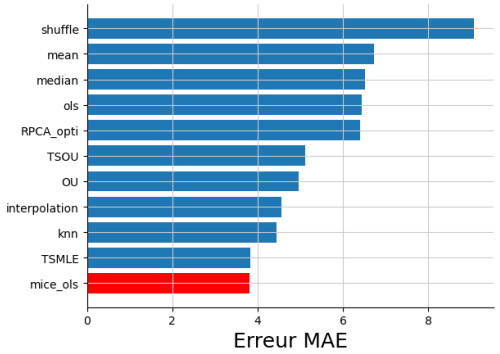
Erreurs de reconstruction estimées sur une validation croisée

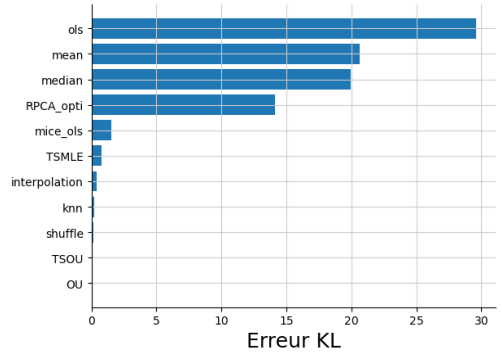


Méthodologie

- La loi de génération des trous doit être représentative de la réalité
- L'imputation finale est recalculée sans ajouter de trous
- Le nombre de trous ajoutés doit être assez grand pour avoir un estimateur fiable pas assez petit pour ne pas trop modifier le taux de valeurs manquantes

- **Optimisation adaptative des hyperparamètres** des méthodes d'imputation
- Choix du **modèle de génération** de **données manquantes**
- **Plusieurs métriques** de comparaison :

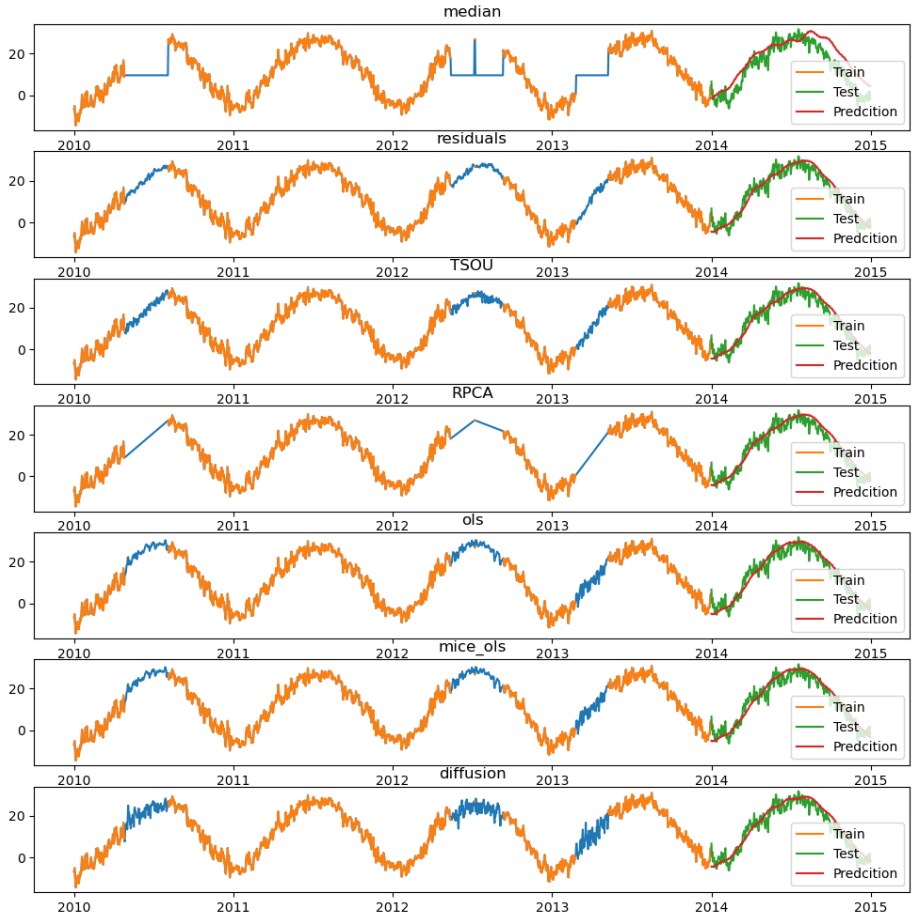




Comment définir la qualité d'une imputation ? Faut-il reproduire les fluctuations observées dans les données, ou bien les minimiser ?

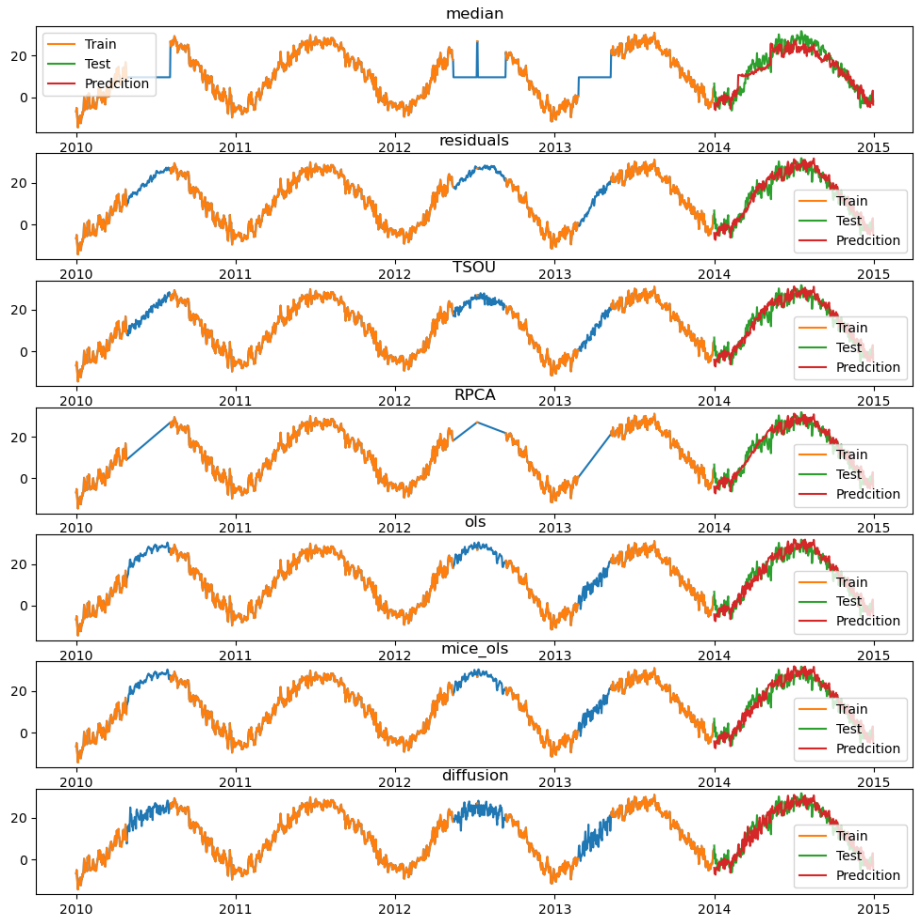
Quel impact sur les performances de prédiction ?

Prophet



	median	residuals	TSOU	RPCA	ols	mice_ols	diffusion
mae	4.905450	2.386468	2.488762	2.563329	2.366268	2.368731	2.458612

Exponential Smoothing



	median	residuals	TSOU	RPCA	ols	mice_ols	diffusion
mae	3.157005	2.463152	2.452445	2.389292	2.585602	2.585601	2.558953

Quantmetry
Part of Capgemini Invent

Pourquoi Qolmat ?

Une imputation négligée peut avoir de **forts impacts sur la performance finale** des modèles

Un **enjeu de poids** lors de **l'industrialisation** des modèles sur des données réelles

Les outils disponibles en **python** sont **dispersés**, et nécessitent un effort de prise en main

La **qualité de l'imputation** est très rarement évaluée

QOLMAT

Objectifs

- Qolmat est une libraire python qui permet :
  - **d'uniformiser** les méthodes d'imputation
  - de les **comparer et de recommander** des usages,
  - d'en **estimer les hyperparamètres** pour **optimiser** les performances
  - de **s'intégrer** dans les pipelines standards.

- **Tutoriel** notebook pour une prise en main en 30min
- **Documentation** détaillée

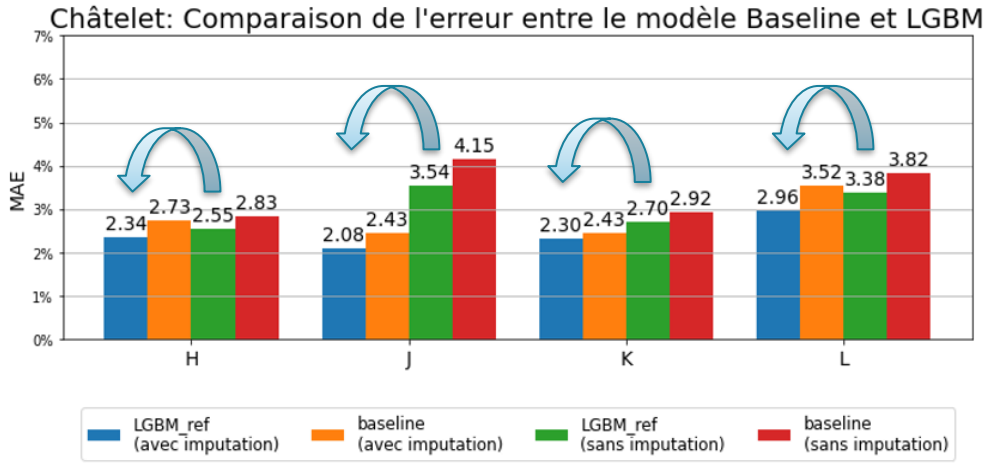
Implémentation

- Mise en place **rapide** (compatibilité **Scikit-learn**)
- Imputation par **régression** : linéaire, gradient boosting, avec ou sans fluctuation
- RPCA, échantillonnage **Bayésien**, …
- Interpolation, moyennes glissantes, …

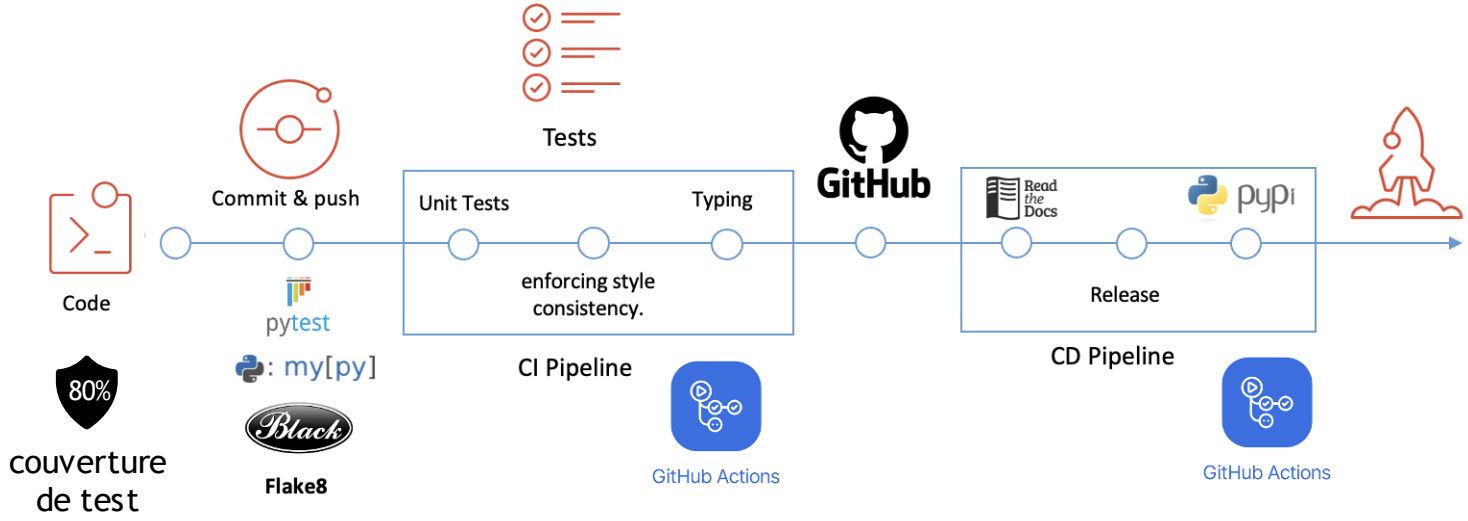
Et en pratique ?

Des performances…

L'imputation a permis de réduire l'erreur de prédiction de **10%** en mission.



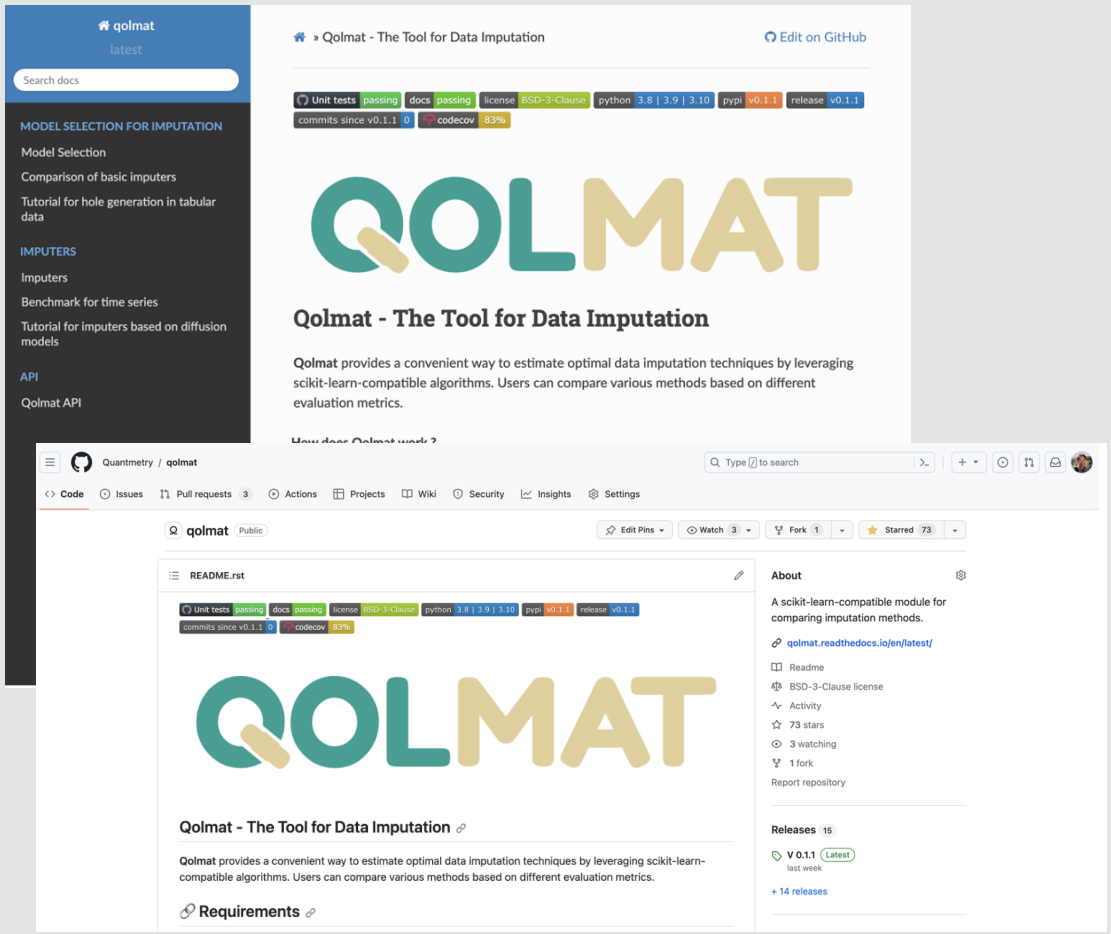
dans un package open-source robuste !



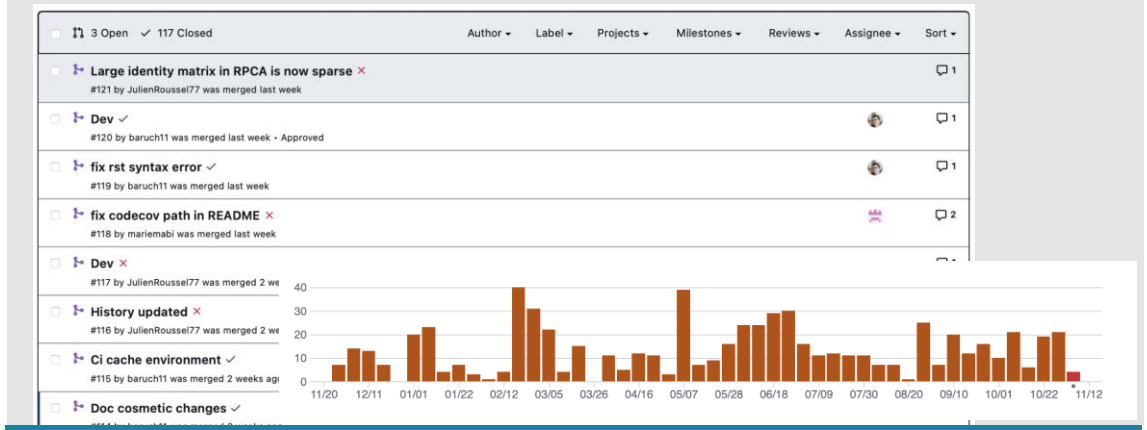
Quantmetry
Part of Capgemini Invent

Ecosystème de Qolmat

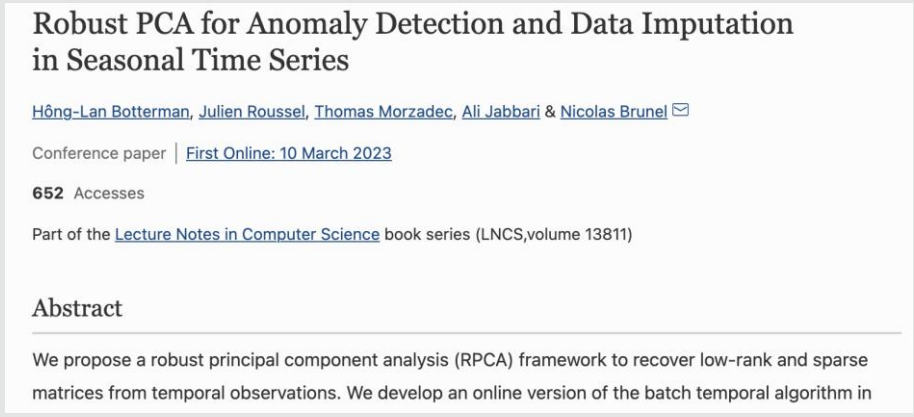
Github du package



Maintenance du code



Publications Scientifiques



Robust PCA for Anomaly Detection and Data Imputation in Seasonal Time Series

Hông-Lan Botterman, Julien Roussel, Thomas Morzadec, Ali Jabbari & Nicolas Brunel

Conference paper | First Online: 10 March 2023

652 Accesses

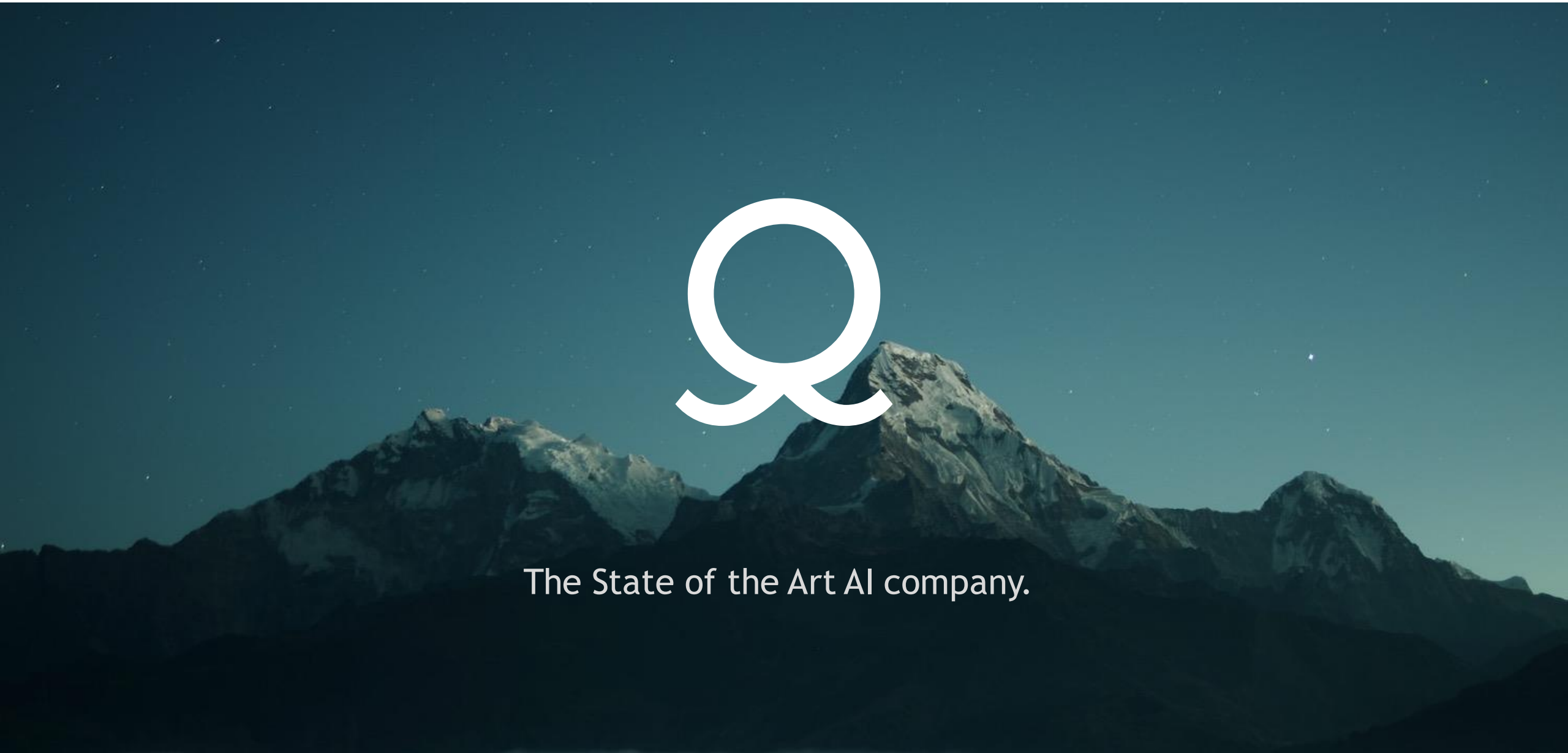
Part of the Lecture Notes in Computer Science book series (LNCS,volume 13811)

Abstract

We propose a robust principal component analysis (RPCA) framework to recover low-rank and sparse matrices from temporal observations. We develop an online version of the batch temporal algorithm in

Quantmetry
Part of Capgemini Invent

Quantmetry



Plusieurs packages d'imputation de données

Packages	Temporel	Catégoriel	Multivarié	Multiple/Simple	Langage	Méthodes	Evaluation	Compatible **Scikit**
Qolmat	☑	Bientôt	☑	multiple	python	Stats simples, EM, RPCA, diffusion **models…**	☑	☑
R-miss-tastic	☑	☑	☑	simple	R	Beaucoup	☑ (sans V.C.)	✗
missingpy	✗	☑	☑	multiple	python	• missForest • KNN	✗	☑
amelia	☑	☑	☑	multiple	R	EM (jointly multi. Normal)	✗	✗
missMDA	✗	☑	☑	multiple	R	Méthodes d'analyses factorielles : ACP, ACM	✗	✗
sklearn (fillna)	✗	☑	✗	simple	python	"de base"	✗	☑
sklearn (IterativeImputer)	✗	✗	☑	simple	python	Estimateur sklearn	✗	☑
miceForest	✗	☑	☑	multiple	python	MICE RF	✗	✗
numpyro (Baysian imputation)	✗	✗	☑	multiple	python	Sampling MCMC + multiple kernel	✗	✗
impyute	☑	✗	☑	multiple	python	MICE EM "de base"	✗	✗
PyPots	☑	☑	☑	multiple	python	DL : SAITS, CSDI, BRITS, …	✗	✗

AI fluence

QU'EST-CE QUE C'EST ?

- Projet lancé en partenariat avec la SNCF & ENS Paris Saclay suite au gain du challenge « AI for Industry » de la région Ile-de-France
- L'objectif est de créer un service de prédictions de l'affluence dans les trains à destination des voyageurs et des acteurs du transport

CHALLENGES CLÉS

- **Sources de données hétérogènes** aux qualités variables (anomalies, données manquantes, …)
- Développer **une solution unifiée** pour des périmètres de données disponibles différents (gares, trains, voitures, + 5 min, +5 jours, etc.)
- Décrire des **comportements complexes et exceptionnelles** (ex. évènements de type JO 2024)
- **Mesurer l'incertitude** pour des séries temporelles

Prédire des séries temporelles complexes, construire des assets pour corriger des données et automatiser les tests de modèles

NOS REALISATIONS

- Modèles de prédiction de l'affluence performants sur l'ensemble du réseau francilien, indépendamment des sources de données disponibles
- Benchmark SOTA des **méthodes de détection d'anomalie** et d'imputation de données manquantes
- **Méthodes de forecasts complexes** applicables à d'autres opérateurs de transports ou d'autres contextes (ex. Smart Grids, détection de panne, …)
- Enrichissement de MAPIE, package Open Source, pour mesurer l'incertitude des séries temporelles

Quantmetry
Part of Capgemini Invent

Région îledeFrance

SNCF