Improving data for the asset management of the water supply network of the Walloon Water Company

Nicolas Rodriguez, Aurélien Mirebeau, Alain Husson, Marie Collet, Eddy Renaud, Yves Le Gat

Improving data for the asset management of the water supply network of the Walloon Water Company (SWDE)

GePaME project between SWDE and INRAE Bordeaux – Task 01

with contributions from A. Mirebeau, A. Husson, M. Collet, E. Renaud, and Y. Le Gat

Nicolas RODRIGUEZ
Service Statistiques et Contrôle des Données
Processus Transformation et Data

Presentation contents

SWDE in a few words

1

Walloon Water Company (SWDE)

Major water utility in Wallonia (Belgium) :

- 67 % of connection points, > 1 million user water meters, > 2.5 million users supplied
- 190/262 cities in Wallonia supplied
- Nearly 40 000 km of pipes
- Over 162 million m^3 of water introduced every year into the pipe network
- Over 150 M€ invested every year for infrastructure maintenance

The GePaME project between SWDE and INRAE-ETTIS

2

GePaME Project

3-year (2020-2023) applied research project between INRAE (ETTIS research unit) and SWDE

INRAE: French public research institute (Agriculture, Food & Water Supply, Environment)

ETTIS: Research unit among which engineers and researchers work on the **Asset Management of Water Infrastructures**

GePaME : multi-scale asset management of drinking water networks

Aim: help SWDE improve its asset management from the pipe to the whole network scale, from the short-term to the long-term

INRAE

ETTIS
Environnement Territoires en Transition Infrastructures Sociétés

La société wallonne des eaux

Task 01: dealing with data

3

Dealing with data – More than 50% of project time

Why do we need to manipulate the data?

Data has been (historically) gathered and structured to answer daily business needs

Asset Management requires large records of data specifically formatted (e.g., models, statistical analyses)

What are the 3 main problems to deal with?

1. Detecting the underlying DB structure in the various files (GIS, CMMS, CSV)
2. Detecting and correcting data which are not consistent
3. Imputing missing data

Tools used and why : R, Excel, QGIS

Unlike analysts employed at SWDE, we did not have access to online (up to date) databases

Statistical analyses and models are easily made with R code

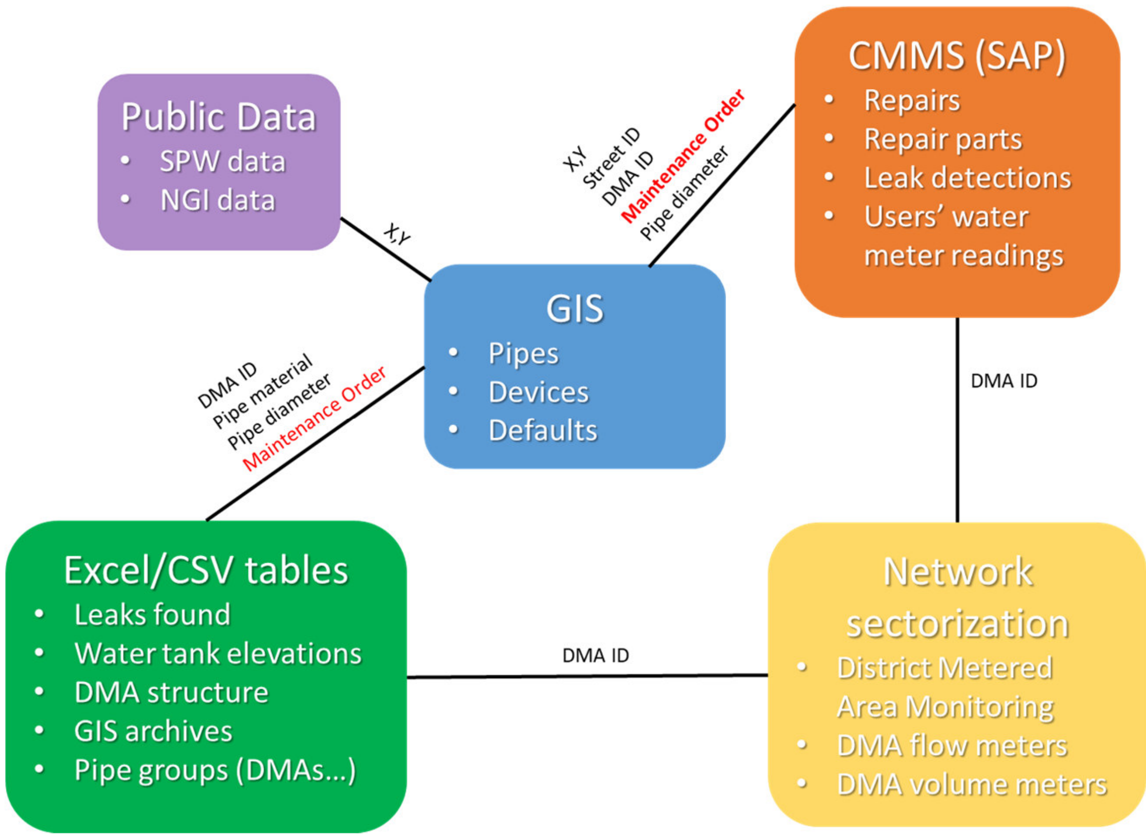
How much data are we talking about?

- 500k GIS pipe segments
- 2 500 DMAs with 1 to >5 water meters, 1 measurement/15 min since ~2016
- 100 000 repairs on the pipe network since 2011
- 90 000 leak inspections since 2016
- Over 1 million user water meters with yearly readings
- Millions of address points for users
- Thousands of streets represented as MultiLineStrings
- …

Problem n°1: identifying the Database structure

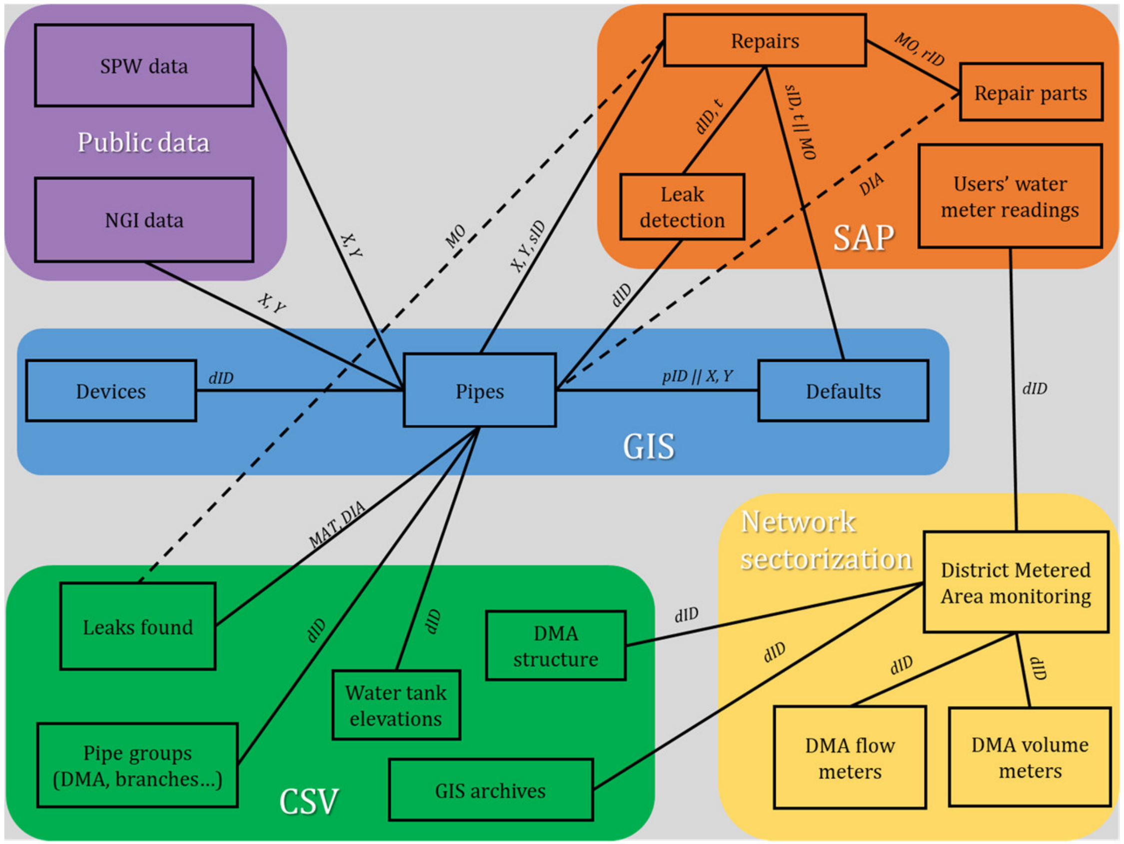
What does the data used in the project look like?

- Not all data from SWDE is used
- 3 main sources (SAP, Elyx Aqua, Perf'O)
- Many separate files
- Keys linking the files sometimes complex



CMMS: Computerized Maintenance Management System
GIS: geographic information system
CSV: comma separated values
SPW (Service Public de Wallonie)
NGI (National Geographic Institute)

What does the data used in the project look like? (internally)



SAP: enterprise resource planning software
GIS: geographic information system
CSV: comma separated values
SPW (Service Public de Wallonie)
NGI (National Geographic Institute)

X, Y: geographic coordinates
MO: Maintenance Order (SAP concept)
ID: (d=DMA, p=pipe, r=repair part, s=street)
t: time
DIA: pipe diameter
MAT: pipe material

Methods used to identify the DB structure

Task	Functions used
Finding the identifiers Removing duplicates	unique() ; duplicated() ; is.na()
Creating new identifiers	setkey() ; DT[,id:=…,by=.(name,date)]
Matching tables using identifiers	merge(x,y,all.x=T/F,all.y=T/F,by=…)
Matching tables using coordinates	st_nearest_feature(x,y)

Challenge: the network changes everyday, identifiers change…

Advantages: automate and make everything reproducible
ex: interpolate water consumption daily over 6 years for > 1 million water meters -> few min (*rolling join*).
Drawbacks: code is not very visual, code syntax is package-specific

Take-home messages:

- **Archive regularly (more than yearly) an image of the whole network or a history of changes**
- **Make sure the link between pipes in service and out of service is kept**
- **Record and store the history of changes in the DMA structure**

Problem n°2: detecting and correcting data errors

5

Typical errors to correct: text data

Manually input text can be inconsistent with the “constraints” (data type, null values forbidden, list of authorized values, upper/lower cases, local characters…) **-> ideally, avoid manual text!**

Task	Solution	Function used	Raw data	Output data	Matched with
• Extract street names, material names, diameters… • Regroup similar values	Regular expressions	str_detect()	fg 50 fonte 1960 dn50 Rue du bidule 80b Avenue_machin face 52 a	FONTE GRISE 50 FONTE GRISE 50 rue, du bidule, 80b avenue, machin, 52a	
Associate similar names	String distance Fuzzy matching	stringdist() amatch() toupper() stri_trans_general(,"Latin-ASCII")	London, Baker st. London, Bqker str FLENU CELL 1 Flénu CEL 01		London, Baker street FLENU CEL 01

Problem n°3: imputing missing data

6

Data imputation: automatized

Imputation of “simple” pipe attributes that should already be available:

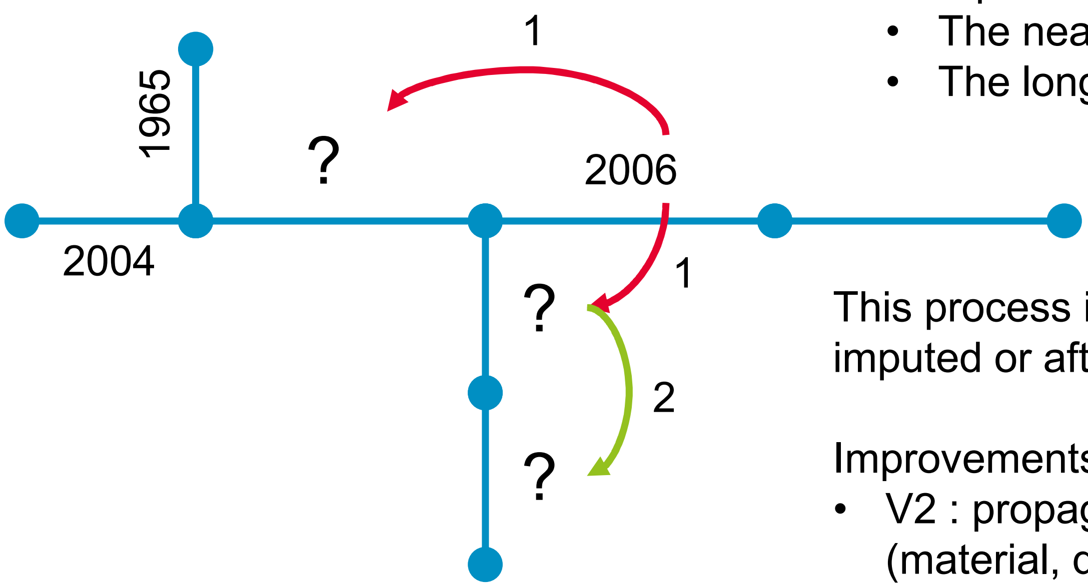
- Pipe materials, diameters: sometimes missing (few %), not always required
- **Pipe installation dates: often missing (>10-20%), almost always required (pipe age)**

Imputation of more complex data not necessarily available:

- Average static service pressure for the pipe (complex R algorithm)
- Number of connection points and number of users for the pipe (geomatics)
- Type of soil, position under the road, azimuth…

Husson, A. et al. (2022). Gestion Patrimoniale Multi-Echelles des réseaux d’eau potable (GePaME) - Rapport intermédiaire 2022. https://hal.inrae.fr/hal-03836891
Rodriguez, N., Husson, A., Le Gat, Y., Renaud, E. (2022). Automatized data preparation methods for the asset management of water supply networks. TSM10 (2022), 71-87.

How pipe installation dates are imputed (initial algorithm V1)



Assumptions:
- Pipes are installed by batches, usually by street
- The nearest known date is the most likely
- The longest neighboring pipe has the most reliable data

This process is repeated iteratively until all missing data is imputed or after a maximum number of steps imposed.

Improvements:
- V2 : propagate information only if N other attributes (material, diameter…) match between neighbors
- V3: similar to V1 but « holes » do not block propagation

Benchmark: comparison with a « naive » method V0 replacing unknowns by median values

Comparing imputation completeness and accuracy (and precision)

Completeness: how much missing data is imputed? (pipe length imputed / unknown)

Accuracy: is the imputed data close to the truth? (pipe length correctly imputed / unknown)

Precision: how much do results change when repeated many times?

Pipe ID	Original data						
a	1950						
b	2000						
c	2010						

Comparing imputation completeness and accuracy (and precision)

Completeness: how much missing data is imputed? (pipe length imputed / unknown)

Accuracy: is the imputed data close to the truth? (pipe length correctly imputed / unknown)

Precision: how much do results change when repeated many times?

Pipe ID	Original data	« Missing » data #1					
a	1950	?					
b	2000	2000					
c	2010	2010					

Comparing imputation completeness and accuracy (and precision)

Completeness: how much missing data is imputed? (pipe length imputed / unknown)

Accuracy: is the imputed data close to the truth? (pipe length correctly imputed / unknown)

Precision: how much do results change when repeated many times?

Pipe ID	Original data	« Missing » data #1	« Missing » data #2				
a	1950	?	1950				
b	2000	2000	?				
c	2010	2010	2010				

Comparing imputation completeness and accuracy (and precision)

Completeness: how much missing data is imputed? (pipe length imputed / unknown)

Accuracy: is the imputed data close to the truth? (pipe length correctly imputed / unknown)

Precision: how much do results change when repeated many times?

Pipe ID	Original data	« Missing » data #1	« Missing » data #2	Imputed by V0 #1		Imputed by V1 #1	
a	1950	?	1950	1950		1950	
b	2000	2000	?	-		-	
c	2010	2010	2010	-		-	

Comparing imputation completeness and accuracy (and precision)

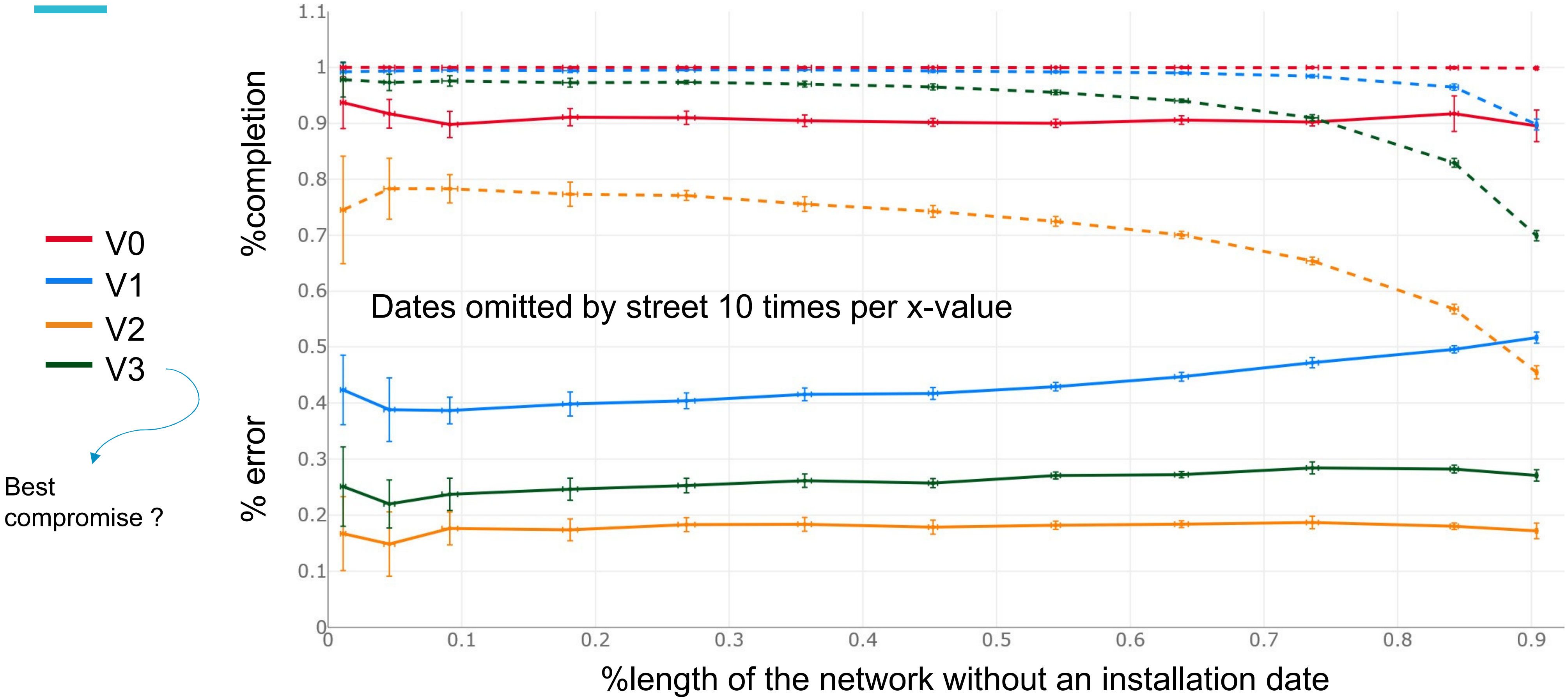
Completeness: how much missing data is imputed? (pipe length imputed / unknown)

Accuracy: is the imputed data close to the truth? (pipe length correctly imputed / unknown)

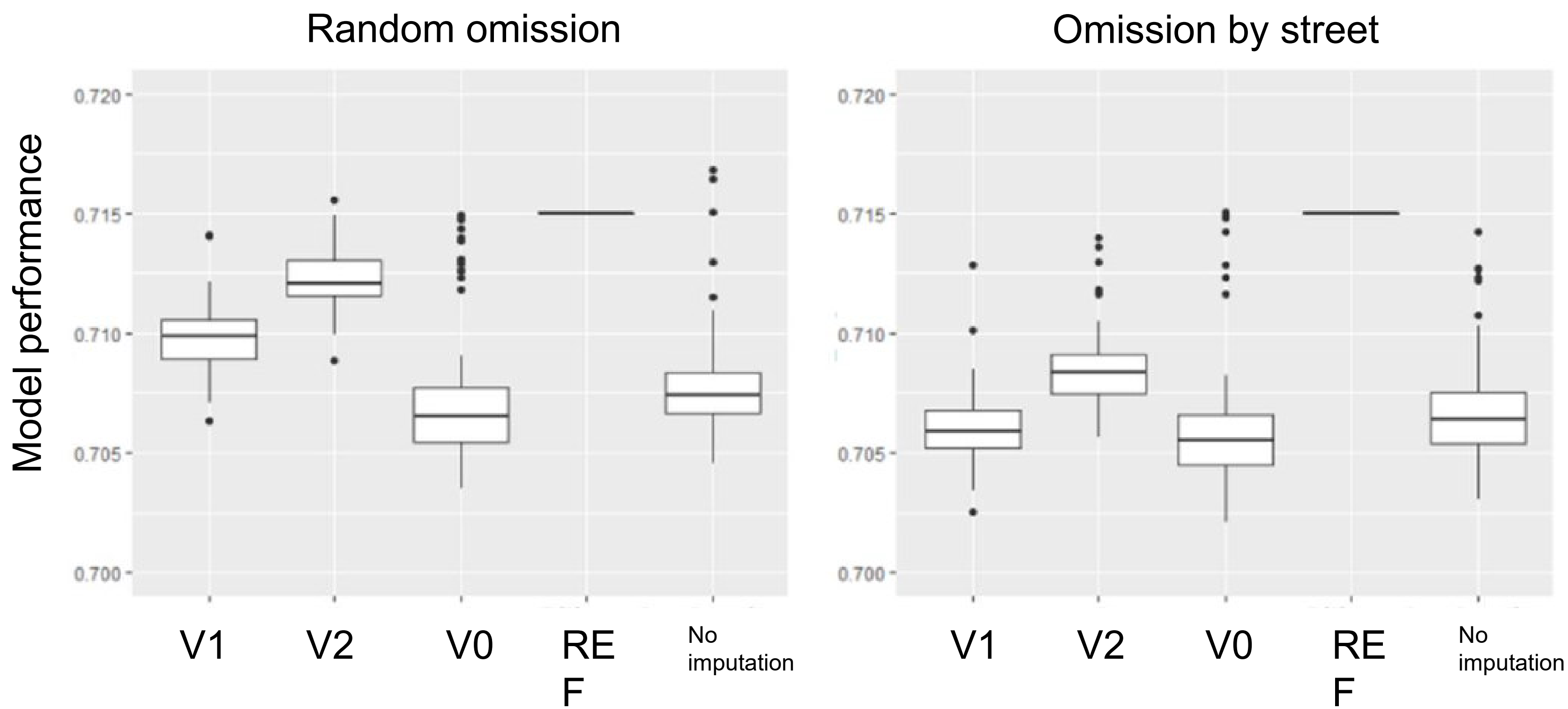
Precision: how much do results change when repeated many times?

Pipe ID	Original data	« Missing » data #1	« Missing » data #2	Imputed by V0 #1	Imputed by V0 #2	Imputed by V1 #1	Imputed by V1 #2
a	1950	?	1950	1950	-	1950	-
b	2000	2000	?	-	2005	-	?
c	2010	2010	2010	-	-	-	-

Comparing imputation completeness and accuracy (and precision)



What is the impact on the prediction of failures?



40% materials omitted

100 repetitions for each imputation method

No data is sometimes better than bad data!

Husson, A. et al. (2022). Gestion Patrimoniale Multi-Echelles des réseaux d’eau potable (GePaME) - Rapport intermédiaire 2022. https://hal.inrae.fr/hal-03836891

Conclusions

These are the take-home messages:

01

ASSET MANAGEMENT

Requires large amounts of formatted, interoperable data.

02

DATA MANIPULATION

Best done with ETL tools with online databases but powerful “offline” alternatives exist (R and its packages).

03

DATA IMPUTATION

The most complex part of data manipulation. Requires intelligent design and rigorous testing.

04

RESULTS AND PERSPECTIVES

Many lessons learned for SWDE data management and improved ability to do asset management.

Thank you for your attention

SWDE
Esplanade René Magritte 20
6010 Charleroi, Belgium
nicolas.rodriguez@swde.be
0032 479 88 63 09

We thank all the members of the GePaME project from INRAE-ETTIS and from SWDE!