infraPLAN
Water Mains R&R & AI

Ask the Experts! #6

Incidental Data Issues -
Missing & Incoherent Values

Click below to read all ATE articles.
Article 1: 1pc Water Main R&R Rate
Article 2: LoF Scores
Article 3: LoF Case Studies
Article 4: Data Needed
Article 5: Abandoned Pipes
Article 6: Missing Data
Article 7: Structural Data Issues

Pipes and breaks with issues weaken break predictions. It is therefore important to limit their number.

The data needed to predict future breaks pertain to the pipes, their breaks, and their operational and environmental context, as described in Article #5. Data issues may bethe result of incidental human errors (missing or incoherent values) or faulty data recording and management processes (structural issues). **In this article we describe incidental pipe and break data issues**, how to identify and correct them. **Structural data issues** are addressed in Article #7.

Incidental data issues

Incidental data issues include missing (for example, no date of install, no material, no diameter), illogical (for example, date of break before date of install) and incoherent values (plastic dated in 1920). These issues tend to be isolated, resulting mainly from missing records, accidental human errors, or poor record keeping. Sometimes a utility acquires a neighboring system with problematic data.

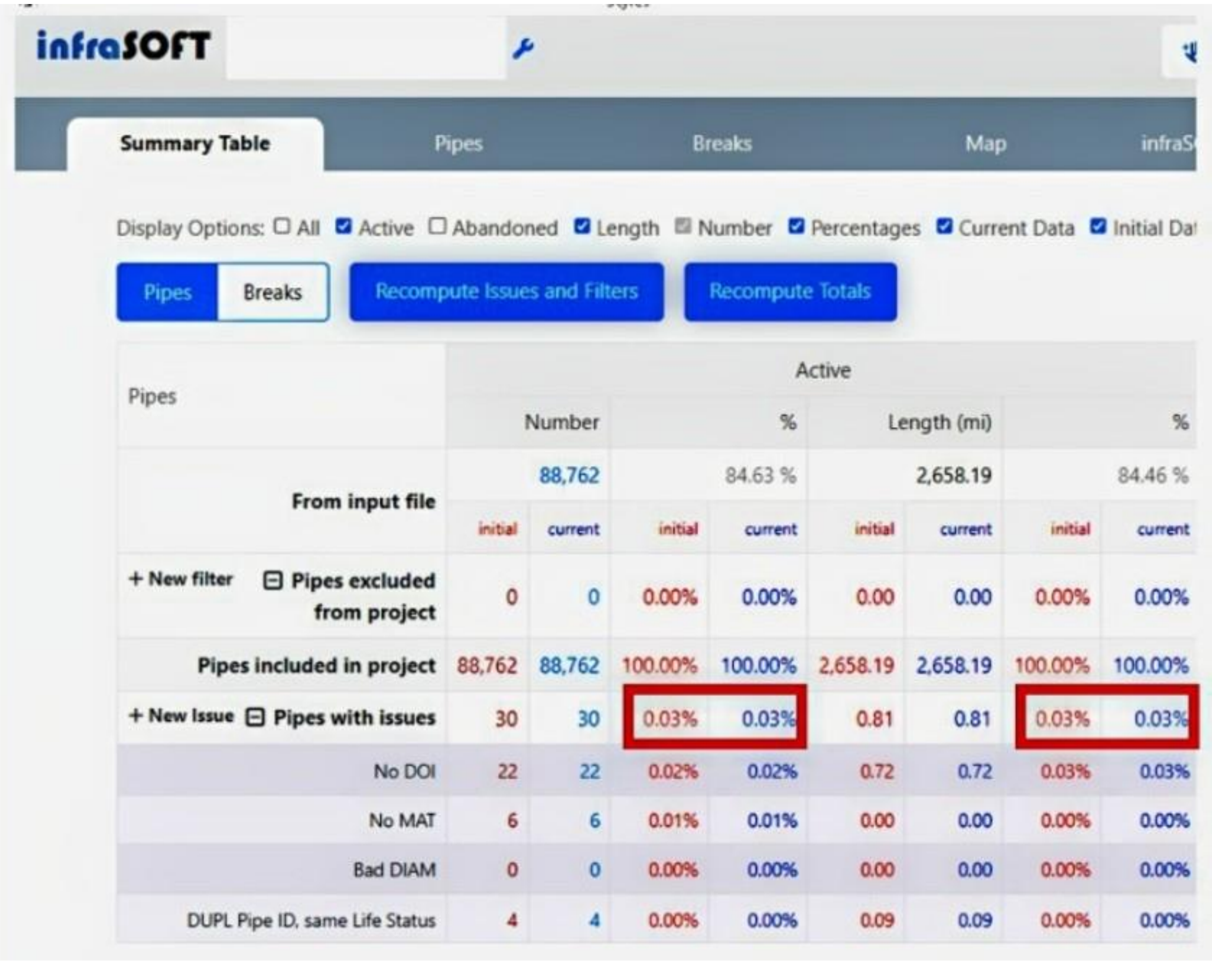
An incidental data issue is not the result of a faulty process; once corrected, the issue is gone.

How to detect data issues

Identifying missing or illogical values is straightforward. However, it can be labor-intensive and time-consuming. Furthermore, it requires a knowledgeable reviewer. Thegood news is that once an incidental data issue is corrected, the issue is gone once for all. The **CLEAN** module of **infraSOFT**, **infraPLAN** water pipes R&R management platform, repertories all the issues **infraPLAN** ever encountered in the hundreds of data sets we have analyzed over time.

> ***The CLEAN module of infraSOFT instantaneously identifies all possible pipe and break issues upon loading the input files.***

The screenshot below shows possible results in the **CLEAN** module of **infraSOFT**, at the system level; 30 pipes (0.03% of the 88,762 pipes), or 0.81 mile (0.03% of the total length of 2,658.2 miles) experience issues that include no Date of Install (DOI), No Material (MAT), Bad Diameter (DIAM), and Duplicate Pipes (DUPL). Many more built-in or utility-specific issues can be detected in **CLEAN**.

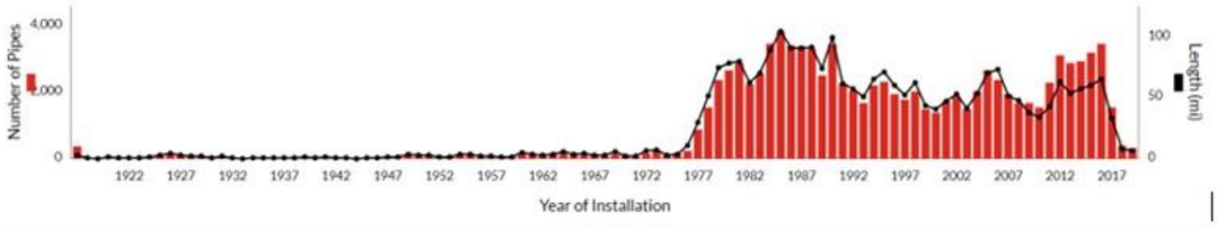


Display Options: ☐ All ☑ Active ☐ Abandoned ☑ Length ▨ Number ☑ Percentages ☑ Current Data ☑ Initial Dat

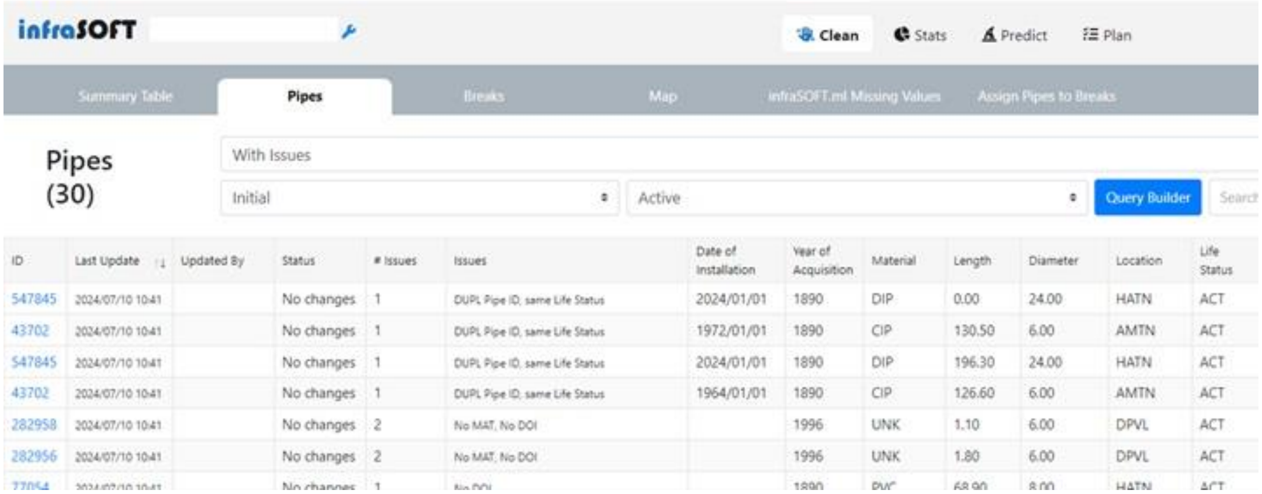
Pipes	Active							
	Number		%		Length (mi)		%	
From input file	88,762		84.63 %		2,658.19		84.46 %	
	initial	current	initial	current	initial	current	initial	current
+ New filter ⊟ Pipes excluded from project	0	0	0.00%	0.00%	0.00	0.00	0.00%	0.00%
Pipes included in project	88,762	88,762	100.00%	100.00%	2,658.19	2,658.19	100.00%	100.00%
+ New Issue ⊟ Pipes with issues	30	30	0.03%	0.03%	0.81	0.81	0.03%	0.03%
No DOI	22	22	0.02%	0.02%	0.72	0.72	0.03%	0.03%
No MAT	6	6	0.01%	0.01%	0.00	0.00	0.00%	0.00%
Bad DIAM	0	0	0.00%	0.00%	0.00	0.00	0.00%	0.00%
DUPL Pipe ID, same Life Status	4	4	0.00%	0.00%	0.09	0.09	0.00%	0.00%

The progress made on data cleaning can be monitored by comparing the values in the ”initial” column (red - refers to the data that was initially loaded) and the “current” column (blue - registers the data after cleaning - see next section).

Analyzing data trends (in the **STATS** module of **infraSOFT**) helps identify utility-specific outliers. For example, the figure below shows the number and length of pipes based on the year of installation (YOI). It shows that most Ductile Iron (DI) pipes wereinstalled after 1976. The utility was therefore able to create an issue in **CLEAN** that flags DI pipes with recorded year of installation prior to that year (some as early as1915 which we know cannot be accurate).



The screenshot below shows how issues are displayed at the pipe level in**CLEAN**.



Pipes (30)

With Issues

Initial

Active

ID	Last Update	Updated By	Status	# Issues	Issues	Date of Installation	Year of Acquisition	Material	Length	Diameter	Location	Life Status
547845	2024/07/10 10:41		No changes	1	DUPL Pipe ID, same Life Status	2024/01/01	1890	DIP	0.00	24.00	HATN	ACT
43702	2024/07/10 10:41		No changes	1	DUPL Pipe ID, same Life Status	1972/01/01	1890	CIP	130.50	6.00	AMTN	ACT
547845	2024/07/10 10:41		No changes	1	DUPL Pipe ID, same Life Status	2024/01/01	1890	DIP	196.30	24.00	HATN	ACT
43702	2024/07/10 10:41		No changes	1	DUPL Pipe ID, same Life Status	1964/01/01	1890	CIP	126.60	6.00	AMTN	ACT
282958	2024/07/10 10:41		No changes	2	No MAT, No DOI		1996	UNK	1.10	6.00	DPVL	ACT
282956	2024/07/10 10:41		No changes	2	No MAT, No DOI		1996	UNK	1.80	6.00	DPVL	ACT
77054	2024/07/10 10:41		No changes	1	No DOI		1890	PVC	68.90	8.00	HATN	ACT

For example, ID 43702 (2nd and 4th row) was given to two pipes which is a problem. The material and the date of installation of pipe ID 292956 (5th row) are unknown.

Similar results are provided for breaks, at the system or break level.

Issues can then be cleaned by the utility at the source, or **infraPLAN** proposes variousways data can be cleaned within the tool, including machine learning-powered.

How to repair missing and incoherent data Issues - The machine learning CLEAN module of infraSOFT

While incidental issues can be corrected by looking at the map (embedded in **infraSOFT**), or outside of **infraSOFT**, by consulting source documents if available, only one issue at a time can be addressed using those manual approaches; it is time-consuming especially for systems that have a high percentage of data issues.

To save time and resources, automated remedies can be activated within **CLEAN.** This is done after a full analysis of the data and with the approval of the utility. Remedies allow:

- Removing all duplicates, pipes or breaks
- Assigning the break to the most likely pipe (when duplicates are all valid and cannot be removed)
- Setting a statistical rule. For example, “pipes located in a certain district registered as DI with YOI prior to 1976 that have been abandoned should be changed to Cast Iron (CI)”.
- Finally, using **ML CLEAN**, **infraSOFT** machine learning-powered cleaning module, to fill in missing values

The **ML CLEAN** module was used for a recent project with 2% of the length of pipe missing a material; 4.5% a year of installation; and 0.02% a diameter. Model validationrelies on selecting 20% of the pipes that were not missing any value, deliberately and randomly removing some of those values, then estimating what they actually were, using **ML CLEAN**. Results were as follows:

- **The material was properly predicted for 98.3%**of the sample.
- The average difference (absolute value) between actual and predicted **years of installation was 1.08 year.**
- The average difference (absolute value) between actual and predicted **diameters was 0.59 inches.**

It is recommended that the percentage of issues be brought below 10% before undertaking break predictions. This can be easily and quickly achieved with infraSOFT ML CLEAN.

Click below to read all ATE articles.
Article 1: 1pc Water Main R&R Rate
Article 2: LoF Scores
Article 3: LoF Case Studies
Article 4: Data Needed
Article 5: Abandoned Pipes
Article 6: Missing Data
Article 7: Structural Data Issues

Contact us for a free discussion on using advanced analytics to maximize your R&R plan!

infraPLAN-LLC helps water utilities, large and small, achieve savings on CIP expenses using our ground breaking platform, **infraSOFT.**

www.infraPLAN-llc.com
(917) 349-6386

Annie Vanrenterghem, PhD, CEO

infraPLAN | 5 Union Square West #1139 | New York, NY 10003 US