infraPLAN
Water Mains R&R & AI

Ask the Experts! #7
Structural Data Issues

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

The promise of machine learning is that break predictions will get better as more data become available. However, if structural flaws in data processes exist, over time, break predictions may actually become worse.

The data needed to predict future breaks pertain to the pipes, their breaks, and their operational and environmental context; they are described in Article #4. Breaks assigned to the wrong pipes lead to erroneous Likelihood of Failure (LoF) scores, which could result in costly unnecessary replacements or breaks. It is therefore critical to limit the number of issues.

In this article we describe pipe and break data resulting from flawed data recording processes, how to identify and correct them. Less punishing incidental data issues are addressed in Article #6.

Structural data issues

Once an incidental data issue is fixed, the missing value is filled in, or an incoherent value is corrected, the correction is permanent. It is not the case for an issue that is the result of a structural flaw: if not modified at the source, that type of issues will keep appearing after each labor-intensive and time-consuming data clean up. This means that, as more data become available, machine learning will not provide increasingly better break predictions. On the contrary, predictions will worsen.

There are several major data management processes that lead to that type of issue:

- Failure to record (or properly record) abandoned pipes (to understand why abandoned pipes matter see Article #5)
- Incorrectly linking breaks and pipes

- Recycling pipe IDs after full or partial replacement
- Failure to report in the CMMS, changes of pipe ID occurring in GIS

The consequences may be as follow:

- Breaks assigned to the wrong pipe: that pipe (and similar pipes) end up having a higher LoF score than they should have; and, as a result, may be prematurely replaced.
- Breaks not assigned/lost to the pipe that had broken: that pipe (and similar pipes) end up having a lower LoF score than they should have; and may break again before they are replaced.

Case study

We track the history of pipe 101 (100-ft long with a diameter of 8") starting in 1910 when it was installed.
Active pipe = ACT
Abandoned pipe/no longer in service = ABN
Ductile Iron = DI
Cast Iron = CI

Pipe ID	Material	Year Install	Diameter	Length	Life Status	Year Abandon
?	CI	1910	8"	40'	ABN*	2000
?	DI	2000	8"	40'	ACT*	
?	CI	1910	8"	60'	ACT*	

What should the pipes IDs be?

Options:
(1) if DI is given the recycled ID 101, the breaks would be wrongly assigned to DI and found to be incoherent because they occurred between 1980 and 1995, prior to installation in 2000. **Those breaks would be discarded from the analysis**. **CIABN would be wrongly assigned no break.**
(2) If CIACT keeps the old ID 101, CIABN is given a different ID or CIABN is not recorded, the breaks would be wrongly assigned to CIACT. The LoF analysis would **wrongly attribute a very high score** to that pipe (that actually never broke!). As a result, a potentially good pipe may end up being replaced.
(3) In some cases when ID recycling is permitted, two (2) pipes may be given the ID 101, CIABN keeping its old ID, and CIACT or DI. Without additional data treatment, **the breaks would end up being discarded** as a break cannot have occurred on 2 pipes.

Best practices:

- IDs are not recycled.
- CIABN conserves its original ID (101) which is in agreement with the assignment of B1, B2, and B3 (to 101).
- CIACT and DI are given a different ID, 103, and 102, respectively.

How to detect data issues

In Article #6 we described how the **CLEAN** module of **infraSOFT**, infraPLAN water pipes R&R planning platform, instantaneously identifies all possible pipe and break issues

upon loading the input files. Some of the issues automatically identified actually point to data structural flaws:

- Issue is "No pipe": This issue appears if the ID of the pipe associated to a break is not included in the pipes database. This likely indicates that the break occurred on an abandoned pipe that was not recorded.
- Issue is "DUPL pipe": Appears if a pipe shares an ID with at least one other pipe. If a break is associated to that pipe ID, assigning it to the right pipe may require additional tests, or not be possible at all.
- Issue is "DOB<DOI": Appears if the date of the break (DOB) is before the date of installation (DOI). This most likely indicates that the break did not occur on that pipe but, if the pipe is a replacement pipe, possibly on the original pipe.

This is the extent to which this part of the analysis can be automated. At this point, a careful review of the processes leading to recording breaks, and managing pipe IDs in GIS and CMMS must take place before using advanced analytics and machine learning to assign LoF scores to water pipes. Otherwise, not only could some results be erroneous, but they will worsen over time as more issues will appear.

How to repair structural data issues

Once structural data issues are identified, actions to be taken are twofold:

- Correct the past issues resulting from the data structural flaw, and
- Correct the data structural flaw to ensure that no such data issues appear in the future.

Examples of data structural flaws correction:

- If the utility is not currently recording abandoned pipes, the decision may be made to retrieve a few years of past abandonment as well as to set up the process to record future abandonment. While best practices exist to both retrieve old abandonments, and also record future abandonments, the recommended approaches need to be adapted to the specific capacities and limitations of the data management system in place. For example, some systems can not control the ID given to a pipe. In such case, the break should be reassigned manually after a replacement.
- For utilities that do record abandonment but recycle IDs, and therefore have a substantial number of pipes that share an ID with at least one other pipe (with a different life status; at least one ACT pipe, and one ABN pipe), we developed an algorithm within **infraSOFT, the DUPL algorithm,** that assigns any break associated to a duplicate pipe ID to the most likely pipe. The logic is selected together with the utility so that it reflects their practices. Note that this remains a best estimate that strikes a reasonable balance between level of effort and data certainty. We also recommend that the practice of recycling IDs be ended if the system allows.

Back to the case study to illustrate options within the **DUPL algorithm**.

Dealing with past issues: We saw previously that breaks B1, B2, and B3 are associated to pipe 101 in the CMMS. If both CIABN and CIACT are given the ID 101 (see table below), **infraSOFT DUPL algorithm** will assign B1, B2, and B3 following one of the optional logics described below. The option is selected by the utility.

Pipe ID	Material	Year Install	Diameter	Length	Life Status	Year Abandon
101	CI	1910	8”	40’	ABN*	2000
102	DI	2000	8”	40’	ACT*	
101	CI	1910	8”	60’	ACT*	

Options include:

- All breaks are assigned to the ABN pipe (CIABN) when the utility finds it more likely that the pipe that experienced the breaks is the one that was then abandoned.
- All breaks are assigned to the longest pipe (CIACT). This option is retained by a utility that is aware small segments of pipes with no break may be opportunistically replaced because they were located in the trench of an ABN pipe.
- The breaks are assigned proportionally to the length. In our example, it would be one break to CIABN and two breaks to CIACT.

Avoiding future issues

We reiterate best practices :

- CIABN conserves its original ID (101) which is in agreement with the assignment of B1, B2, and B3 (to 101).
- CIACT and DI are given a different ID, 103, and 102, respectively.

This would guarantee that the breaks be assigned to the right pipe, a key factor for successful break predictions.

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 maximixe 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