infraPLAN
Water Mains R&R & AI

Ask the Experts! #4
Data Needed

Data needed to predict water main breaks using machine learning

In previous articles, we showed how machine learning provides better water pipe break predictions (Article #3), and is easier to deploy (Article #2).

In the upcoming series of articles, we discuss data, and more specifically:
Article #4: Data needed
Article #5: Abandoned pipes
Article #6: Data issues - missing and incoherent values
Article #7: Data issues - structural flaws

Which data?

The data identifying the pipe or expressing the factors that affect its degradationshould be collected including:

Pipes physical or operational characteristics:
*factors are "nice to have"; **factors are indispensable.

- Pipe ID**
- Pipe type*
- Diameter*
- Material**
- Date (or year) of installation**

- Length**
- Life status (active or abandoned) ** - see Article #5 "Why abandoned pipesshould be included in the break prediction model."
- Date or (year) of Abandonment**
- Location* (district, community board, neighborhood)
- Date of acquisition* (if new systems have been acquired over time)
- Pressure*

Environmental factors

- Soil
- Rail
- Traffic
- Groundwater
- Construction
- Wetland
- Water
- Structure
- Seismic

The above data should be available for all the pipes, whether they broke or not in the past**.**

Breaks

- Break ID*
- Date (or year) of break**
- Pipe ID of the pipe the break occurred on**
- Break type**

Not all pipes and breaks will be included in the study. For example, the following may need to be filtered out:

- Pipes not targeted for R&R based on, for example, location, material, type, or diameter.
- Breaks unrelated to pipe degradation. For example, leaks on joints or appurtenances; third party breaks.

How much?

- At least 100 miles of pipes
- At least 5 years of significant and consistent break data corresponding to a minimum of 0.02 breaks/mi/yr
- Smaller systems or systems with little or poor data can eventually benefit from data from other systems.

- If the value pertaining to an indispensable variable (referred as** in the “Which data?” section) is missing or suspicious, the pipe or break will be eliminated fromthe study. That pipe will have no LoF score.
- A model always benefits from more and better data. If a *factor is not yet available, the utility may want to start collecting them, but not before a Cost/Benefit analysis of collecting additional data has been undertaken. Having run hundreds of models, we understand which data generate more analytical value and therefore should be collected or cleaned in priority. We will not encourage the utility to collect new data at great cost, when the benefit is limited.
- Pipes with missing or suspicious values for a non- indispensable or environmental factor, are not eliminated from the study; an average value is assigned.
- Last, some factors such as traffic, construction density, pressure or the definition of a break, may have changed over time. It is important to select a period of break observation during which the factors that matter were relatively stable, and reflect the current conditions.

How good?

Pipes or breaks with data issues cannot be included in the study, which weakens the results. Articles #6 and #7 describe the issues pipe and break data may experience, and how to clean them. They include isolated issues due to human errors (missing or incoherent values), or structural issues resulting from flawed data collection and management.

It is recommended that, at the onset of a project, all issues be identified as a percentage of the number/length of pipes, and of the number of breaks. Then, when it comes to data, a project is typically organized in 3 phases:

- Phase 1: bring percentage of pipe and break data with issues below 10%. Then reliable break predictions can be generated.
- Phase 2: provide a road map to further lower the percentage of isolated pipe andbreak issues. See Article #6.
- Phase 3: provide a road map to modify the processes leading to structural data issues. See Article #7.

What source?

- The physical characteristics of a pipe are typically available in GIS.
- Operational data may come from the hydraulic model.
- Environmental factors can be pulled from publicly available GIS layers maintained by local, state or federal agencies (DOT, USGS, etc.).

- Breaks are often extracted from the CMMS work orders.

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