infraPLAN Ask the Experts! #5_1

Machine Learning to Predict Water Pipe Breaks
The Ultimate Data Series

In previous articles, we showed how machine learning provides better water pipe break predictions, and how it is also easier to deploy[1][2].
In this series of articles, we discuss data, and more specifically:
Part 1: Data needed to predict water main breaks using machine learning
Part 2: Why abandoned pipes should be included in a break prediction model
Parts 3-4: Problems that may be encountered with data and how to fix them

(1) Article 1
(2) Article 2

Part 1: Data needed to predict water pipe breaks using machine learning

Which data?

The factors that affect the degradation of a pipe should be collected. They include:

- **Pipes physical or operational characteristics**
  - Pipe ID**
  - Pipe type*
  - Diameter**
  - Material**
  - Date (or year) of installation**
  - Length**
  - Life status (active or abandoned)** - see Part 2 of this Data Series: "Why abandoned pipes should 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*

** indispensable; *Nice to have

- **Environmental factors***
    - Soil
    - Traffic
    - Groundwater
    - Construction density
    - Zoning
    - Other

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

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

Not all pipes and breaks should 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.
- Work Orders that were not issued because of pipe degradation. For example, leaks on joints or appurtenances; third party breaks.

How much?

- At least 100 miles of pipe, and 5 years of significant and consistent break data.
- Smaller systems can benefit from data from other systems.
- Some factors are indispensable (referred as **** in the "Which data?" section); If not yet available, the utility may want to start collecting the *factors
- A model always benefits from more and better data. However, having run hundreds of models, we understand which data generates more analytical value and therefore, should be collected or cleaned in priority.
- Incomplete data may in some cases be acceptable.
- Last, some factors such as traffic, construction density, pressure or even 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. Parts 3 and 4 of this "Data Series" 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 issues below 10%. Then reliable break predictions can be generated.
Phase 2: provide a road map to further lower the percentage of isolated pipe and break issues.
Phase 3: provide a road map to modify the processes leading to structural issues.

For Phases 2 and 3, see Parts 3-4 of this Data Series: "Problems that may be encountered with data and how to fix them".

What Source?

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

Over the next 3 weeks we will address the following:
Part 2: Why abandoned pipes should be included in the model?
Parts 3-4: Problems that may be encountered with data and how to fix them

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