A Random Forest algorithm

applied to condition-based wastewater deterioration modeling and forecasting

Diogo Vitorino (Addition)
Sérgio T. Coelho (LNEC)
Steven Sheets (GCDWR)
Brad Jurkovac (GHD)
Conceição Amado and Pedro Santos (IST-UL)

WDSA
Bari, Italy, 14-17 July 2014

Presentation outline

1. Introduction
2. Problem outline
3. Random Forest: in short
4. Inputs
5. Results
6. Prioritization performance
7. Mutual information
8. Data driven approach
9. Applicability

1. Introduction

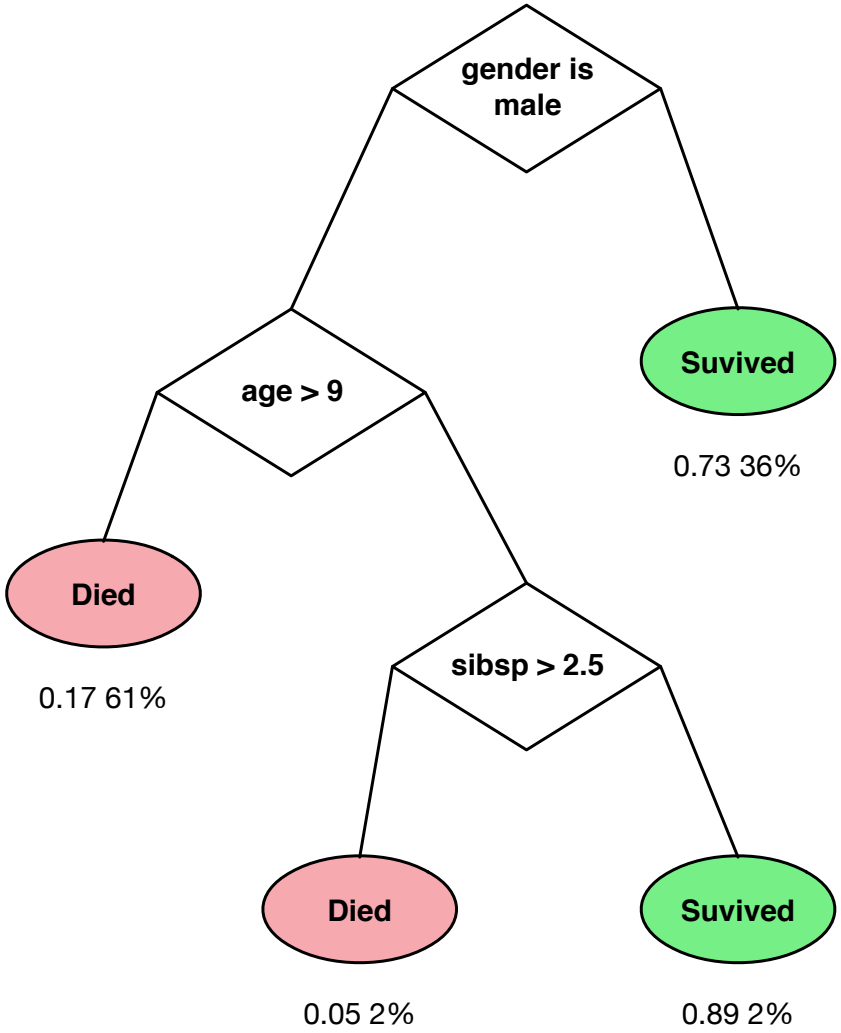
- Water Environment Research Foundation (WERF) and U.S. Environmental Protection Agency (USEPA) funded research
- Application of AWARE-P software in a US context
- Development of new tools for wastewater drainage systems
- R&D team includes GHD, LNEC, IST-UL and Addition
- Application in Gwinnett County Department of Water Resources (GA)

2. Problem outline

- Utilities make a remarkable investment in acquiring large volumes of high-quality data for their sanitary sewer system. Namely in:
  - Asset inventory (GIS or other)
  - CCTV inspections covering significant proportions of total network length
- Current practice focuses on direct evaluation of assets
- Ranking/prioritization of assets is often based on risk matrix crossing:
  - PoF ~ likelihood of failure (age/material-based decay curves), with
  - Consequence of failure, measured by compounding several variables

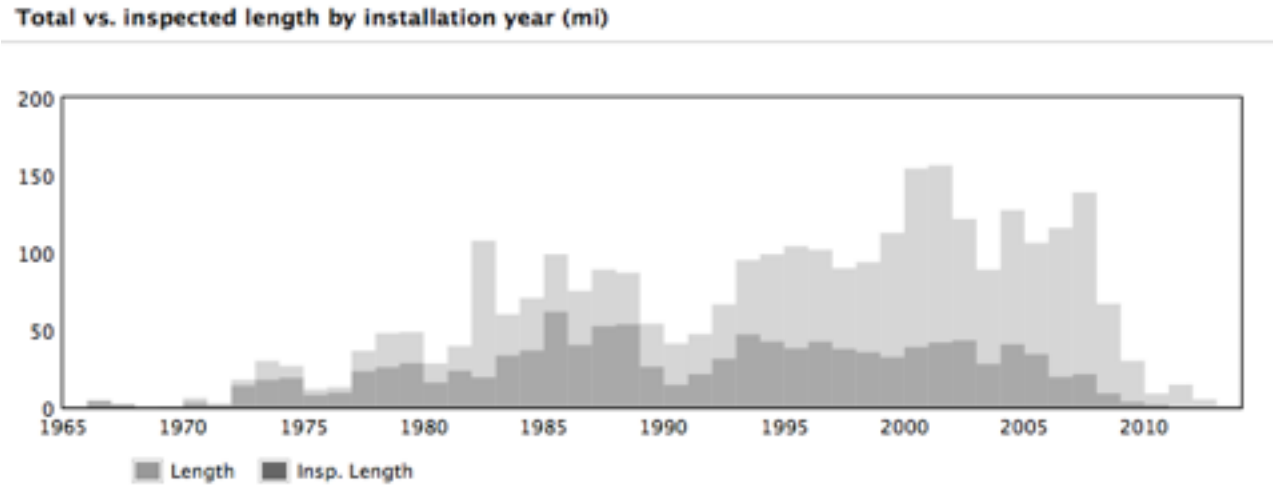
3. Random Forest: in short

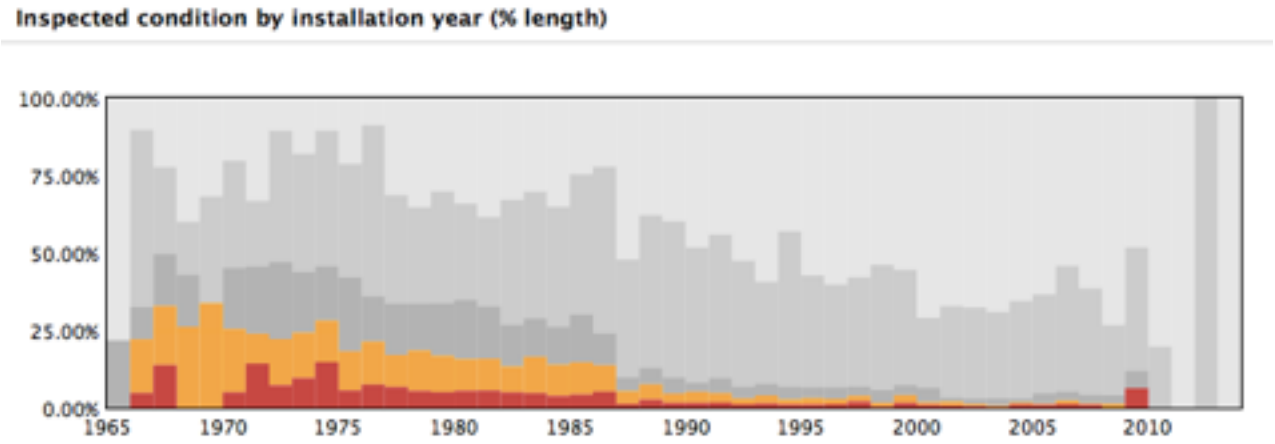
- A classification problem: what's the probability of pipe X being critical today? In 10 years' time?
- Decision trees work by looking at data and recursively splitting the given information by choosing the most relevant variable on each step.
- On each "leaf", an estimated probability is given.
- Method makes no assumptions and no inferences (e.g., a given probability distribution): results are only as good as the data.



4. Inputs

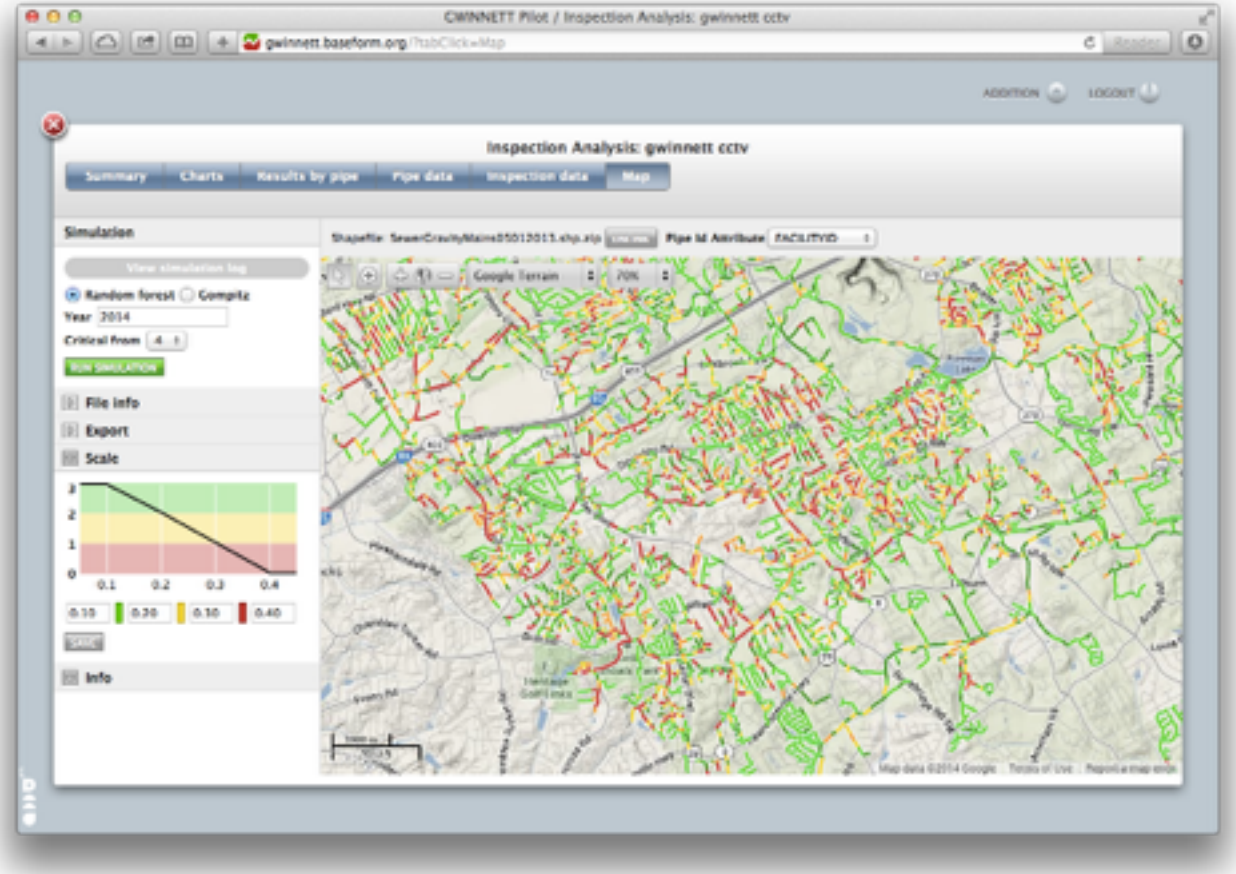
- 2 input files needed:
- Pipes
  id, material, length, diameter, date
- Inspections
  pipe id, date, condition
- Have a look at the data. Results are only as good as the data feeding them.





5. Results

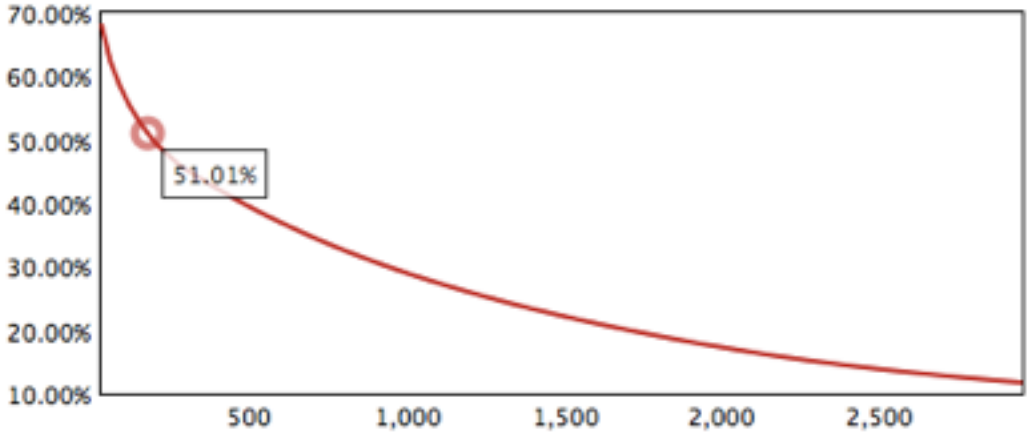
- Individual, pipe-by-pipe and zonal predictions are produced
- GIS-based maps allow for visualizing estimated criticals.



6. Prioritization performance

- A measure of the benefit of inspecting a given length of sewers by using the software is detailed on the generated chart.
- In the figure, non-directed inspections would find 11% critical pipes (the average) whereas using RF ranking to inspect the first 200mi would yield >50%.
- $/critical-finding is increased 5x.

Critical pipes by accumulated inspected length (est. probability)

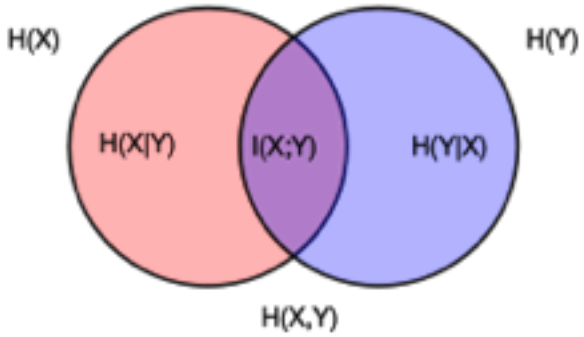


Aggregated estimated probability distribution of critical assets by length (mi)

Material	Est. Probability	Length (mi)
STEEL	9.50%	0.04
HDPE	13.42%	0.13
Not Known	27.94%	0.97
CIP	31.83%	3.05
RCP	23.89%	16.37
RPM	21.27%	63.08
VCP	22.97%	80.84
DIP	12.08%	106.97
PVC	6.14%	81.80
Total	11.98%	353.27

7. Mutual information

- The mutual information of two variables or factors is a measure of their mutual dependence
- Striving to find the explanation for pipe condition is key to being able to understand and predict its future condition, or that of its peers.
- The user can add potential explanatory variables and directly measure its contribution to pipe behavior
- Examples of possible explanatory variables: soil type, traffic, tree cover, land use, I&I, zoning, slope, etc.



Co-variable mutual information

Variable	Information gain
Zone	0.00%
Material	9.16%
Diameter	1.45%
Length	1.41%
Age	21.77%
Prev. Inspection	0.32%
Age Prev. Inspection	0.30%
Co-variable 1	1.06%
Co-variable 2	0.00%
Co-variable 3	0.00%

8. Data driven approach

- Mining big data, using novel statistical approaches, can bring forth reliable predictions on sewer condition.
- Traditional assumptions on service life can be validated or challenged by the utility's own existing data.
- A data-driven solution should make no assumptions: it will look at existing data and extract as much factual information as possible.
- Different types of information can be analyzed by measuring their contribution to explaining pipe condition over time.
- Moving on from asset-by-asset to zonal prioritization yields more reliable assessments and more actionable results.

9. Applicability

- IAM planning: Prioritize future actions, both to schedule inspections and to plan rehabilitation interventions
- Outcome can be used directly as a prioritization or ranking scheme, based on estimated probability of critical condition.
- Usage in short-, medium- and long-term prediction of condition:
- Prediction of future pipe condition degradation, based on direct, measurable, analysis of the utility's own data.
- Accurate estimation and planning for future levels of service and rehab needs.

Acknowledgments

The authors acknowledge all **team members, participants** and **sponsors** of the following projects:

- **AWARE-P**

  www.aware-p.org

- **INFR5R12**

  www.werf.org/a/k/Search/ResearchProfile.aspx?ReportID=INFR5R12

- **Baseform**

  www.baseform.org (software implementation details)