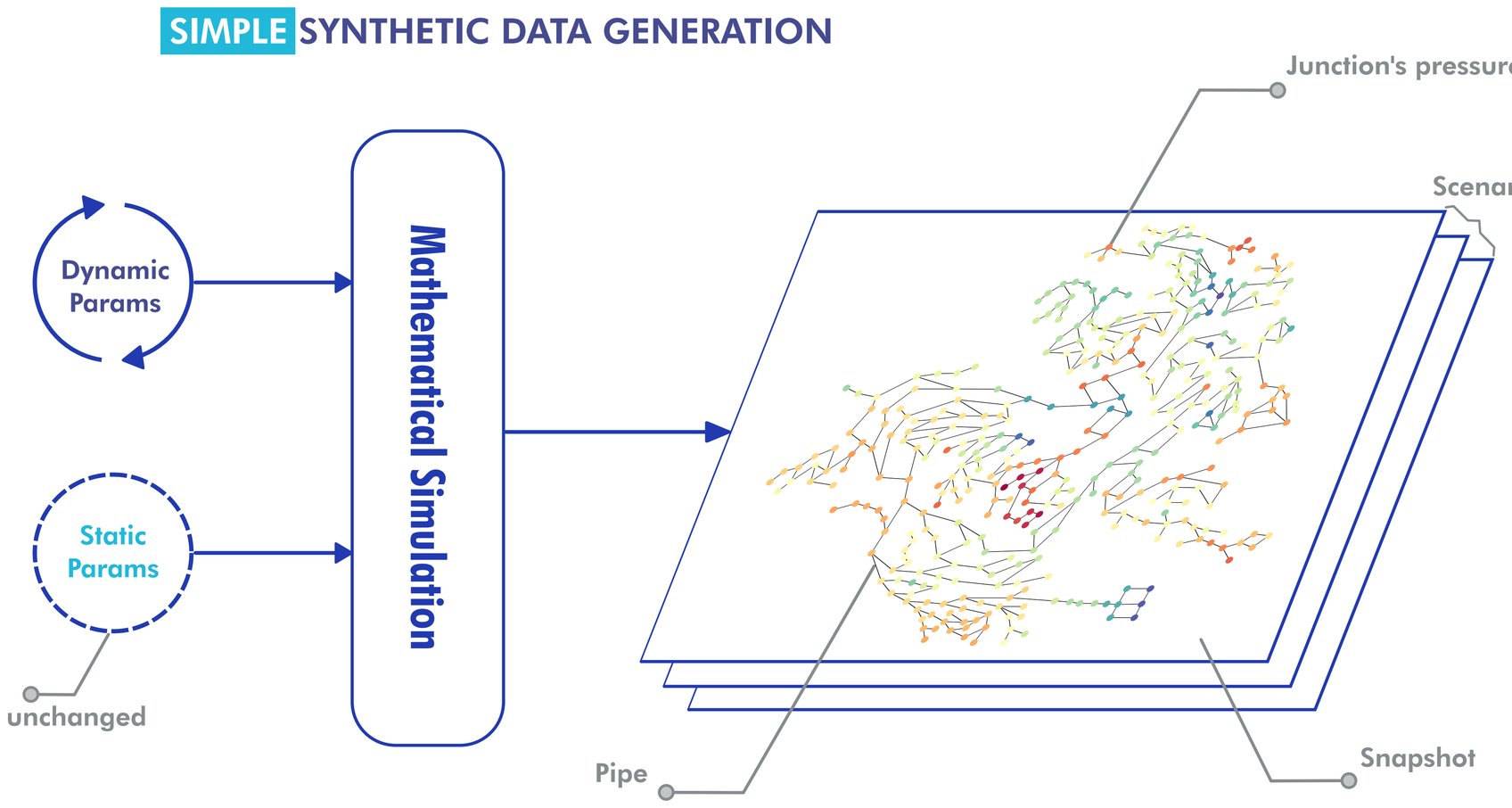
Distorted or high-fidelity dataset for data augmentation in Graph Neural Networks?

Improving pressure estimation in Water Distribution Systems

Huy Truong[1*], Andres Tello[1*], Victoria Degeler[2], Alexander Lazovik[1], Julian Koellermeier[1]

[1]University of Groningen, [2]University of Amsterdam, [*]Equal contribution

SIMPLE SYNTHETIC DATA GENERATION



Background

The objective is to estimate pressure values at unknown junctions given existing sensors.

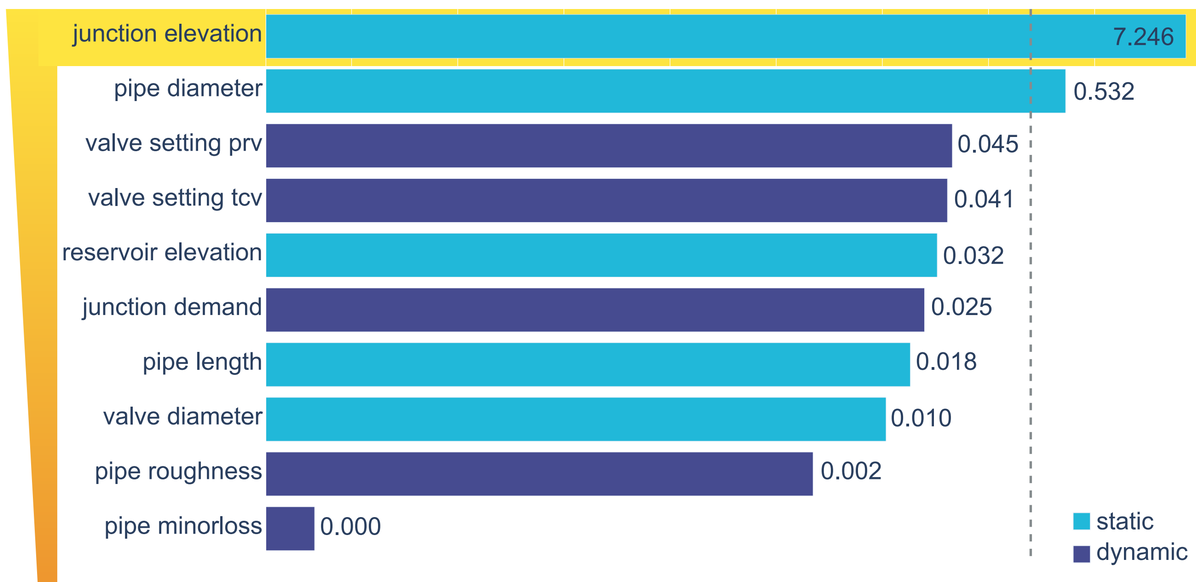
A **Simple** dataset is created using only dynamic parameters that can be changed in reality.

Issues

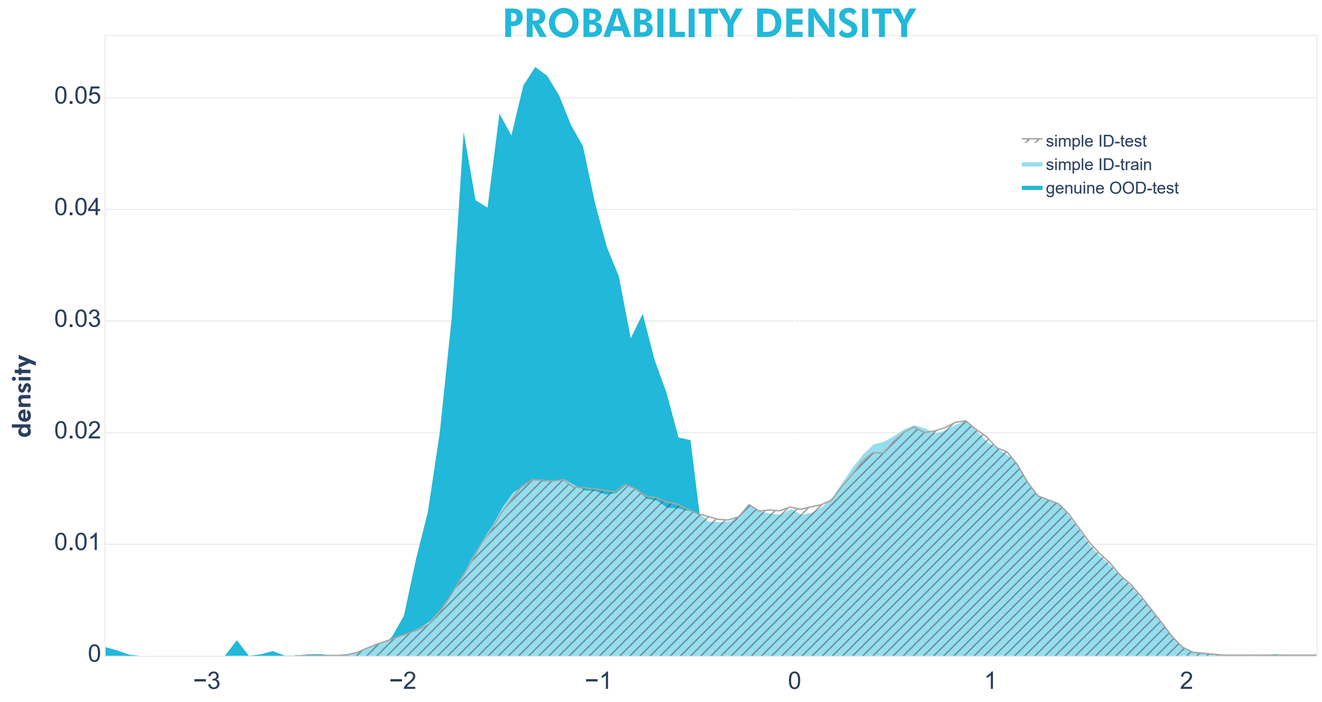
- Limited number of sensors
- Few In-Distribution (ID) scenes
- Test on Out-Of-Distribution (OOD) scenes

SENSITIVITY ANALYSIS

Sensitivity score: $|\text{Neighbor.Var}_{\text{factor}} - \text{Neighbor.Var}_{\text{unchange}}|$

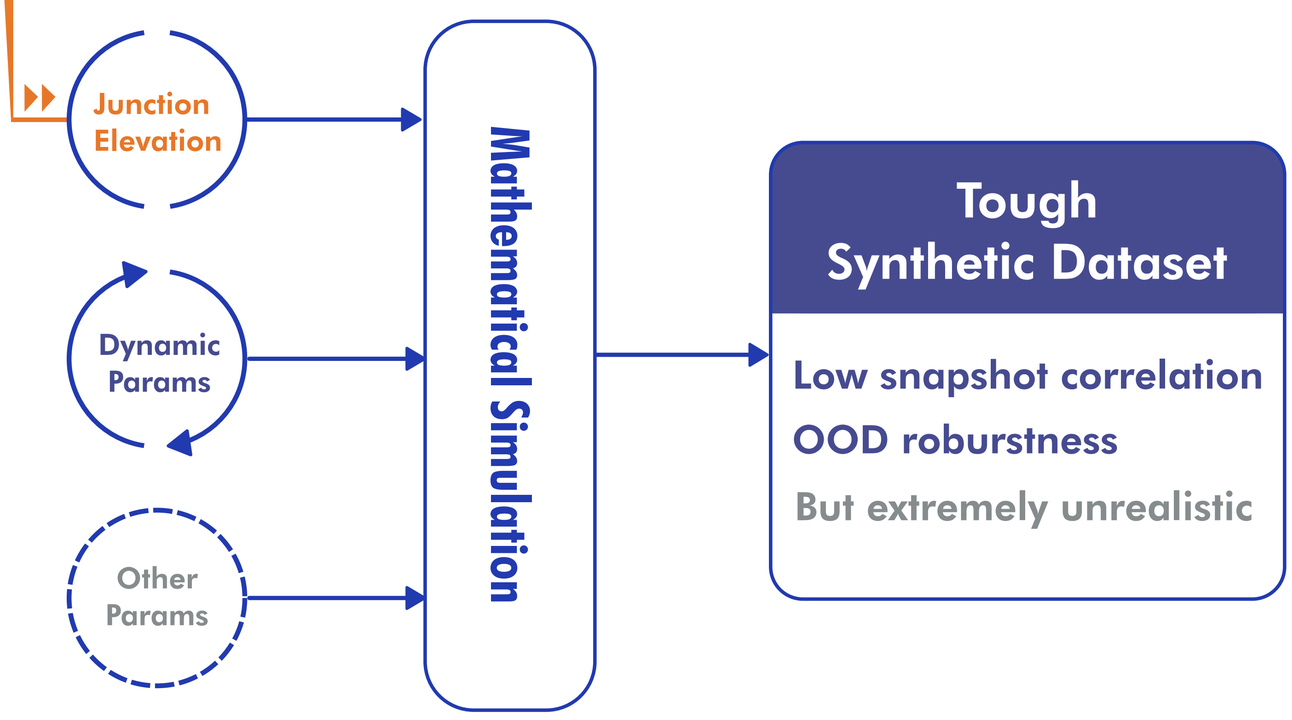


PROBABILITY DENSITY



Idea: Distorting the most sensitivity parameter in the data generation phase

ELEVATION-DISTORTED(TOUGH) SYNTHETIC DATA GENERATION



Methodology

Apply an algorithm to strategically distort the elevation parameters at junctions to provide a distorted **Tough** dataset.

Use the dataset to train a Graph Neural Network to reconstruct pressure values at unknown positions in an unsupervised setting by masking out 95% of nodes and using visible nodes as input.

Test on an OOD data that the model has not encountered during training.

Results on OOD data

The model trained on the **Tough** dataset produces similar results on virtual sensors but shows a noticeable discrepancy on practical sensors compared to the model trained on **Simple** dataset.

Trained on	Virtual Sensors			Real Sensors		
	MAE↓	MAPE↓	NSE↑	MAE↓	MAPE↓	NSE↑
Simple	1.937	0.070	0.773	1.427	0.038	-1.294
Tough	2.017	0.075	0.790	0.713	0.026	0.818

Discussion

- Relationship to Generative Model
- Relationship to Active Learning

university of groningen

ECML PKDD '23
TORINO,ITALY

h.c.truong@rug.nl