Neural Wave Hackathon: Bosch Challenge
Root Cause Analysis of Scrap Increase in Manufacturing

Team: Operators of the Neuron

This concise report highlights our approach, findings, and challenges in identifying root causes of increased scrap rates. Full implementation and all other resources can be found on our GitHub repository or on the lightning.ai platform.

Objective

This project addresses the rise in scrap rates in a manufacturing environment. As production quality decreased, the goal became to identify the root causes of this drop, specifically focusing on factors influencing `Station5_mp_85`, a critical quality metric tied to increased scrap. Accurately identifying these factors allows targeted interventions to minimize scrap and improve efficiency.

Our approach

Two datasets (`low_scrap.csv` and `high_scrap.csv`) provide measurements from 2500 parts before ($\mathcal{D}_{\text{low}}$) and after the quality drop ($\mathcal{D}_{\text{high}}$). Each dataset column represents a specific physical quantity (e.g., `Station2_mp_9`) measured across manufacturing stations. The stations are measured sequentially, imposing a constraint on causal direction between stations but not within.

A directed acyclic graph (DAG) models causal relationships between the physical quantities that were measured. Nodes represent the quantities and a directed edge from one node to another represents causal influence.

Causal Inference

The causal DAG is inferred using conditional independence tests and causal inference methods. We opted for the NoGAM algorithm[1], as implemented in the `dodiscover` Python library, which utilizes 5-fold cross-validation. This algorithm allows us to incorporate our prior knowledge that measurements at station n cannot influence those at station $m < n$, enforcing a causal direction consistent with the sequential nature of the manufacturing process.

We assume that there is one true causal dependency graph underlying both data sets. Ideally, we would train a model on $\mathcal{D}_{\text{low}} \cup \mathcal{D}_{\text{high}}$. Due to time and memory constraints, we were not able to do this and opted for just using $\mathcal{D}_{\text{low}}$.

[1] https://openreview.net/forum?id=rVO0Bx90deu

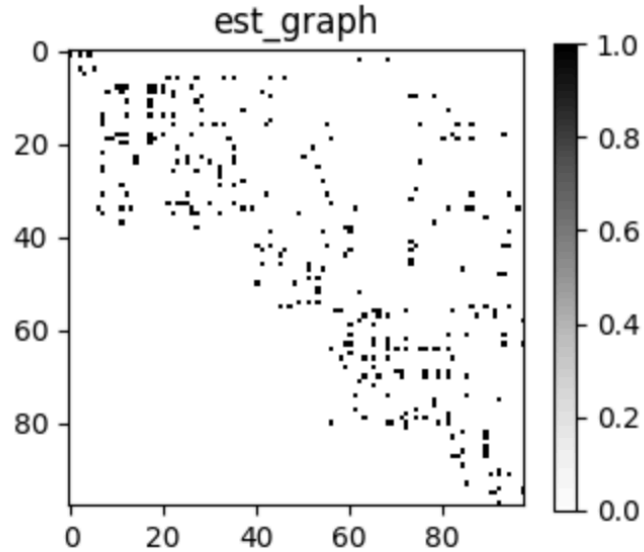

Figure 1: The adjacency matrix of our estimated DAG

As shown in Figure 1, the adjacency matrix of our estimated DAG demonstrates the block-upper diagonal structure, reflecting the sequential measurement process and respecting our prior knowledge constraints.

Root Cause Detection

We analyzed influence rankings of measurements on `Station5_mp_85`, identifying primary contributors to the scrap increase. We utilized the `py-why` library to perform root cause detection based on the anomalous data (`high_scrap.csv`), identifying measurements that significantly contributed to the increase in `Station5_mp_85`.

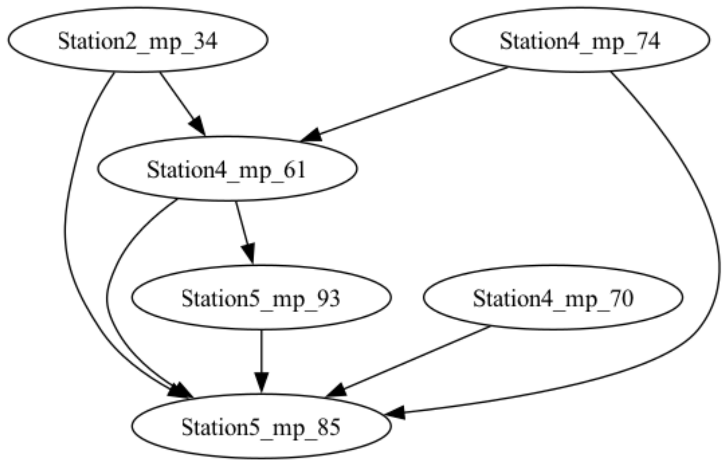

Figure 2: The parent subgraph of our estimated DAG

Determining Effective Interventions

Additionally, we explored Bayesian Networks using the `causalnex` library to compute marginal distributions and simulate interventions, further aiding in isolating key variables affecting the scrap rates.

Results

- Graph Accuracy: A causal DAG comparison of our inferred one with Bosch's ground truth yielded a satisfactory score, confirming relevance for critical relationships.

- Root Cause Identification: Measurement variables were ranked by influence on `Station5_mp_85`. The key contributors upstream of the Node 85 were identified. Root Cause Analysis on the Ground-Truth DAG has resulted in node `Station2_mp_13` being attributed the most significant contributor.

Challenges

- Constraint on Causality through Sequential Dependencies: Temporal ordering between stations added complexity by necessitating the enforcement constraints only allowing for causation flow from upstream to downstream stations.
- Optimization of Causal Graph: Balancing minimal Hamming distance and agreement of ancestors of `Station5_mp_85` to the ground truth turned out to be challenging, when solving for the DAG based on the given data.