Detection of Point-Wise Anomalies in Large-Scale Hierarchical Multivariate Time Series

Author: Magdalena Marie Sarapatkova | Supervisor: Ing. Pavel Zimmermann, PhD. | Prague University of Economics and Business | Faculty of Informatics and Statistics | 2025

Why it matters

Motivation

Modern retailers generate massive, highly seasonal data that is hierarchical (SKU → product → department → store → national) and multivariate (holidays, markdowns, etc.). Yet most detectors struggle with scale, hierarchy, interpretability, or all three. Traditional pipelines stop at forecasting and underuse anomaly detection for decision support.

Managers need timely, interpretable signals about where and why performance deviates. Spotting business-relevant anomalies early enables better stock, promotions, and financial control. This work bridges that gap with an explainable, hierarchy-aware framework.

Research Questions

1. Which anomaly-detection methods suit large-scale, hierarchical, multivariate time series?
2. How do approaches compare in terms of accuracy, scalability, and interpretability?
3. Can a practical, explainable framework that works in real operations be built?

Data

421,570 data points · 3,331 department-level weekly series · 45 stores · 81 departments · 143 weeks.

How it works

The study spanned 51 tuned configurations (17 models × 3 decompositions), plus ensemble strategies including business logic, evaluated in windows with controlled changes to anomaly standard deviation, prevalence, and random seeds.

This thesis introduces a novel hierarchical detection framework and delivers a operational pipeline that accurately detects anomalies in real-world retail data at scale.

1. **Decomposition:** Separate strong seasonality from noise using Prophet to produce residuals for anomaly detection.
2. **Residual scaling:** Normalize residual magnitudes across series with RobustScaler.
3. **Anomaly detection with interpretability:** Run best performing detectors (Thresholding, Isolation Forest, KNN, HDBSCAN, Gaussian Mixture Model, Mahalanobis Distance).
4. **Ensemble with interpretability:** Combine top detectors via Soft Union to maximize recall.
5. **Business filtering:** Enforce materiality (|residual| ≥ $5,000) and context rules (spikes only without holiday/markdown, drops only with holiday/markdown) to boost precision.
6. **Hierarchical reconciliation:** Propagate to store/national only when ≥5% of departments/stores agree on spike/drop, yielding coherent higher-level alerts with explicit drill-down.
7. **Visualization & reporting:** Provide Monday-morning summaries per store and drill-down dashboards; run incrementally with centralized outputs and basic monitoring.

Achievements

✓ Accuracy

- Pre-filter (Soft-Union) recall = 0.93–0.96 across configurations; Post-filter F1 = 0.11–0.30, depending on injection difficulty.
- Harder injections (lower anomaly std) reduced F1 to 0.1091; easier increased it to 0.3003.
- Varying injected percentage from 0.5% → 2% moved F1 from 0.1234 → 0.2817; random seed changes were negligible.

✓ Scalability & Cost

- Runs on Deepnote L4 GPU in ~14 min total; decomposition ≈ 7.5 min, each anomaly detection model < 1 min.
- ~$0.39 per weekly run — negligible relative to operational value.

✓ Business Use

- Monday morning summaries per store: auto-generated, plain-language narratives from pipeline outputs (no manual curation).
- High explainability with simple rules: materiality (|residual| ≥ $5,000), context (spikes only without holiday/markdown; drops only with holiday/markdown), and hierarchy (propagate only if ≥5% of lower hierarchy agree).

SCALE
3,331 series
421k+ points

EXPERIMENT
17 detectors ×
3 decompositions
= 51 configs

RUNTIME & COST
~14 min/run (L4 GPU)
~$0.39/run

PRE-FILTER
Precision 0.016
Recall 0.923
F1 0.032

POST-FILTER
Precision 0.145
Recall 0.358
F1 0.202

EXPLAINABLE
SHAP-like attributions
Simple rules

ACTIONABLE
Hierarchy-aware alerts
Monday summaries

Limitations

- Validation used injected anomalies (pattern-free), which can favor simpler models over deep learning.
- Short history (one full + two partial years) constrained some decomposition methods.
- Compute limits restricted deep models (LSTM AEs, Deep SVDD) at full scale.
- No hierarchical decomposition available; hierarchy handled via post-detection reconciliation.
- No live stakeholder review yet (interpretability is technical; business validation pending).

Future work

- Productionize: dashboards, alerting, incremental pipeline, monitoring.
- Explore daily/SKU granularity; expand anomaly types.
- Add feedback loops/active learning; incorporate labels over time.
- Re-evaluate advanced decomposition and deep models with more data/compute.
- Keep searching for suitable hierarchical decomposition method.