Optimizing Data Quality in Real-Time, Time Series ETL Pipelines Using Monte Carlo Methods and Machine Learning

Robert W. Bakyayita

Uganda Martyrs University Nkozi
Faculty of Science
Department of Computer Science and Information Systems
MSc. Information Systems

July 2025

Outline

Background

Real-time ETL pipelines are critical for processing time-sensitive data in many industries. However, ensuring data quality is challenging due to the high velocity, volume, and variability inherent in time series data streams.

- Traditional batch-based data cleaning methods are often inadequate for real-time requirements.
- Monte Carlo methods offer a powerful way to model uncertainty in data.
- Machine learning techniques provide adaptive capabilities to detect and correct anomalies.
- Combining both approaches enables automated, probabilistic, and adaptive data quality control.

Problem Statement

- There is a lack of adaptive frameworks specifically designed for managing data quality in real-time time series pipelines.
- Despite its potential, Markov Chain Monte Carlo (MCMC) methods remain underutilized in streaming data contexts (Brooks et al., 2011).
- Machine learning techniques for learning from historical error patterns are not widely integrated into real-time data quality solutions.
- There is an absence of standardized benchmarks and metrics to evaluate data quality in real-time ETL pipelines.

Research Objectives

- Develop an MCMC-based module to simulate realistic data corruptions and correct quality issues in real-time.
- Integrate machine learning models that learn from past anomaly corrections to improve prediction accuracy.
- Design and implement a prototype real-time ETL pipeline combining MCMC and ML techniques for data cleaning.
- Evaluate the system's effectiveness using established time series anomaly detection metrics and benchmarks.

Methodology

- **Tools:** Python, Nifi, InfluxDB, Grafana, Scikit-learn, TensorFlow, PyMC3.
- **Approach:**
  1. Data ingestion using Nifi.
  2. Anomaly injection and simulation using MCMC.
  3. ML-based prediction module trained on previous anomalies.
  4. Dashboard visualization with influxQ and Grafana.

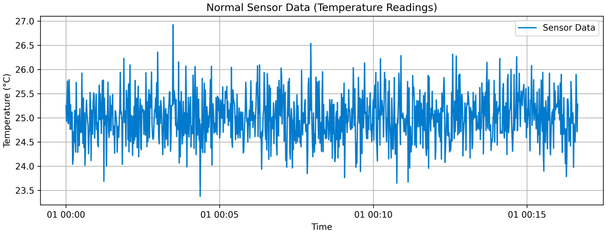
Key Findings

- MCMC helps detect and fix errors in time series data.
- Using past errors improves prediction accuracy.
- Combining methods is more reliable than simple rules.
- Current data quality metrics need improvement.

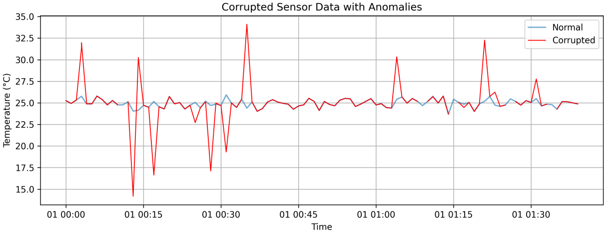
Contributions

- Developed a flexible, containerized system using MCMC and ML for real-time data pipelines.
- Proposed a method for detecting and fixing anomalies in streaming data.
- Tested the system with simulated sensor data.
- Advanced research on data quality in streaming environments.

Simulation Results (1/2)

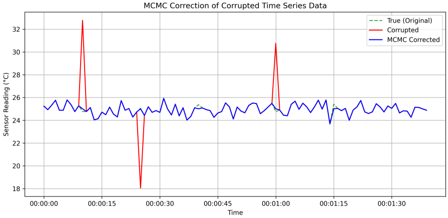


(a) Normal Data

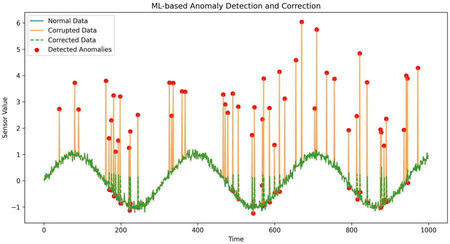


(b) Corrupted Data

Simulation Results (2/2)



(c) MCMC Output



(d) ML Correction

Conclusion and Future Work

Conclusion:

- Demonstrated a practical application of MCMC and ML in enhancing time series data quality.
- Proposed a framework suitable for real-time ETL environments.

Future Work:

- Incorporate federated learning for decentralized anomaly detection.
- Expand evaluation to real industrial data streams.
- Develop standard benchmarking suite for real-time ETL pipelines.

References I

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Robert, C. P. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer.