44th IEEE -- International Performance Computing and Communications Conference

Assessing and Visualizing Completeness, Co-Coverage, and Scalability in Multivariate Time-Series Data

Long Vu∗, Madeline Frank∗, Honghui Xu∗, Sisi Chen†, Tu N. Nguyen∗, Selena He∗, Bobin Deng∗, Kun Suo∗

Outline

1. Introduction

Introduction

- We explore data quality challenges in multivariate time-series datasets
- Focus on air-quality data collected from multiple global locations
- Our goal: understand, quantify, and visualize missingness and co-coverage
- We draw ideas from the visualization research community to approach this problem
- Finally, we present our framework, metrics, and visual tools: combining human insight with automated filtering for reliable analysis

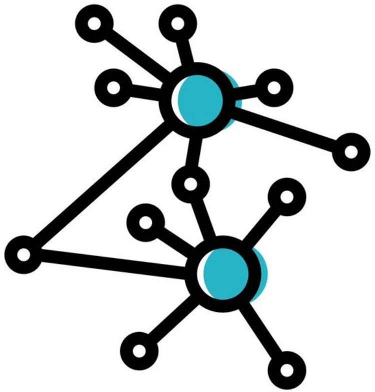
2. Motivation

Motivation — Original Research Questions

- Explore indoor air-quality datasets through data mining across global locations

→ Goal: to uncover environmental and behavioral insights.

Q1. How do key indoor pollutants ($PM_{2.5}$, NO_2, CO, O_3, NH_3, CO_2) interact and correlate under different indoor conditions?



Motivation — Original Research Questions

- Explore indoor air-quality datasets through data mining across global locations

→ Goal: to uncover environmental and behavioral insights.

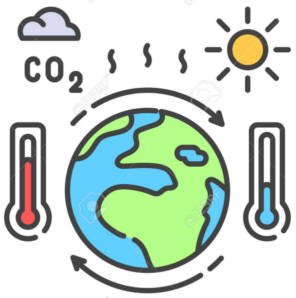
Q2. Are there distinct patterns or pollution levels between countries and geographic regions?

Motivation — Original Research Questions

- Explore indoor air-quality datasets through data mining across global locations

→ Goal: to uncover environmental and behavioral insights.

Q3. How do environmental factors like temperature and humidity affect pollutant variation across time and season?

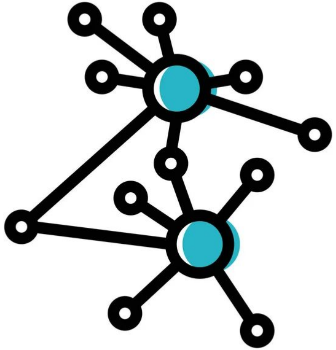


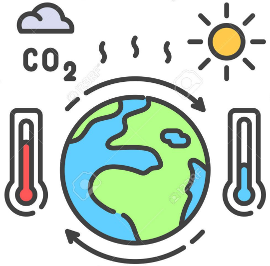
Motivation — Original Research Questions

- Explore indoor air-quality datasets through data mining across global locations

→ Goal: to uncover environmental and behavioral insights.

And many other research questions…





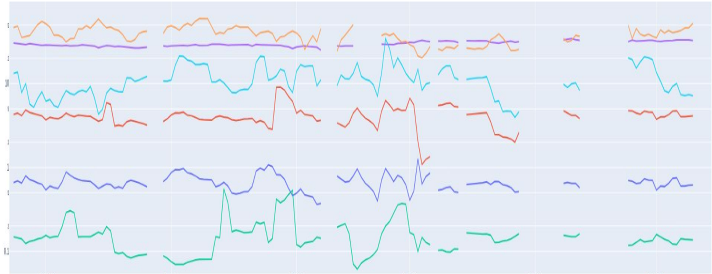
Motivation — Original Research Questions

- Explore indoor air-quality datasets through data mining across global locations

→ Goal: to uncover environmental and behavioral insights.

- We investigated 6 multivariate time-series datasets
  - *India, Mexico, Italy, Sweden, California–Home, California–Apartment*

Dataset Group	Total Rows	Columns (avg±std)	Duration (days)	Sampling Rate
India	173,465	9.0 ± 0.0	603.0 ± 0.0	every minute
Mexico	7,406	10.0 ± 0.0	367.0 ± 0.0	hourly
Sweden*	498,672	7.0 ± 0.0	1370.5 ± 0.5	every 10 minute
Calihome*	672,510	31.1 ± 3.9	6.1 ± 0.5	every minute
Caliapt*	228,083	50.1 ± 10.5	6.3 ± 0.8	every minute
Italy*	28,177,936	18.0 ± 0.0	348.3 ± 2.4	every 2 second

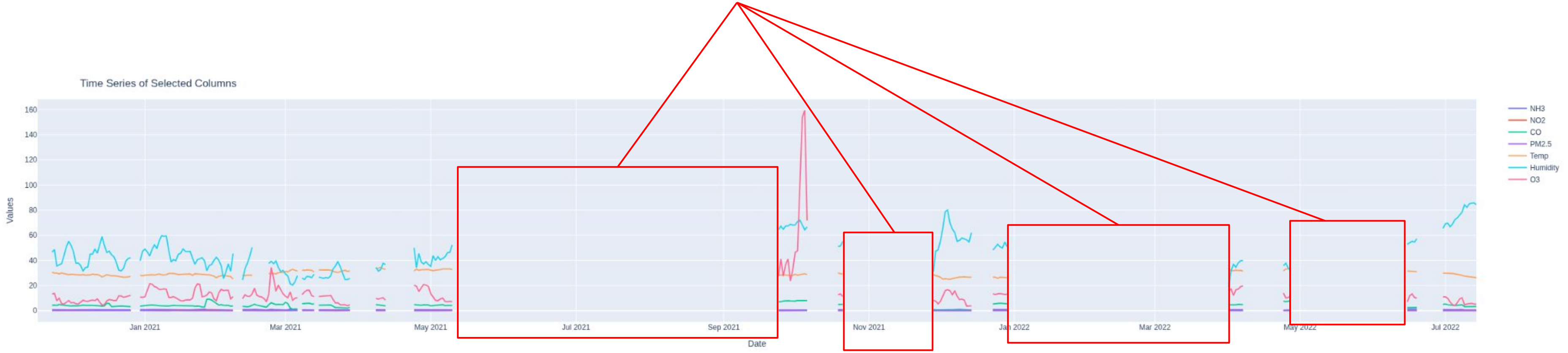


India, Nov 2020 - May 2021

Motivation — The Problem with Air Quality Datasets

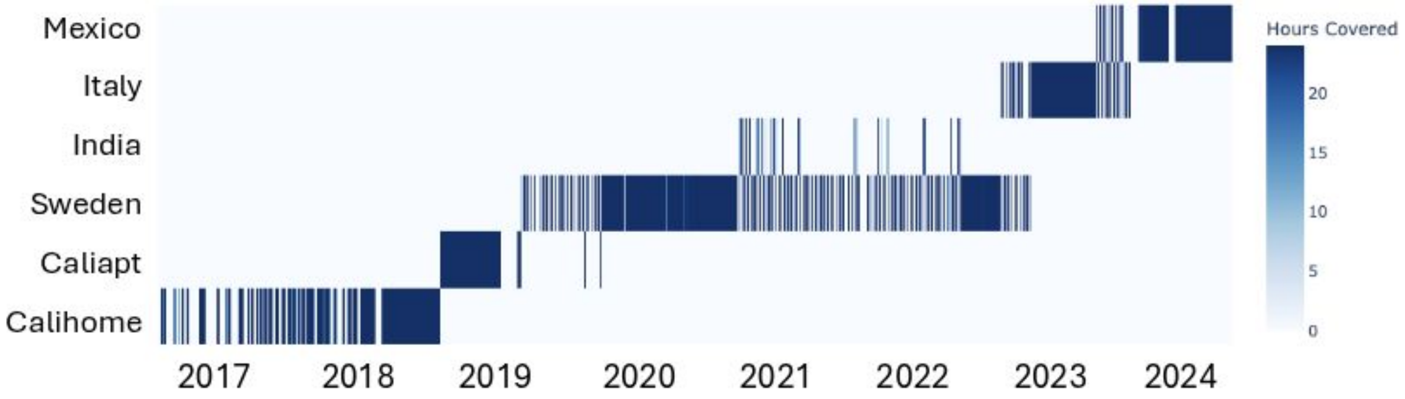
- Massive missing data from sensor failures, inconsistent sampling, and maintenance gaps

Missing data



Motivation — The Problem with Air Quality Datasets

- Massive missing data from sensor failures, inconsistent sampling, and maintenance gaps
- Uneven feature availability – not all pollutants recorded at the same time



Motivation — The Problem with Air Quality Datasets

- Massive missing data from sensor failures, inconsistent sampling, and maintenance gaps
- Uneven feature availability – not all pollutants recorded at the same time
- No clear communication of how severe or structured the missingness is

→ Hard to know which data are usable or how trustworthy conclusions are

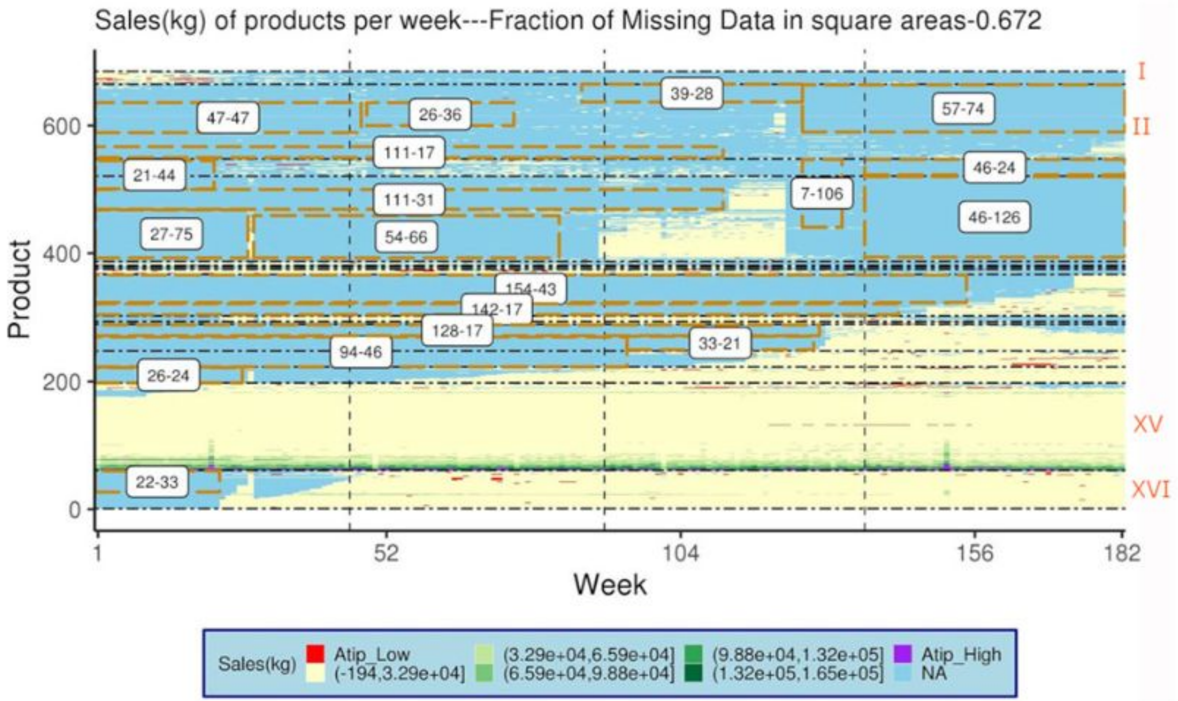
3. Related Work

Related Work — Visualization Community

- The visualization field has systematized approaches to understanding complex data quality issues.
- "Communicating data quality" is an object of study
- Researchers have developed:
    - Taxonomies and frameworks for data completeness and uncertainty
    - Design studies on how humans interpret incomplete data
    - Evaluation methods linking visual design to analytical accuracy

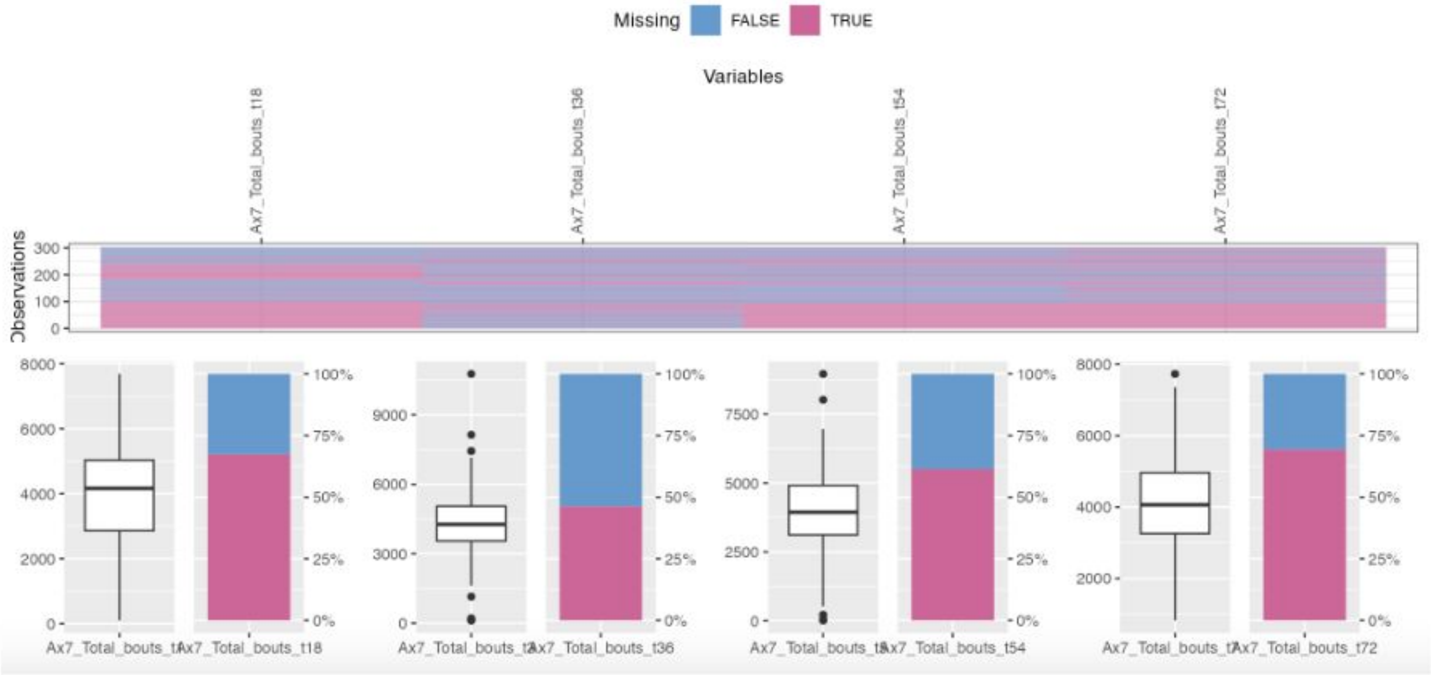
Related Work (cont.)

E. Jimenez and R. Macıas, "*Graphical tools for visualization of missing data in large longitudinal phenomena,*" in Computer Graphics Forum, vol. 41, pp. 438–452, Wiley Online Library, 2022



Related Work (cont.)

S. Alsufyani, M. Forshaw, S. Del Din, A. Yarnall, L. Rochester, and S. J. Fernstad, *"Multi-level visualization for exploration of structures in missing data,"* CGVC. The Eurographics Association, 2024.

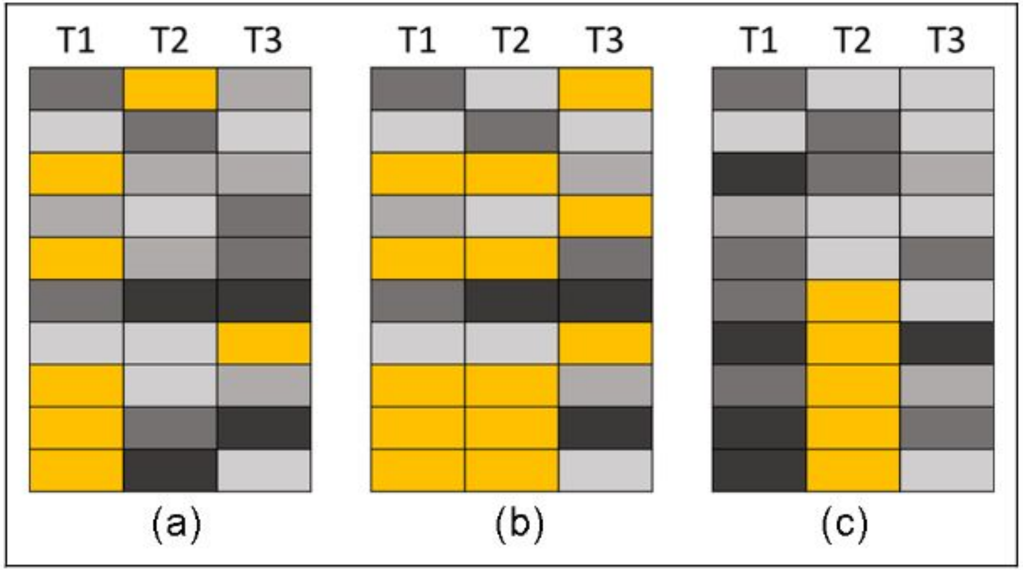


MissVisG prototype

Related Work (cont.)

S. J. Fernstad, *"To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization,"* Information Visualization, vol. 18, no. 2, pp. 230–250, 2019.

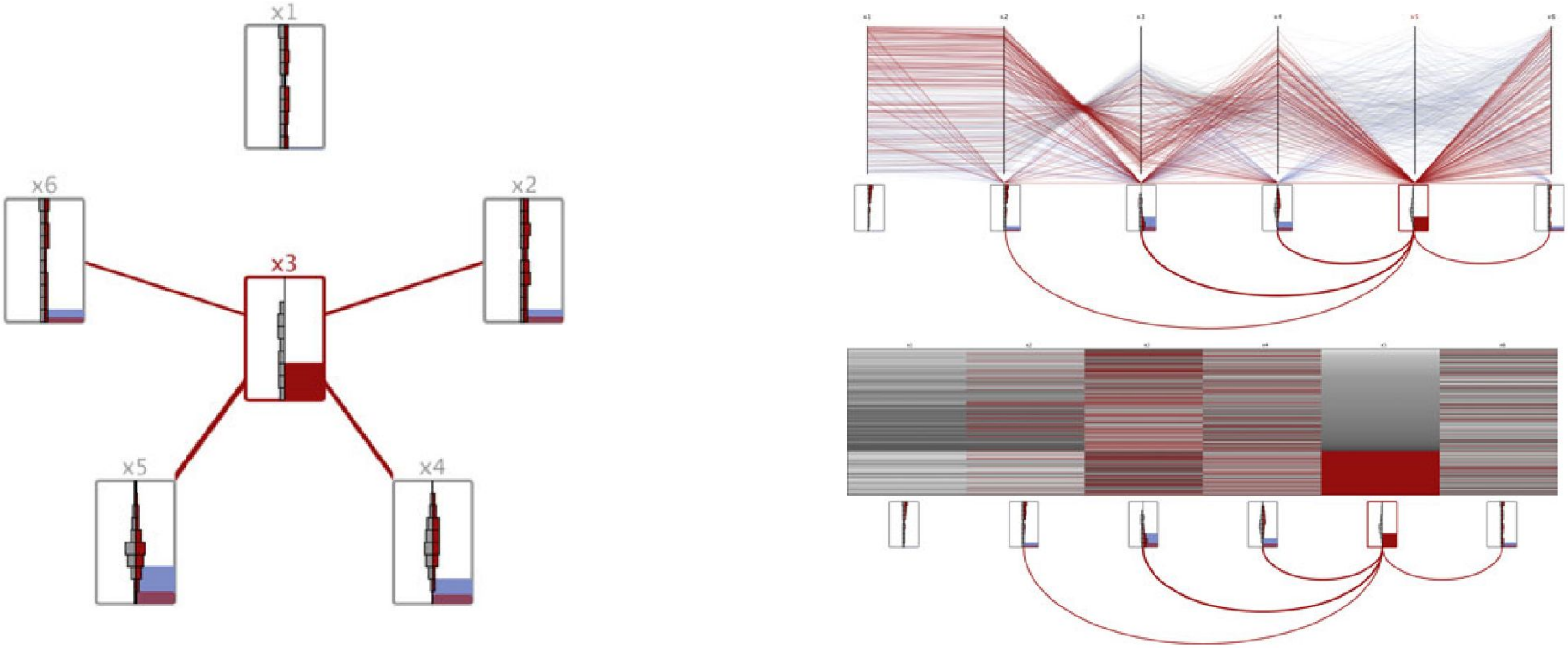
formalized *missingness as analyzable structure*



3 missingness patterns

Related Work (cont.)

S. Alsufyani, M. Forshaw, S. Del Din, A. Yarnall, L. Rochester, and S. J. Fernstad, *"Multi-level visualization for exploration of structures in missing data,"* CGVC. The Eurographics Association, 2024.



Different layouts for analyzing missingness

Related Work — Insights from Literature

1. Data Quality in Information Visualization
2. Visualizing Missing Data
3. Missingness as a Feature

Not designed for multivariate time-series

Rarely capture temporal granularity or cross-feature co-coverage

4. Methodology

Design Goals — What We Aim to Achieve

Challenge	Goal
Large Amounts of Missing Data	Reveal Missingness Patterns and Trends
	Quantify Missingness Impact
Multivariate Co-Coverage	Measure Feature Availability Consistency
	Guide Feature Selection
Scalability of Detection and Verification	Support Multi-Scale Data Summarization
	Enable Efficient Interaction

Our Proposal — Framework and Visualization Toolkit

1. Examine techniques for identifying, quantifying, and visualizing missingness
2. Co-coverage analysis: Assess how consistently features are recorded together and guarantee reliable multivariate relationships
3. Address scalability by introducing an automated framework that standardizes, validates, and visualizes large-scale datasets, aiming at fulfilling design requirements

Assessing Missingness for Time Series Data

1. Identifying Missingness Patterns and Trends
    a. Variable-Specific Trends
    b. Simultaneous Missingness
    c. Temporal Patterns
    d. Correlations with Other Variables

Assessing Missingness for Time Series Data (cont.)

2. Quantifying Missingness
    a. Gap Length Statistics
    b. Temporal Coverage
    c. Period-Specific Missingness

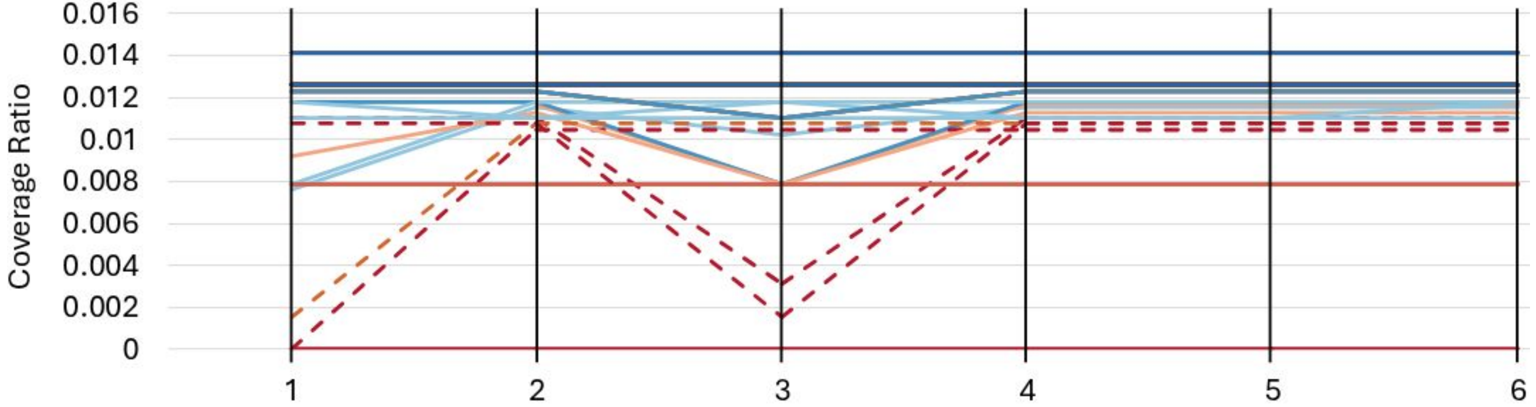
Temporal Coverage Heatmap

Understanding the Data Co-Coverage

1. Measuring Feature Co-Coverage

   co-coverage matrix

2. Identifying Reliable Feature Relationships



Parallel Coordinates Plot of Average Co-Coverage Values for Features Across Datasets

5. Results

Heuristic Filtering Model

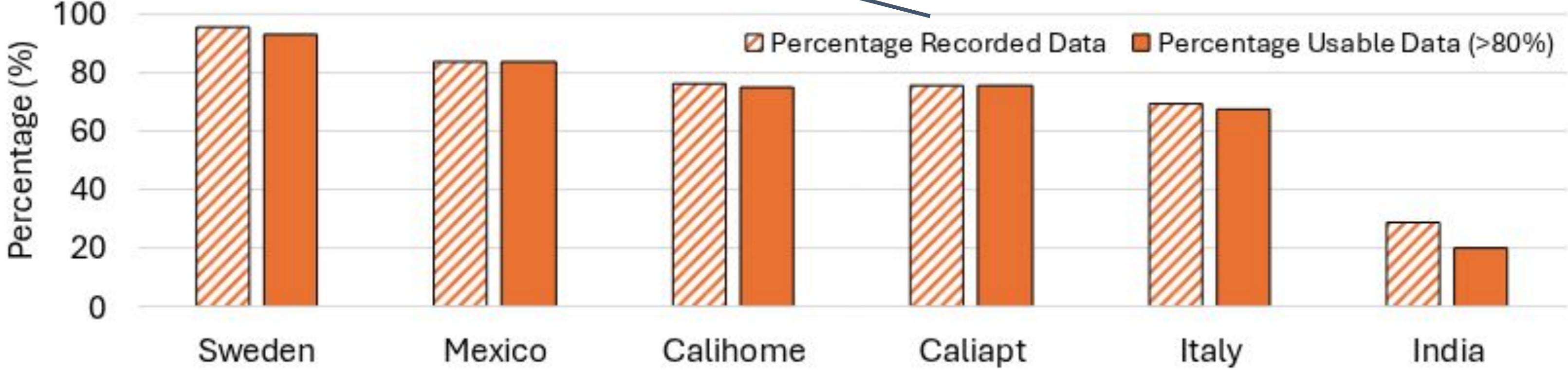
Core Idea

- Split the time series into small intervals
- For each interval:
  - Check if most features are present
  - Check if they overlap in time
  - Keep only intervals that pass both checks

Output: A filtered dataset that preserves meaningful structure while discarding noise and incomplete records.

Results

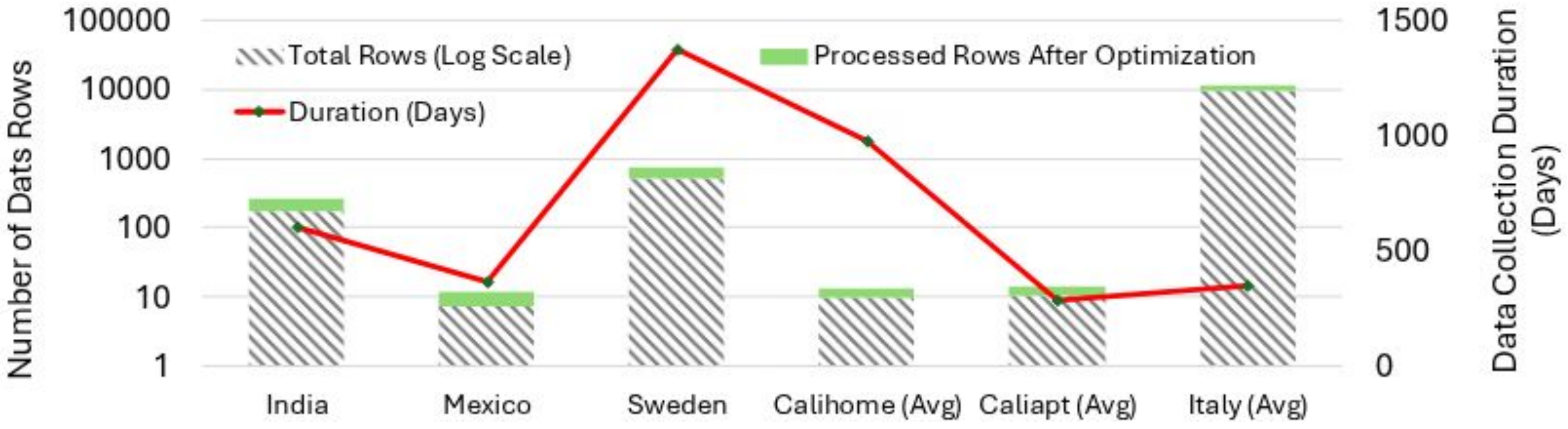
how much data was
originally collected



Baseline Missingness Assessment Across Datasets

Datasets with seemingly adequate data overall can have significantly less actual usable data

Results



Data Reduction across Datasets via our Framework

The final processed data is **significantly smaller**, retaining only the most reliable segments.

6. Conclusion & Future Work

Conclusion & Future Work

- Proposed a human-centric, scalable framework for data completeness assessment
- Combines metrics + visuals + heuristics → interpretable results
- Future directions
  - Dynamic interactive plots (zoom / filter / cluster)
  - AI-based imputation & anomaly detection
  - Broader application → IoT, healthcare, environmental monitoring

Thank You!