Traffic analysis and prediction using Spark and Big Data Algorithms

Authors

Andreea-Daniela Lupu (MIAO)

Nechifor Alexandru (MIAO)

Năstasă Baraș Luca (LC)

Roșcan Teodor (MISS)

Contents

Acknowledgement

The data processing and analysis were carried out with the support of the SMIS 124759 project – RaaS-IS (Research as a Service Iași), funded through the Operational Programme Competitiveness.

("Procesarea și analiza datelor a fost realizată cu sprijinul proiectului SMIS 124759 - RaaS-IS (Research as a Service Iasi), finanțat prin Programul Operațional Competitivitate")

Project Objectives

The main objectives of this project are:

- To analyze large-scale spatio-temporal traffic data using distributed processing techniques.
- To identify temporal, spatial, and infrastructural traffic patterns through scalable analysis.
- To engineer interpretable features that capture periodic traffic behavior and spatial context.
- To apply simple machine learning, graph-based, and similarity-based methods for traffic analysis rather than detailed forecasting.
- To evaluate the applicability of different analytical approaches to large spatio-temporal datasets.
- To study the resulted graphs in order to perform some human level prediction about future traffic

Dataset Description

[Dataset GitHub Link]

Overview

The dataset used in this project is **LargeST (Large-Scale Spatio-Temporal)**, a publicly available traffic dataset designed for large-scale traffic analysis. It contains multi-year traffic measurements collected from thousands of sensors deployed across the California road network, with values recorded at regular time intervals. In addition to traffic measurements, the dataset includes metadata describing sensor location and road characteristics, such as geographic coordinates, number of lanes, direction of travel, and administrative region. The sensors are also connected through an underlying road network structure, which enables graph-based analysis. Due to its size and spatio-temporal nature, the LargeST dataset is well suited for distributed processing and scalable analytical methods.

The dataset consists of ≈ 8,600 traffic sensors deployed on highways and major roads across California and it has 5 years of historical data (2017–2021), providing long-term temporal coverage. Measurements recorded at regular time intervals (e.g., every 5 minutes) are giving very high temporal resolution.

1. Traffic Data (Time-Series)

This is the core dynamic data containing historical traffic measurements (likely traffic flow or speed) recorded by sensors over time.

- `sensor_id`: A unique identifier for the traffic sensor.
- `timestamp`: The specific date and time of the recorded measurement. In the raw data, this may be a float/unix timestamp, which is later converted to a standard timestamp format.
- `value`: The recorded traffic metric (e.g., average traffic flow or speed) for that time interval.
- `year`: A derived column used for partitioning the data storage by year.

2. Station Metadata (Sensor Information)

This table provides static geographical and structural details for each sensor, allowing for spatial analysis (e.g., by county or freeway).

- **`ID`**: The sensor ID (matches `sensor_id` in the traffic data).
- **`Fwy`**: The freeway number where the sensor is located (e.g., 5, 405, 101).
- **`Direction`**: The direction of traffic flow (e.g., N, S, E, W).
- **`County`**: The county in California where the sensor is located (e.g., Los Angeles, Orange).
- **`District`**: The administrative district ID associated with the sensor.
- **`Lng`**: Longitude coordinate of the sensor.
- **`Lat`**: Latitude coordinate of the sensor.
- **`Lanes`**: The number of lanes at the sensor's location, used to analyze traffic volume relative to road capacity.

Data Processing Pipeline (4.406.085.187 rows)

1. **Ingestion**: Raw HDF5 files are read, which contain traffic matrices.
2. **Enrichment**: The traffic data is joined with the metadata using the sensor ID to add location details (County, Freeway, etc.) to every traffic record.
3. **Cleaning**: Timestamps are normalized, and rows with missing values are removed.
4. **Analysis**: The data allows for analyzing hourly traffic patterns, comparing weekday vs. weekend traffic, and visualizing the spatial distribution of congestion across California.

```
Sample data:
+---------+-----+----+-------+---------+-----------+--------+---------+------+-----+--------+---------+----+-------------------+
|sensor_id|value|year|ID     |Lat      |Lng        |District|County   |Fwy   |Lanes|Type    |Direction|ID2 |timestamp          |
+---------+-----+----+-------+---------+-----------+--------+---------+------+-----+--------+---------+----+-------------------+
|1118894  |282.0|2019|1118894|32.729729|-117.107696|11      |San Diego|I805-S|5    |Mainline|S        |7600|2019-12-07 00:00:00|
|1119383  |105.0|2019|1119383|32.729896|-117.107497|11      |San Diego|I805-S|5    |Mainline|S        |7601|2019-12-07 00:00:00|
|1123210  |197.0|2019|1123210|32.740965|-117.116526|11      |San Diego|I805-S|4    |Mainline|S        |7602|2019-12-07 00:00:00|
|1111543  |194.0|2019|1111543|32.748425|-117.12253 |11      |San Diego|I805-S|5    |Mainline|S        |7603|2019-12-07 00:00:00|
|1111542  |204.0|2019|1111542|32.756165|-117.12539 |11      |San Diego|I805-S|5    |Mainline|S        |7604|2019-12-07 00:00:00|
+---------+-----+----+-------+---------+-----------+--------+---------+------+-----+--------+---------+----+-------------------+
only showing top 5 rows
```

State of the Art (SOTA)

The LargeST dataset is primarily used in the literature as a benchmark for large-scale traffic forecasting, especially for evaluating complex deep learning architectures designed to model spatio-temporal dependencies. Due to its size, long temporal span, and explicit road network structure, the dataset is particularly suited for advanced neural models that combine temporal modeling with spatial graph representations.

Recent state-of-the-art approaches typically rely on graph-based neural networks (GNNs), spatio-temporal convolutional models, and attention-based architectures, which aim to capture both local road connectivity and long-range temporal correlations. These models are often trained to predict future traffic values several time steps ahead and are evaluated against LargeST to demonstrate scalability and accuracy on real-world data. In addition to forecasting, some studies also use the dataset to address missing data imputation, applying interpolation techniques or neural models to reconstruct incomplete sensor measurements.

While these approaches achieve strong predictive performance, they require substantial computational resources, including GPU acceleration and long training times, especially given the scale of the LargeST dataset. Training and tuning such deep neural architectures falls outside the scope of this project, whose focus is on distributed data analysis, feature extraction, and scalable analytical methods rather than state-of-the-art forecasting accuracy. As a result, complex neural forecasting models were not implemented, and the project instead emphasizes interpretable, computationally efficient techniques aligned with the objectives of a Big Data Analysis course.

Useful Links

The following works illustrate how the LargeST dataset is used in state-of-the-art research:

1. **LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting**
   The original paper introducing the dataset and its use for evaluating large-scale spatio-temporal models.

2. **STGCN: Spatio-Temporal Graph Convolutional Networks for Traffic Forecasting**
   A foundational graph-based neural architecture widely used as a baseline on traffic

datasets, including LargeST-like benchmarks.

3. **Graph WaveNet for Deep Spatial-Temporal Graph Modeling**
   An advanced GNN-based forecasting model that captures both adaptive graph structure and temporal dynamics.

4. **Spatio-Temporal Transformer Networks for Traffic Forecasting**
   An example of attention-based architectures applied to large-scale traffic prediction and missing data handling.

Similar Products & Relevant Links

1. **LargeST (official repo)**
2. **Apache Sedona (formerly GeoSpark)**
   a. A cluster computing system that extends Apache Spark to efficiently process and analyze massive-scale spatial data (like traffic sensor logs) across distributed machines.
3. **Bigscity-LibCity**
   a. A unified open-source library that reproduces over 70 state-of-the-art spatio-temporal data mining models, allowing researchers to benchmark new traffic forecasting algorithms against standard baselines.
4. **Eclipse GeoMesa**
   a. A distributed spatio-temporal database built for massive datasets, enabling rapid querying and real-time analysis of streaming sensor data by leveraging cloud-native storage like Accumulo or HBase.

List of Scripts and Notebooks (Project Structure)

The project consists of the following Jupyter notebooks and supporting scripts inside the `Traffic_Analysis_Project` folder:

1. **`00_data_preparation.ipynb`**
   Converts the original LargeST data from HDF5 format into Parquet files and stores them on the cluster for efficient processing with Apache Spark.

2. **`01_data_visualization.ipynb`**
   Performs exploratory analysis to visualize spatial, infrastructural, and temporal characteristics of the traffic data.

3. **`02_spatio_temporal_analysis.ipynb`**
   Examines traffic behavior across space and time by categorizing sensors based on flow, variability, and congestion.

4. **`03_simple_ML_algorithms.ipynb`**
   Applies simple machine learning techniques to assess the relevance of engineered features and traffic patterns.

5. **`04_graph_community_finding.ipynb`**
   Uses graph-based methods to identify communities within the traffic sensor network.

6. **`05_lsh_algorithm.ipynb`**
   Applies Locality-Sensitive Hashing to identify sensors with similar temporal traffic profiles.
7. **Utility script (`utils.py`)**
   Contains shared helper functions used across multiple notebooks.

Results

This section presents the main findings obtained from the large-scale analysis of the LargeST dataset. The results are organized thematically to reflect the different analytical perspectives applied throughout the project, ranging from exploratory analysis to similarity search, while emphasizing scalability, interpretability, and structural insights.

Spatial and Infrastructural Characteristics of the Traffic Network

The analysis reveals a highly uneven spatial distribution of traffic sensors across California. Sensor deployment is strongly concentrated in large metropolitan counties, particularly Los Angeles, Orange, and San Diego, while rural and less populated regions exhibit significantly sparser coverage. A similar concentration pattern is observed at the freeway level, where a limited number of major highways (e.g., US-101, I-5) host a disproportionate share of sensors. This confirms that the dataset primarily captures traffic behavior along critical transportation corridors.

Geographic visualizations show that sensors are not randomly distributed but form continuous spatial chains aligned with road infrastructure. These chains closely follow freeway layouts and urban road networks, highlighting the inherently graph-structured nature of the dataset. Sensors belonging to the same administrative districts and freeways tend to cluster spatially, reinforcing the importance of spatial context and motivating the use of graph-based analytical methods.

Infrastructure characteristics also play a significant role in traffic behavior. Roads with a higher number of lanes consistently exhibit higher average traffic flow, although the relationship is not strictly linear at higher lane counts. This suggests that while infrastructure capacity is a key determinant of traffic intensity, additional factors such as location, connectivity, and demand patterns must also be considered.

Temporal Traffic Dynamics and Periodicity

Strong and consistent temporal patterns are observed across the entire dataset. Hourly traffic profiles exhibit a clear daily cycle, with low traffic during nighttime hours, sharp

increases during morning commute periods, sustained daytime activity, and evening peaks followed by gradual declines. This pattern remains stable across multiple years, with only moderate variations in overall intensity.

A clear distinction emerges between weekday and weekend traffic behavior. Weekdays display pronounced morning and evening peaks corresponding to commuting patterns, whereas weekend traffic is more evenly distributed throughout the day, with reduced morning activity and later peak times. These findings confirm the relevance of temporal features such as hour-of-day, day-of-week, and rush-hour indicators for traffic analysis.

When traffic behavior is examined across predefined time periods (weekday peak, weekday off-peak, weekend), weekday peak periods consistently exhibit the highest average flow, highest congestion levels, and lowest proportion of empty-road conditions. In contrast, off-peak and weekend periods show lower average flow, higher variability, and significantly reduced congestion. These results highlight the strong interaction between time-of-day and traffic intensity and justify the separation of traffic behavior into temporally meaningful regimes.

Below is a detailed explanation of our spatio-temporal analysis describing our 24-hours approach. Later on we will discuss the peak/off-peak approach.

We analyzed traffic flow in 2 stages:

- First, we aggregated 5 years of 5-minute interval readings into sensor-level statistics, computing average flow per lane and the coefficient of variation (CV = σ/μ) for stability across timestamps.
- Sensors were then classified into six distinct patterns based on flow intensity thresholds (80, 140, 180 veh/5min/lane) and stability metrics (CV threshold of 0.5).

The prevalence of "Light & Variable" (69.2%) over "Light & Stable" (13.6%) suggests that most low-flow corridors experience significant temporal fluctuations, which will become clearer in the subsequent peak-hour analysis.

After classifying individual sensors, the analysis aggregates data to the highway corridor level to identify which routes experience the most widespread congestion. This shift from sensor-level to highway-level analysis provides strategic insights for regional transportation planning.

The dataset includes dataset 132 highways representing major arterials and freeways with adequate sensor coverage.
Overall System Health Score: 97.7% Normal

- Critical highways (>15% congested): 0 (0.0%)
- Warning highways (5-15% congested): 3 (2.3%)
- Normal highways (<5% congested): 129 (97.7%)

These 3 warning highways are:

- SR142-E (Congestion: 13.7% of time at/over capacity,Avg flow: 107.1/veh/5min/lane, Capacity utilization: 59.5%)
- SR142-W (Congestion: 11.5% of time at/over capacity, Avg flow: 94.1 veh/5min/lane, Capacity utilization: 52.3%)
- SR25-N (Congestion: 9.7% of time at/over capacity, Avg flow: 68.3 veh/5min/lane, Capacity utilization: 38.0%)

Only **7 sensors (0.08% of the network)** experience congestion more than 15% of the time under 24-hour averaged conditions, indicating that chronic capacity failures are rare and geographically isolated. I-605 ranks first place as most congested.
The proximity of these hotspots (I-605, I-5, I-105 all in southeastern LA County within ~10 miles) suggests a regional bottleneck complex rather than isolated point failures.

Our initial 24-hour averaging approach produced misleading results. By aggregating traffic flow across all hours of the day over 5 years, we smoothed out critical peak-hour congestion events. For example, highways with severe rush-hour bottlenecks (e.g., 200 veh/5min/lane during 8am) were diluted by low overnight flows (e.g., 20 veh/5min/lane at 3am), resulting in moderate averages (~100 veh/5min/lane) that masked operational problems.
Revised Approach:
We addressed this by disaggregating data into
distinct time periods:

- Weekday Peak (7-9am, 4-7pm): Rush hour traffic
- Weekday Off-Peak: All other weekday hours
- Weekend: Saturday-Sunday traffic

We also conducted regional traffic analysis. Los Angeles is the most stressed regional network as it demonstrates the highest traffic intensity across all time periods and experiences the most dramatic peak-hour degradation:

Capacity Utilization by Time Period:

- Weekday Off-Peak: 40.2% (72.3 veh/5min/lane)
- Weekday Peak: 58.7% (105.6 veh/5min/lane) ← Highest of all regions
- Weekend: 40.7% (73.2 veh/5min/lane)

The comparison of the top 20 most congested sensors during weekday peak hours (7-9 AM, 4-6 PM) versus off-peak periods reveals the dramatic transformation that occurs during commute windows—and exposes how severely 24-hour averaging understates congestion problems - Key Finding: Most sensors experience +20 to +32 percentage point increases in congestion frequency during peak hours. Some sensors that are congested ~25-30% of the time during off-peak hours surge to 50-68% congestion during rush hours.

The I-605 South sensor (766937) stands as the archetypal chronic failure, ranking first in both the 24-hour analysis with 59.2% congestion and the peak-hour analysis with 67.9% congestion. The relatively modest 8.7 percentage point increase from 24-hour to peak-hour conditions indicates that this location operates in a degraded state continuously. Unlike episodic bottlenecks that experience dramatic peak-to-off-peak swings of 25-30 percentage points, the I-605 sensor maintains congestion levels above 50% throughout the day, with peak hours merely intensifying an already failed condition. The off-peak congestion analysis further confirms this characterization, showing that even during non-commute periods, this sensor experiences 57.5% congestion—a level that would qualify as severe peak-hour congestion on most other highways.

Feature Relevance and Low-Dimensional Structure

Simple machine learning models were employed primarily for feature validation and interpretability, rather than for optimizing predictive performance. Linear regression results indicate that infrastructure capacity (number of lanes) and temporal dynamics are the dominant drivers of traffic flow. In particular, the strong coefficients associated with the cyclic hour-of-day features confirm that traffic behavior follows a pronounced daily periodic pattern, while rush-hour indicators further reinforce the role of peak commuting times. Weekend

indicators exhibit a comparatively weaker influence, suggesting that global traffic dynamics are more strongly governed by daily cycles than by weekly distinctions.

Dimensionality reduction through Principal Component Analysis (PCA) further supports these observations. The leading principal component is largely driven by road capacity, while subsequent components capture temporal variation associated with daily traffic cycles. The first few components explain a substantial portion of the overall variance, indicating that traffic behavior—despite being collected across thousands of sensors and time points—can be effectively described using a low-dimensional representation. This structured variance distribution confirms that traffic dynamics are governed by a small set of dominant factors rather than high-dimensional noise.

In addition, comparisons between linear regression and tree-based models (Random Forest) reveal that while linear models capture global trends, non-linear models better accommodate variability under high-traffic and congested conditions. However, the primary purpose of these models remains explanatory, demonstrating that classical machine learning techniques are sufficient to extract meaningful structure and insights from the dataset, without resorting to computationally expensive deep learning approaches.

Graph Structure and Community Detection

Modeling the traffic sensor network as a graph reveals meaningful structural organization. Community detection identifies a large number of network communities, with a strongly skewed size distribution: many small clusters and a few larger, structurally dominant ones. This long-tailed distribution is characteristic of real-world transportation networks.

Spatial visualization of detected communities shows strong geographic coherence. Communities are spatially contiguous and often correspond to specific road segments, highway sections, or urban regions. This demonstrates that graph-based clustering captures real structural divisions within the road network rather than arbitrary groupings.

Traffic behavior varies significantly across communities. Some clusters consistently exhibit higher average flow and congestion levels, while others remain lightly loaded. These differences confirm that traffic dynamics are closely tied to network topology and connectivity, and that graph-based analysis provides complementary insights to purely temporal or feature-based approaches.

Similarity Search Using Locality-Sensitive Hashing

Locality-Sensitive Hashing (LSH) was applied to identify sensors with similar temporal traffic profiles in a scalable manner. Weekly traffic patterns (168-hour cycles) were used as the basis for similarity comparison. Sensors retrieved as similar by the LSH procedure exhibit highly aligned temporal behavior, with synchronized daily peaks, comparable amplitudes, and consistent responses relative to congestion thresholds.

The comparison of weekly profiles demonstrates that LSH successfully captures similarity in temporal dynamics without requiring exhaustive pairwise comparisons. This confirms that similarity search is feasible at scale and provides an additional analytical perspective that complements structural similarity derived from graph analysis. While graph-based methods emphasize physical connectivity, LSH focuses on behavioral similarity, allowing sensors from different regions to be compared based on how traffic evolves over time.

Summary

Overall, the results show that:

- Traffic behavior in the LargeST dataset is highly structured both temporally and spatially.
- Most of the network operates under light or moderate traffic conditions, with congestion concentrated in a very small number of localized segments.
- Temporal periodicity and infrastructure capacity are the most influential factors shaping traffic flow.
- Graph-based and similarity-based methods uncover complementary aspects of traffic behavior that are not visible through traditional aggregation alone.
- Distributed analysis using Apache Spark enables scalable exploration and interpretation of complex spatio-temporal traffic data.

These findings demonstrate the value of combining multiple analytical paradigms when working with large-scale traffic datasets and highlight the suitability of the LargeST dataset for distributed data analysis beyond pure forecasting tasks.

Team Task Distribution

Documentation

- Roșcan Teodor, Nechifor Alexandru, Lupu Andreea-Daniela, Nastasă Baraș Luca

RaaS Setup and Connection:

- Roșcan Teodor, Nechifor Alexandru, Nastasă Baraș Luca

Data Visualization (Preparation-Visualization-Analysis):

- Nechifor Alexandru, Lupu Andreea-Daniela

Clustering:

- Lupu Andreea-Daniela, Nastasă Baraș Luca

ML Scripts:

- Nechifor Alexandru

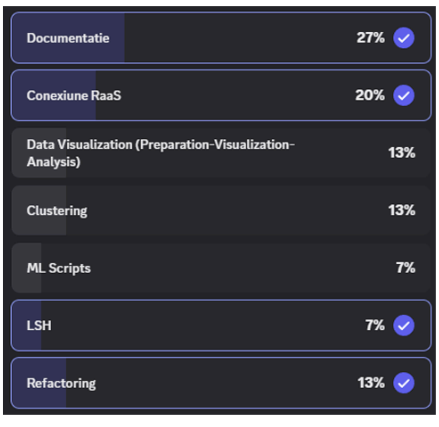
LSH:

- Roșcan Teodor

Refactoring:

- Roșcan Teodor, Nechifor Alexandru

Roșcan Teodor = RT
Nechifor Alexandru = NA
Lupu Andreea-Daniela = LAD
Nastasă Baraș Luca = NBL



RT, NA, LAD, NBL

RT, NA, NBL

NA, LAD

LAD, NBL

NA

RT

RT, NA

Conclusions

After implementing this project, we can conclude the following:

The project utilized Apache Spark to perform scalable analysis on the LargeST dataset (≈8,600 sensors, 2017–2021), focusing on distributed processing and feature extraction rather than deep neural forecasting.

Spatial Findings: Traffic congestion is not random but highly localized to specific metropolitan segments (e.g., Los Angeles), and graph-based community detection successfully grouped sensors by physical road infrastructure.

Scalability: The application of Locality-Sensitive Hashing (LSH) proved effective for finding sensors with similar temporal behaviors without the computational cost of exhaustive pairwise comparisons.

Bibliography

LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting

The original paper introducing the dataset and its use for evaluating large-scale spatio-temporal models.

STGCN: Spatio-Temporal Graph Convolutional Networks for Traffic Forecasting

A foundational graph-based neural architecture widely used as a baseline on traffic datasets, including LargeST-like benchmarks.

Graph WaveNet for Deep Spatial-Temporal Graph Modeling

An advanced GNN-based forecasting model that captures both adaptive graph structure and temporal dynamics.

Spatio-Temporal Transformer Networks for Traffic Forecasting

An example of attention-based architectures applied to large-scale traffic prediction and missing data handling.