Generation of Mobile Phone Traffic Data

Supervisor: Prof. Fabio Pinelli Author: Alessandro Lo Verde

MD2SL 2^{nd} Master Degree in Data Science & Statistical Learning

24 April 2024

Index

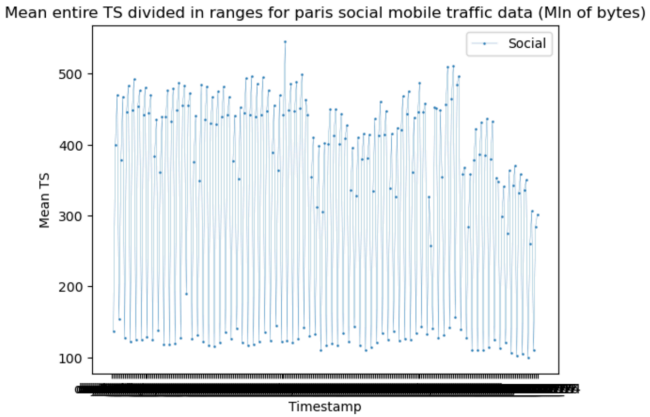
Mobile Traffic data: NETMOB 2023 Dataset

- **Data Source**: NETMOB 2023 challenge website
- **Cities**: Paris, Lyon, Marseille
- **Original Time Span**: 16th March 2019 to 31st May 2019
- **Granularity**: Recorded at 15-minute intervals within 100m x 100m squares

Aggregation Process

1. **Spatial Aggregation: IRIS cell-time series pairing**
   - The squares containing a time series for a certain mobile application have been associated with the IRIS cells with the largest area of intersection. IRIS cells are equal-sized units with approximately 2000 residents each (standard from INSEE).
2. **Categories aggregation within the IRIS cell:**
   - Social Media (Facebook, Instagram, LinkedIn, Pinterest, Snapchat, Twitter)
   - Gaming (Clash of Clans, EA Games, Fortnite, Playstation, Pokemon GO)
3. **Temporal Aggregations**:
   - **Entire Time Series aggregated per 8 hour intervals ranges**:
     - Night (0:00 - 8:00), Morning (8:00 - 16:00), Afternoon (16:00 - 24:00)
   - **Weekly hourly average time series**:
     - Entire time series aggregated hourly and than averaged across all weeks

Time patterns in the Average Time series



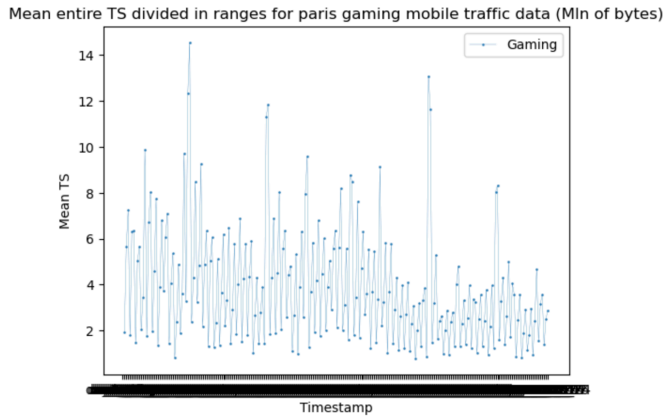
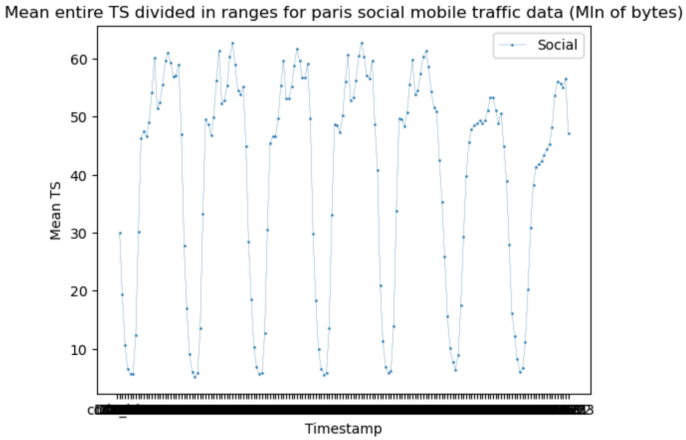


Figure: Average across all IRIS cells of Paris of the Entire Series aggregated into 8 hrs ranges - Social(left), Gaming (right)



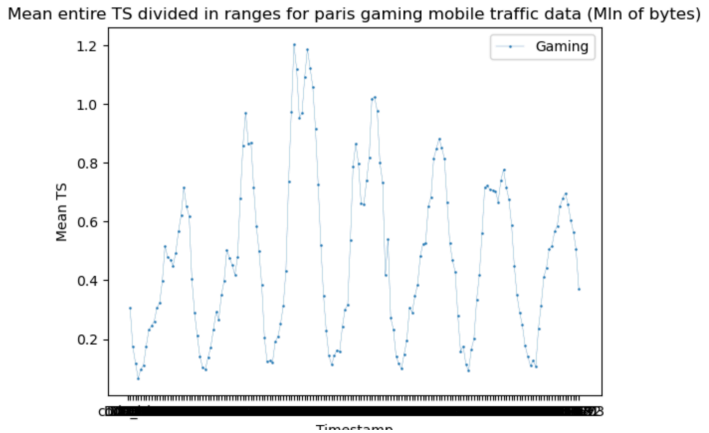


Figure: Average across all IRIS cells of Paris of the Weekly Hourly average TS - Social(left), Gaming (right)

Is it possible to predict mobile traffic data using geographic and demographic features?

Research Questions

- Is it possible to leverage on geographic and demographic areal data to generate a synthetic time series of mobile traffic data consumption?
- Which features are relevant for this purpose?
- Are all the categories of mobile applications (social networks, gaming) similarly predictable?
- Can the model trained in one city be extended to predict the time series of mobile traffic data in another city?

Features selected to predict the time series

- **Cities**: Paris (2635 IRIS cells), Lyon (411 IRIS cells), Marseille (336 cells)

Geographic, Demographic and Infrastructural data from INSEE Dataset (171 features)

- **Area** in m^2 of the IRIS cell
- **Distance from the city centre**
- **Demographic indicators** (e.g., age, gender, job, nationality, income, etc..)
- **Housing** (e.g., total number, type and year of building, etc...)
- **Electricity consumption** (e.g., for residences, for each economic sector, etc...)

Points of Interest (POIs) from OpenStreetMap (157 features)

- **Amenities, Offices and Shops** (e.g. drinking water, restaurant, cafe, school, bank, post box, estate agent, government, clothes, hairdresser, bakery, butcher, etc...)
- **Healthcare** (e.g. pharmacy, hospital, dentist, etc...)
- **Historic and Tourism** (e.g. monuments, ruins, museum, gallery, artworks, etc...)
- **Land use** (e.g. residential, forest, grass, etc...)
- **Leisure** (e.g. garden, sports centre, park, etc...)
- **Public transport and Railway** (e.g. train station, rail, bus stop, subway, etc...)

The larger the IRIS cell the higher the Traffic data consumed

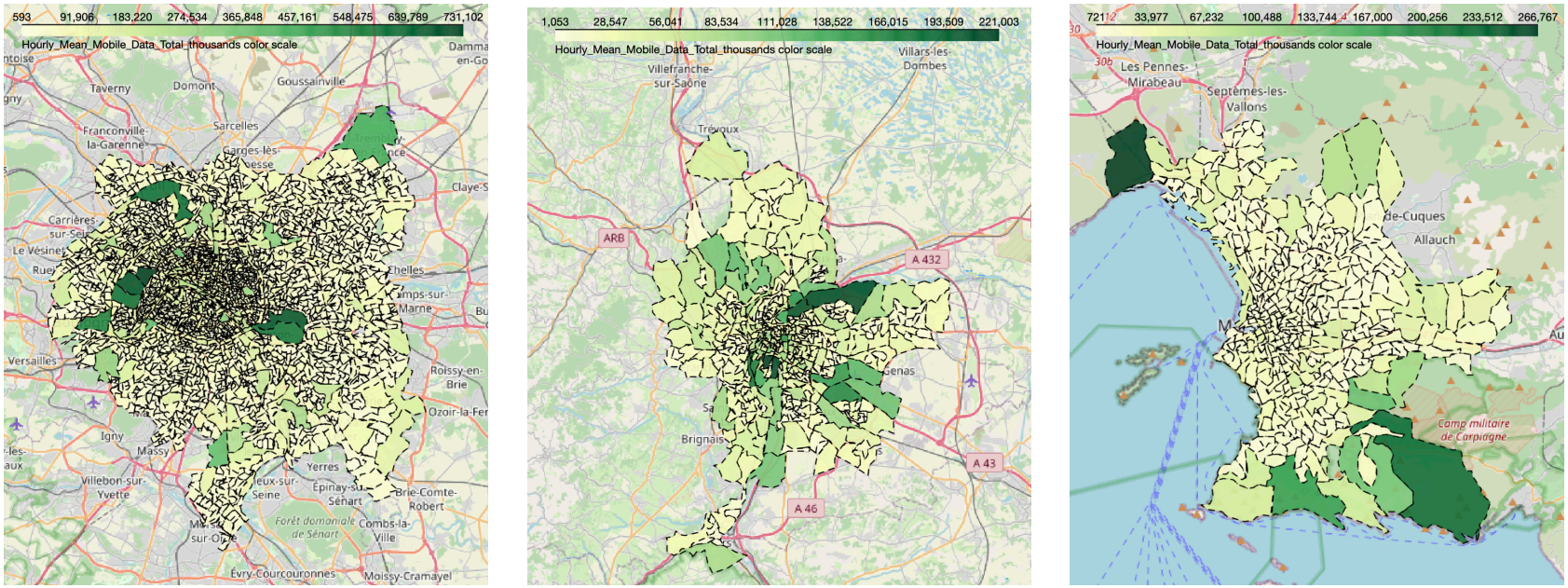


Figure: Distribution of the average traffic data used for the Social TS (thousands of bytes) for Paris (left), Lyon (central), Marseille (right)

Traffic data density (Bytes per 100 m2)

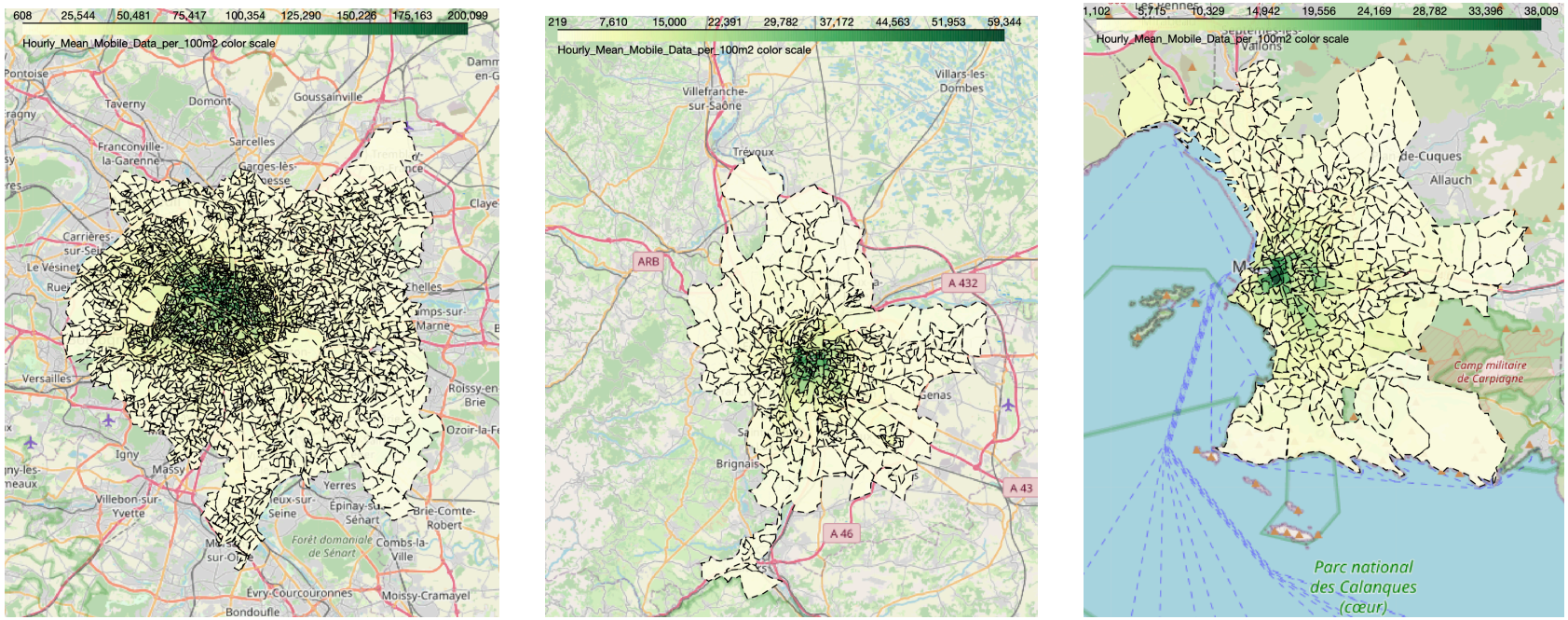


Figure: Distribution of the average traffic data density used for the Social TS (bytes per 100 m2) for Paris (left), Lyon (central), Marseille (right)

Are the features selected relevant to predict traffic data?

KNN Method

- K-nearest neighbors to cluster similar IRIS cells wrt a target cell based on the features available.
- The distance (RMSE) from the target time series time series is computed.
- Table below resumes the average RMSE across all IRIS cells of Paris considering a different number of n similar and n random neighbours.

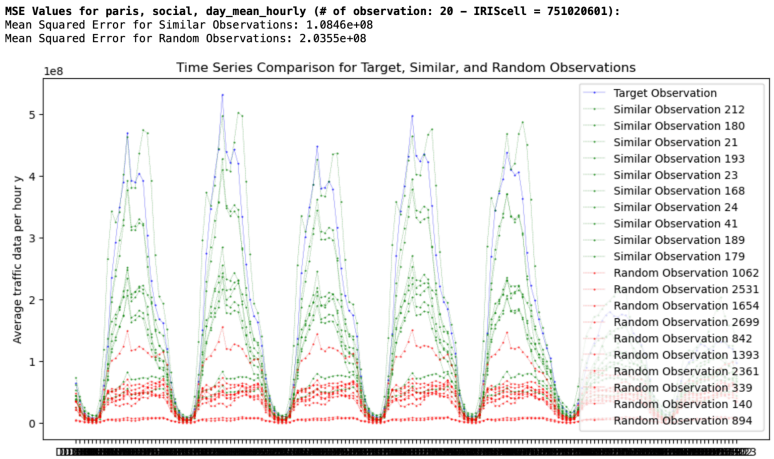


Figure: 10 nearest similar neighbors (green) vs 10 random neighbours (red) wrt a target IRIS cell (blue)

Average RMSE Results for Paris (Weekly Social TS)

Number of Neighbors	KNN RMSE	Random RMSE
26 (1%)	5.46×10^7	7.81×10^7
131(5%)	6.32×10^7	7.89×10^7
163 (10%)	6.74×10^7	7.87×10^7

Problem Definition

Framework

- City partitioned into IRIS cells with equal population, each one characterized by a set of features. For each cell, a fragment of a time series describing the traffic volume used for different categories of mobile applications (Social, gaming,...) is available.
  - City: $C = \{Cell_1, Cell_2, ..., Cell_K\}$
  - Features describing each cell : $F = \{Feature_1, Feature_2, ..., Feature_M\}$
  - Historical TS of Traffic data: $T^g = \{T_1^g, T_2^g, ..., T_L^g\}$.
  - Categories: $g = \{Category_1, Category_2, ..., Category_H\}$

Objective

- Find models $M^g : F \rightarrow T^g$ for each category g of apps which better predict the time series of mobile traffic data in each cell based on the set of features available
- Output should mirror real data's dynamics and consumption levels.

ML Models: Neural Network and XGBOOST

Fully connected Neural Network Model

The **architecture** selected for this work is the following one:

- 3 Hidden layers with 2000 neurons each (Activation function: ReLU, Dropout: 0.2).
- Output layer dimension depends on the time aggregation considered (168 for weekly hourly TS or 228 for the entire TS divided in 8 hrs ranges).

Training (batch size: 100, L1 Loss (MAE) and Adam optimizer): Variable learning rate and number of epochs depending on the category and time aggregation of the TS.

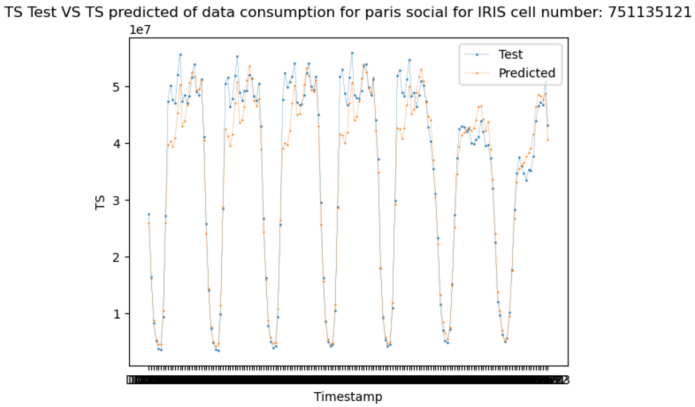
XGBoost Model

Ensemble method: Gradient-boosting with tree-based approach

- MSE as Loss function and evaluated with RMSE.
- Tuning of the parameters via a pairwise 3 fold CV (n_estimators: 2000, max_depth: 5, early stopping rounds = 7, other regularization parameters).

Training: Variable learning rate depending on the category and time aggregation.

Models Trained and Evaluated on Paris: Entire Series vs weekly



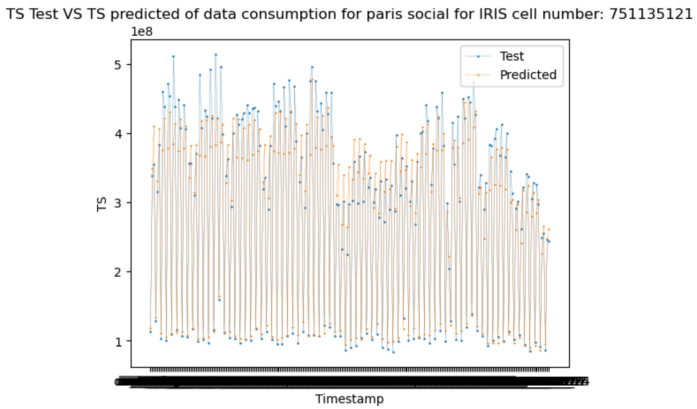
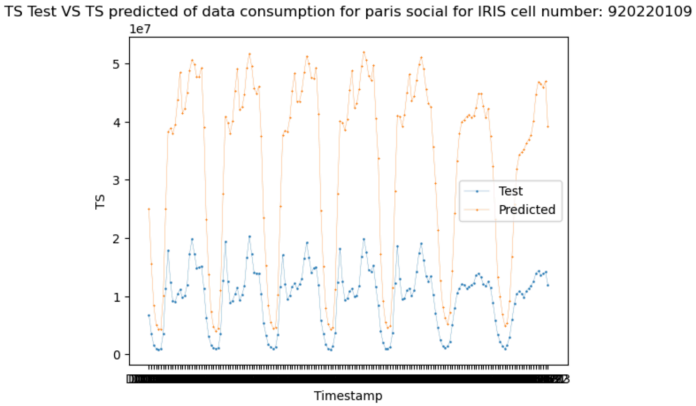


Figure: IRIS Cell: '751135121'. **Good prediction** on the Social time series by the NN model trained on Paris and evaluated on Paris - Weekly TS (left) Entire Series TS (right)



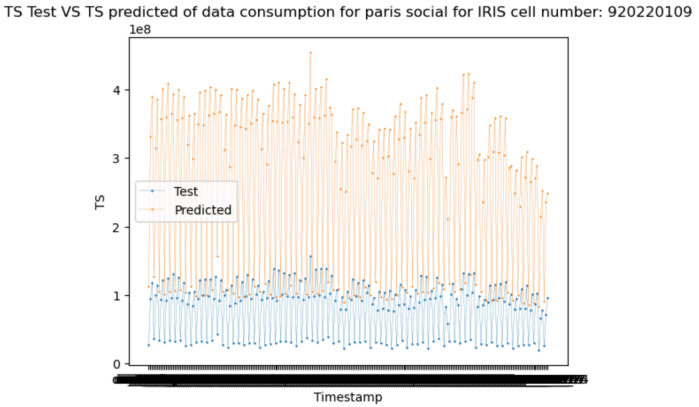


Figure: IRIS Cell: '920220109'. **Overprediction** on the Social time series by the NN model trained on Paris and evaluated on Paris - Weekly TS (left) Entire Series TS (right)

Models Trained and Evaluated on Paris: Social vs Gaming

Weekly TS predicted with NN

Number	Social	Gaming
Average TS	3.75×10^7	4.76×10^5
RMSE	1.86×10^7	8.30×10^5
MAPE	44.46%	88.23%

Weekly TS predicted with XGBOOST

Number	Social	Gaming
Average TS	3.75×10^7	4.76×10^5
RMSE	1.93×10^7	7.73×10^5
MAPE	48.14%	250.23%

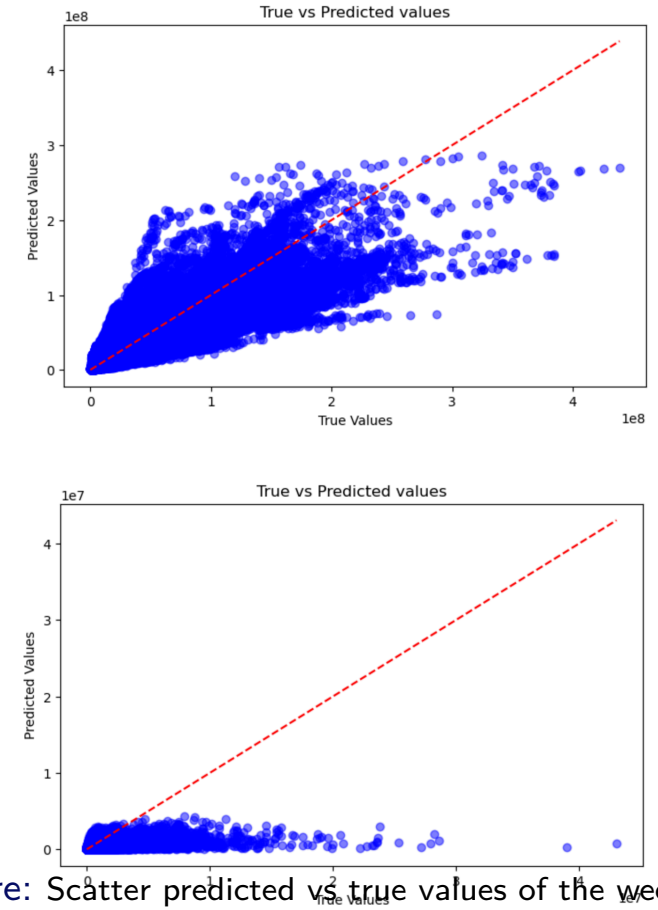


Figure: Scatter predicted vs true values of the weekly time series predicted with the NN model trained on Paris and evaluated on Paris - Social (top) Gaming (bottom)

Models Trained and Evaluated on Paris: Distribution of the Errors

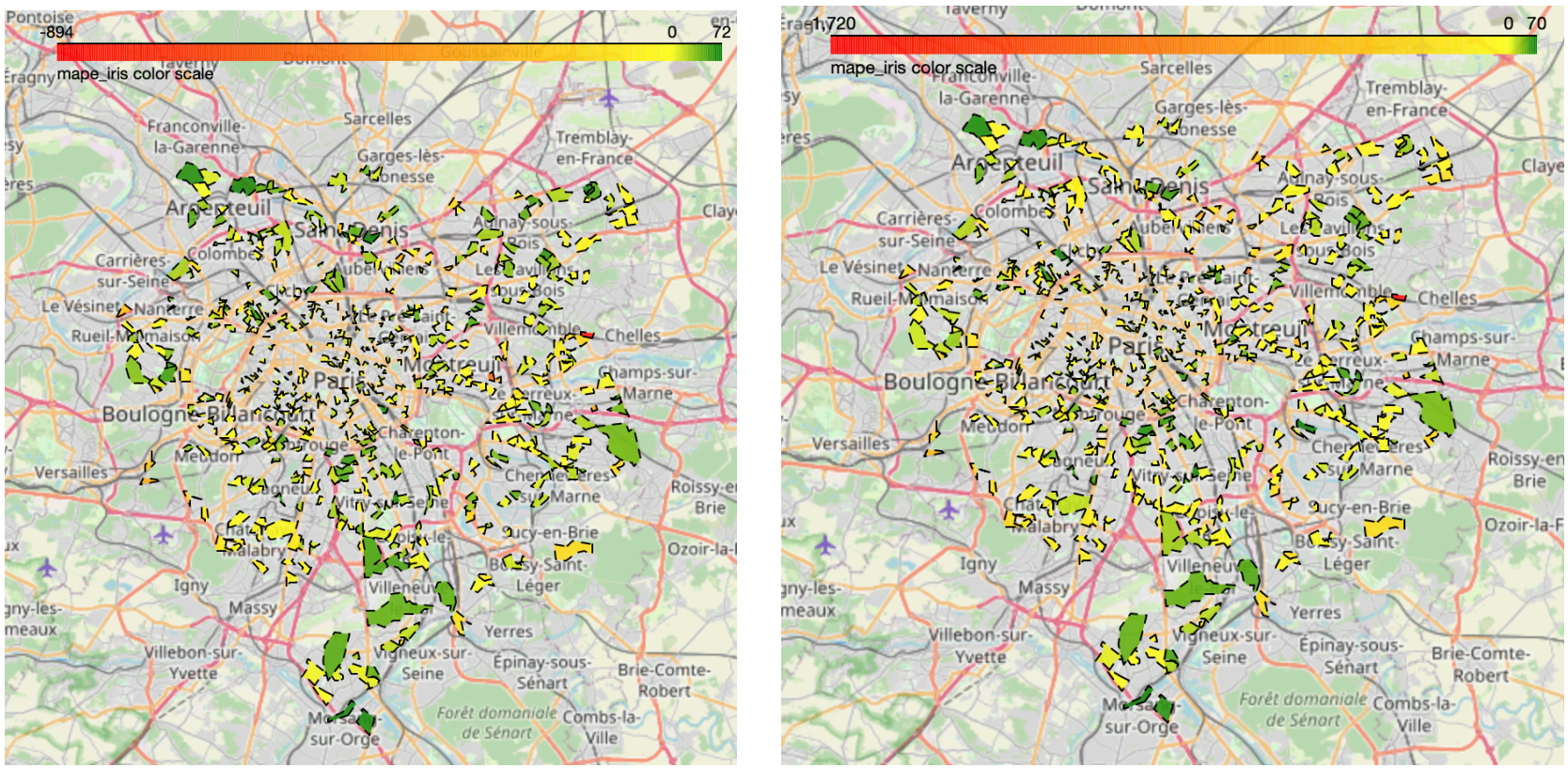
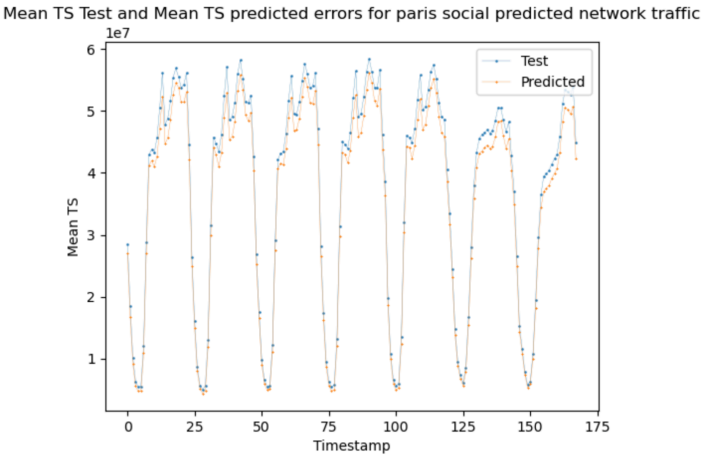


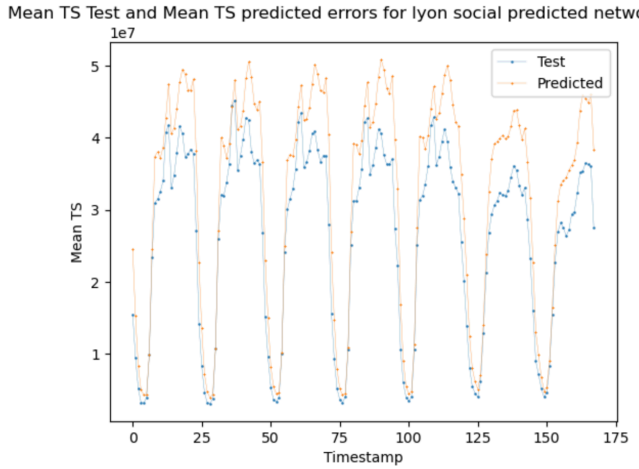
Figure: Distribution of the average MAPE for the weekly time series predicted with the NN model trained on Paris and evaluated on Paris for Social (left) and Gaming (right) - Over-predictions (red - orange), good predictions (yellow), under-predictions (green)

Results from the Feature Importance Methods

- The **Permutation Feature Importance method** was applied to the neural network models. **Shap feature importance was applied to Xgboost**. Both methods were applied to all categories and time aggregations.
- The **reduced models trained on the top 50 features** gave similar results when compared to the complete ones.
- **Important Features Selected**
  - **Area**: A smaller area reduces the traffic data, a larger area increases the traffic data.
  - **Distance from City Centre**: A smaller distance from the city centre increases the traffic data predicted, a larger distance from the city centre reduces the predicted traffic data.
  - **Urbanization and Electricity Consumption**: The higher the electricity consumption and urbanization, the higher the value of the traffic data.
  - **Presence of Attractions**: A greater number of attractions corresponds to a higher the value of the traffic data.

Paris-Trained Models Exhibit Upward Bias on Lyon and Marseille





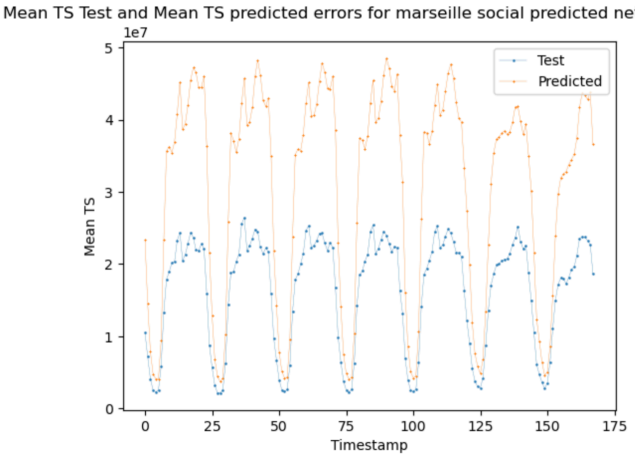


Figure: Weekly hourly TS average prediction across all IRIS cells vs true values of the average TS using the NN model trained on Paris and evaluated on: Paris (left), Lyon (right), Marseille (bottom)

Future developments and improvements

1. **Models trained in one city to predict TS in other cities**:
   - **Adding features related to the consumers preferences for different cities** on mobile traffic data (e.g from survey data)
   - **Multiplying the predictions for a constant proportional factor**: (e.g. the ratio between the average traffic data used in the city of training and test)
2. **Models trained in one part of a city to forecast traffic patterns in other areas within the same city**:
   - **Adding features related to the people dynamics within the different areas of the city**: car traffic or tourism presence in the different areas
   - **Improving the parameters fine tuning for each model configuration**
   - **Exploring alternative models** in particular the ones that takes into account of the Spatial Correlation in the data.

Bibliography

1. Molnar C.(2023). Interpretable Machine Learning, A Guide for Making Black Box Models Explainable
2. Shapley L. (1953) A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317.
3. Martínez-Durive O. E. , Mishra S., Ziemlicki C., Rubrichi S. , Smoreda Z. , and Fiore M. (17 Jul 2023). The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography, arXiv - CS - Networking and Internet Architecture.
4. Chen T., Guestrin C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. doi:10.1145/2939672.2939785
5. Goodfellow I., Bengio Y., Courville A. (2016). Deep Learning. MIT Press.
6. Kinsley H., Kukiela D. (2020). Neural Networks from Scratch in Python. Harrison Kinsley.
7. Štrumbelj E. , Kononenko. I.(2014). "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 : 647-665.