Forecasting COVID-19 Cases using Mobility and Socio-Economic Data

PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Mohamed Diagne, Mélody Allamélou, Aïda Mikou, Timothé Medjbeur

Deep Learning Course – Clément Gorin

Motivation

Predicting **COVID-19 cases** is critical for **public health responses**. Traditional **epidemiological models** provide theoretical insights but struggle to capture rapid changes in **mobility** and behavior. Leveraging real-world **mobility** and **socio-economic data** with **deep learning** offers a more flexible and reactive forecasting approach.

Dataset

- **77 departments** (excluding overseas territories)
- **658 consecutive days** (March 18, 2020 – December 31, 2021)
- **49,819 observations**
- **Target:** Daily confirmed positive **COVID-19 cases** per department

Table 1. Predictive Features Overview

Epidemiology	Mobility	Socio-Economic
COVID-19 test positivity rate (%)	Change in visits to retail/recreation locations (%)	Local unemployment rate (%)
Effective reproduction number	Change in visits to grocery/pharmacy stores (%)	Local poverty rate (%)
Intensive care unit occupancy rate (%)	Change in visits to parks (%)	Median household income (€)
	Change in usage of public transport stations (%)	Local population size
	Change in visits to workplaces (%)	
	Change in time spent at home (%)	

Sources: Mobility data from Google Mobility Reports, socio-economic data from INSEE, and COVID-19 data from DataGouv.

Methodology

Transmission Dynamics and Target Definition

Given the **incubation period**, symptom onset, testing, and **reporting delays**, an inherent lag exists between mobility changes and recorded COVID-19 infections.

To capture this dynamic, we created **lagged target variables**. The optimal lag corresponds to a **4-day shift** in reported positive cases, enabling the **prediction of infections** 5 days after a mobility change.

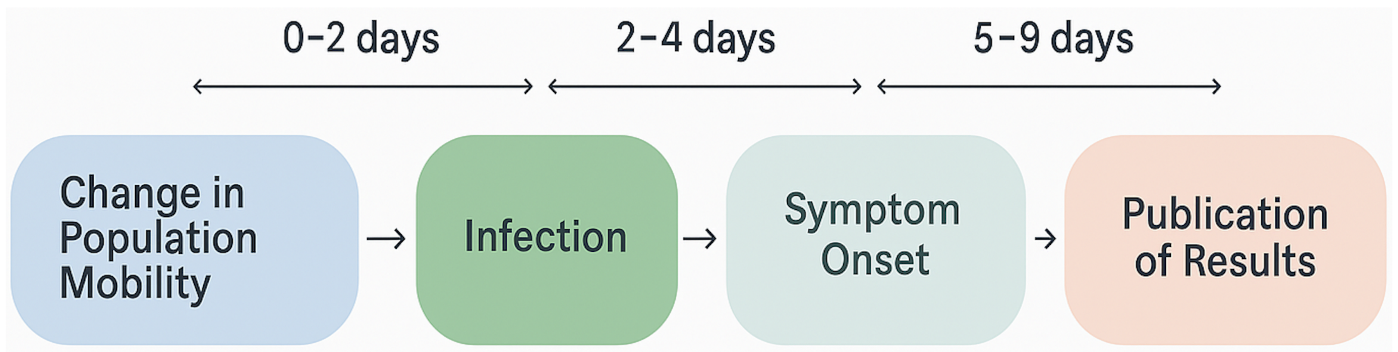
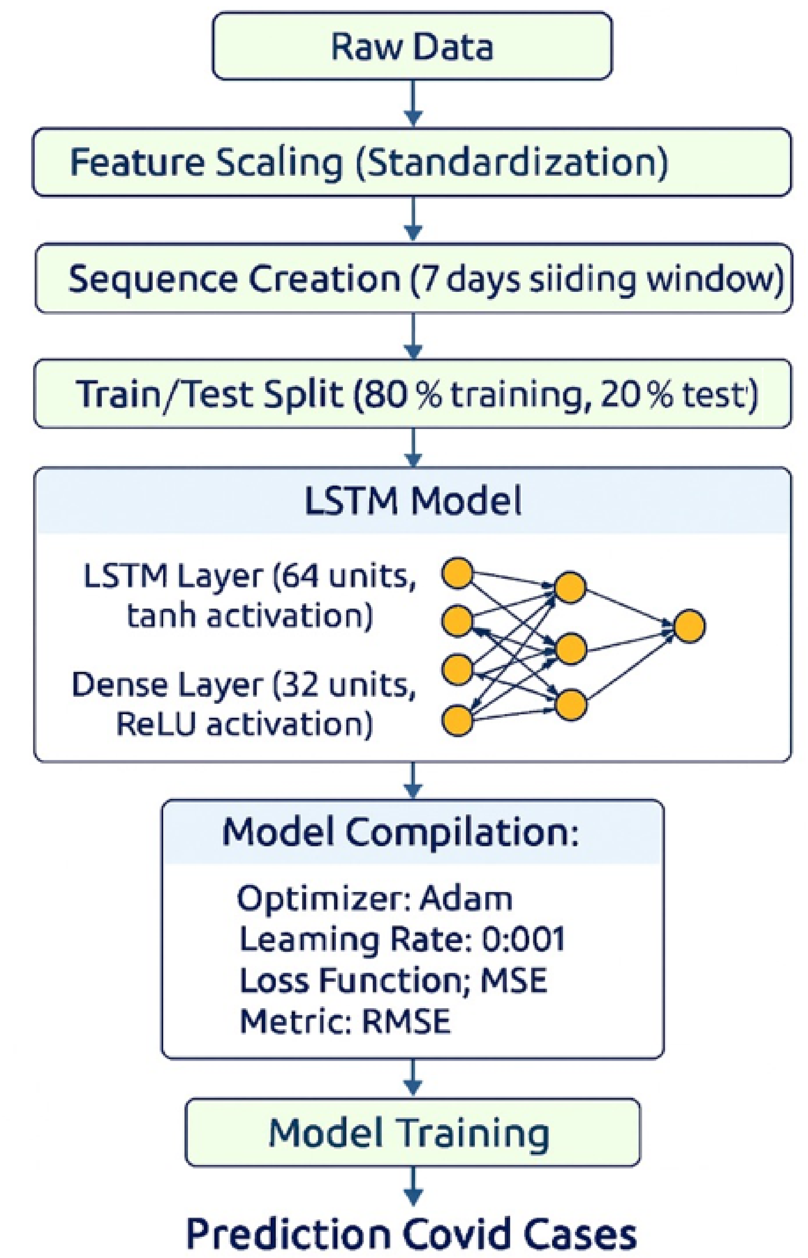


Fig1. Transmission Cycle and Reporting Lag

Model Pipeline and Training

- **LSTM Model:** Captures temporal dependencies in sequential data, making it ideal for forecasting COVID-19 cases.
- **Performance Metric:** Root Mean Squared Error (RMSE), which penalizes large errors, ensuring high prediction accuracy, critical for health forecasts.



Results

Results Summary

Table 2. Model Performance Metrics

Data	Pearson's Correlation	RMSE
Training Set	0.95	144.46
Testing Set	0.92	307.94

- **High Pearson's correlation** between predicted and observed cases (0.95 training, 0.92 testing).
- **RMSE** increased from 144.46 (training) to 307.94 (testing).
- **Model captures trend well** despite moderate accuracy loss on unseen data.

Spatial Error Analysis

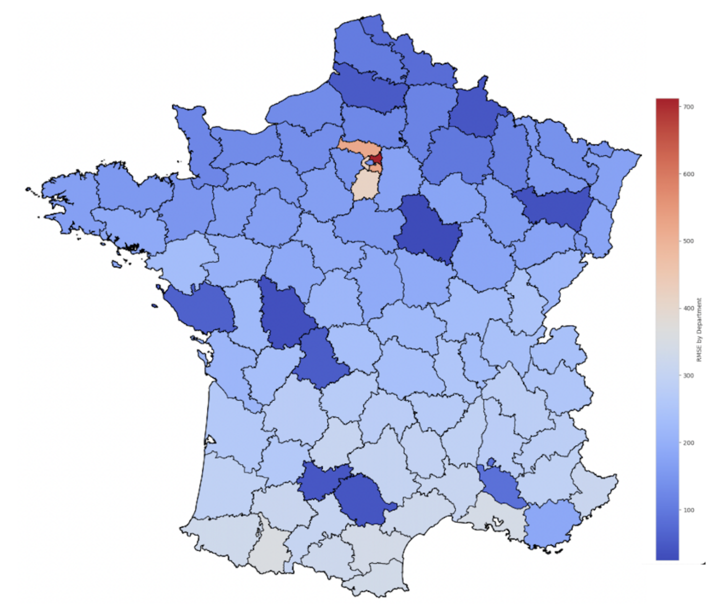


Fig2. RMSE Spatial Distribution

$$I = \frac{N}{W} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Formula: Moran's I Index

- **Higher RMSE values** were observed in **Île-de-France**, likely due to higher population density, mobility complexity, and regional transmission dynamics.
- **Moran's I = 0.42**, statistically significant, indicates **moderate spatial clustering** of errors, suggesting neighboring departments share unmodeled local factors.

Interpretability

We applied **Integrated Gradients** to identify the most influential features across time steps in COVID-19 case prediction.

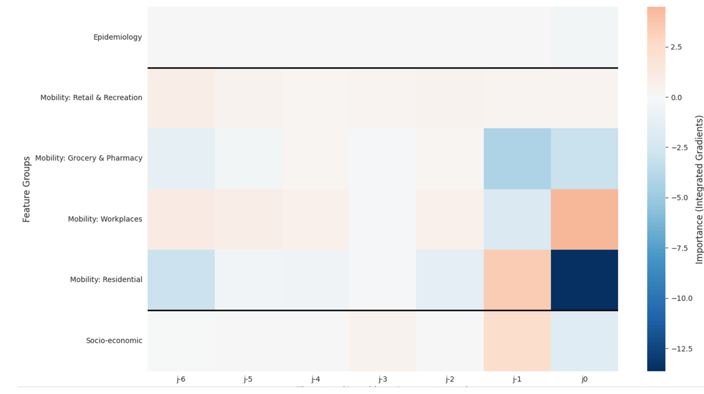


Fig 3. Feature Interpretability Analysis

- Recent mobility changes (Workplaces, Residential) drive COVID-19 case predictions, with negative attributions linking mobility reductions to fewer cases.
- Epidemiological and socio-economic features have limited immediate impact.
- Influence peaks at J-1 and J0, matching expected transmission dynamics.

Conclusion

- **LSTM-based model** successfully predicted COVID-19 cases by capturing temporal dynamics and mobility-driven patterns.
- **Interpretability analysis** revealed that recent mobility changes were the main drivers of prediction, aligning with transmission delays.

Future Work

- **Extend the analysis** to include COVID-19 dynamics from 2022 to capture later pandemic phases.
- **Develop spatially adaptive models** to better handle regional variations in prediction errors.
- **Integrate additional features** such as vaccination rates and public health policy indicators.

To consult the bibliography and code, click on the Google Colab link below.

https://colab.research.google.com/drive/1uh4QEaH0Vao8ZVPLoiz8k4MMC5sGXLql?usp=sharing