Correlations between epidemiological time series forecasting and influence regions of Brazilian cities

Fernando Henrique Oliveira Duarte[1], Gladston J. P. Moreira[1], Eduardo J. S. Luz[1], Leonardo B. L. Santos[2], Vander L. S. Freitas[1]

[1]Department of Computing – Federal University of Ouro Preto (UFOP)
CEP 35400-000 – Ouro Preto – MG – Brazil

fernando.hod@aluno.ufop.edu.br, {gladston,eduluz,vander.freitas}@ufop.edu.br

[2]National Center for Monitoring and Alerts of Natural Disaster (Cemaden)
CEP 12247-016 – Sao Jose Dos Campos – SP – Brazil

leonardo.santos@cemaden.gov.br

Abstract. *The study investigates the correlation between mobility network centralities, demographic features, and RMSE in COVID-19 prediction models (Graph Convolution Networks - GCN, Prophet, and Long Short-Term Memory - LSTM) across Brazilian municipalities. The analysis reveals that betweenness centrality, Degree, Strength, and Municipal Population exhibit positive correlations with RMSE, indicating that municipalities with central positioning, numerous connections, high neighbor flow, and larger populations negatively influence the predictions.*

1. Introduction

Predicting patterns that evolve is a popular area of investigation in data analytics for forecasting future trends and behaviors. Various approaches, including machine learning models, are commonly used to capture the complexity of the series and generate reliable estimates [Smith et al. 2004, Vaishya et al. 2020].

Mobility networks offer a substantial data source for analyzing flow dynamics in complex systems [Albert and Barabási 2002]. This can be exemplified by nodes that represent specific locations connected by edges, possessing weights that determine the movement of individuals between locations within a given time frame [Fanelli and Piazza 2020, Freitas et al. 2020a, Freitas et al. 2020b, Rothan and Byrareddy 2020].

By combining temporal pattern predictions with mobility networks, the temporal and spatial dynamics of events can be objectively analyzed. In this context, Graph Convolutional Networks (GCNs), a machine learning algorithm specifically developed for graphs, facilitate the inclusion of connections between elements to build a complex network. Models such as the Graph Convolutional Long Short-Term Memory (GCLSTM) [Chen et al. 2022] and the Graph Convolutional Recurrent Network (GCRN) [Seo et al. 2018] have recently been utilized for forecasting COVID-19 case time series in Brazil, as described in [Duarte et al. 2023]. They mix GCNs with Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) layers and will be referred to as GCN-based models here.

This study builds upon the foundational work presented in [Duarte et al. 2023] by delving into the intricate relationships between mobility network centrality metrics, demographic and socioeconomic indicators, epidemiological variables, and the prediction errors of COVID-19 time series. In our prior investigation [Duarte et al. 2023], a diverse array of predictive models, including LSTM, Prophet, GCLSTM, and GCRN, were employed. Particularly noteworthy were the outstanding R^2 scores achieved by the GCN-based and Prophet models, surpassing 0.97. The Prophet model, in particular, emerged as the leading performer, attaining a remarkable mean RMSE of 1758.21 with a standard deviation of 430.81. Following closely, GCRN exhibited the second-best performance with a mean RMSE of 2990.40 and a standard deviation of 1035.11, while GCLSTM secured the third position with a mean RMSE of 3535.38 and a standard deviation of 1221.01. In contrast, the LSTM model ranked last, displaying a mean RMSE of 4298.89 and a standard deviation of 1670.56.

2. Methodology

2.1. Data Sources

To depict the spread of COVID-19 in Brazil, we examined its temporal and spatial dimensions. Temporally, we calculated the “Avg Daily Cases”, representing the mean number of daily COVID-19 cases, and “Reported Days”, indicating the number of days COVID-19 cases were reported for each municipality, using the publicly available dataset of COVID-19 daily cases provided by [Cota 2020]. This dataset covers the period from February 2020 - when the epidemic began in Brazil - to November 2022, totaling 1009 consecutive days. It gathers official Ministry of Health data collections, with updates provided asynchronously.

Concerning the spatial dimension, we use the origin-destination survey for “Road and Waterway connections” [IBGE 2017]. In this network, each city represents a node and their weighted connections account for the weekly flow of vehicles between them. The resulting network has $N = 5385$ nodes and $L = 65639$ edges.

The 2022 Brazilian census provides the variable POPMUN, which indicates the population size of municipalities and enables demographic analysis. According to the “Regions of Influence of Cities 2018” (REGIC 2018) survey, documented in [IBGE 2020], VAR03 reflects the Gross Domestic Product (GDP) of each municipality, serving as an economic activity measure. Next, the Territory Management Centrality Score (VAR19) provides insights into the effectiveness of municipal governance through both public and private management centrality indices. Additionally, the General Attraction Score (VAR56) measures the overall attractiveness of municipalities in terms of their ability to attract people and resources. VAR79, the Quantity of Commercial Categories, indicates the range of available services in each municipality, which is often associated with the diversity of commerce. These variables collectively provide significant insights into the distinctive features of Brazilian municipalities.

2.2. Network Metrics

The analysis of mobility networks’ structure and dynamics requires the utilization of network metrics such as Degree, Betweenness, Strength, and Closeness. Since the weights of the mobility network signify the flows of vehicles, the computation of shortest paths for Betweenness and Closeness relies on distances. Therefore, we used the inverse of the flow, whereby larger flows correspond to shorter distances. We used the demographic and flow data presented in Section 2.1 to calculate those metrics.

2.3. Time Series Prediction Models

In [Duarte et al. 2023], we presented two models based on GCNs, the GCRN and GCLSTM, that incorporate a mobility network to forecast COVID-19 cases in Brazil. The network serves as an approximation of the disease path, as shown in [Freitas et al. 2020b, Freitas et al. 2020a]. The models utilize convolutions to capture the interconnections between neighboring municipalities in the graph for making predictions on temporal data. For comparison purposes, we implemented Prophet [Taylor and Letham 2018] and LSTM (Long Short-Term Memory) [Hochreiter and Schmidhuber 1997] models, that do not make use of mobility data.

In contrast, the Prophet [Taylor and Letham 2018] and LSTM [Hochreiter and Schmidhuber 1997] models are solely temporal. LSTM is a type of RNN, a deep learning model characterized by its ability to handle data sequences such as time series. Prophet is an additive regression model extensively employed in time series analysis and data forecasting, recognized for its versatility and effectiveness [Hastie 2017]. Both models can capture complex temporal features appropriate for forecasting series with startling changes, trends, and seasonal variance.

The analysis presented in [Duarte et al. 2023] suggests that the Prophet model has high accuracy in prediction, with exceptional performance in certain regions but not as impressive in others, presenting a large standard deviation. Conversely, the LSTM model exhibits the lowest accuracy levels. The two GCN-based models demonstrate similar performances, with a performance between the Prophet and LSTM models.

2.4. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a commonly used metric to evaluate the performance of prediction models. It is calculated by taking the square root of the average of the squared differences between the predicted value $\hat{y}$ and the actual value y:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \tag{1}$$

where n is the number of data points. The RMSE quantifies the prediction power of the model, with lower values indicating better performance.

3. Results and Discussion

Figure 1 depicts the logarithmic-scale RMSE values for LSTM model predictions across Brazilian municipalities. The displayed map reveals a similar pattern in RMSE distribution among all models. Despite the expectation of identifying a discernible pattern associated with the spread of COVID-19, such a trend proved elusive in the observed data.

Figure 2 illustrates correlation coefficients between RMSE and other variables. Non-significant correlations (p-value > 0.05) are excluded. The results highlight a robust correlation among the RMSE of all models.

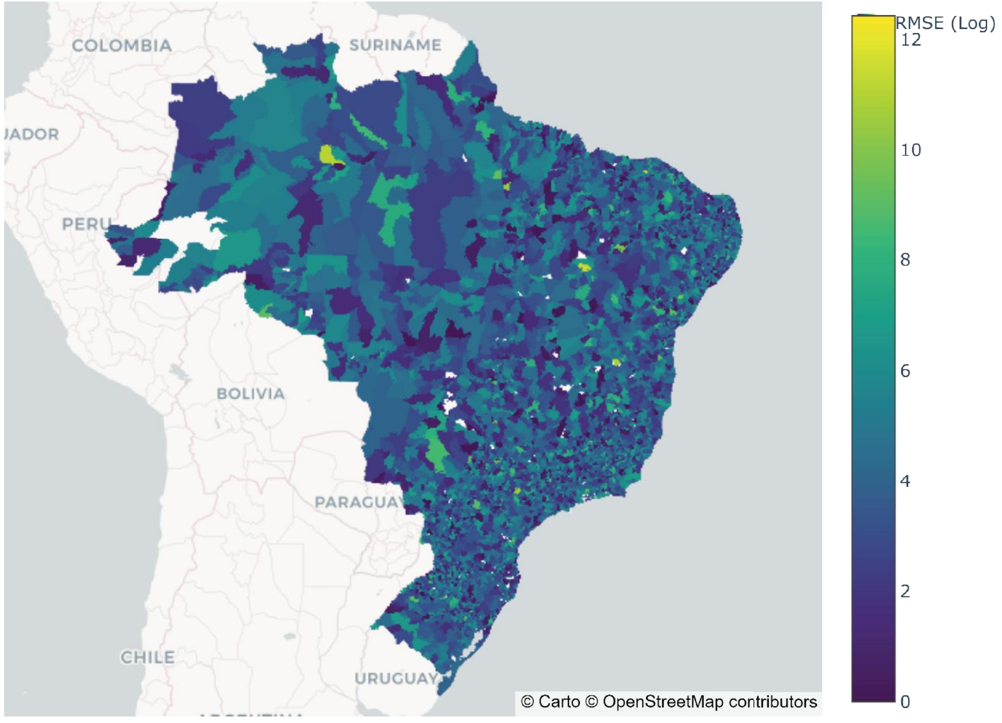
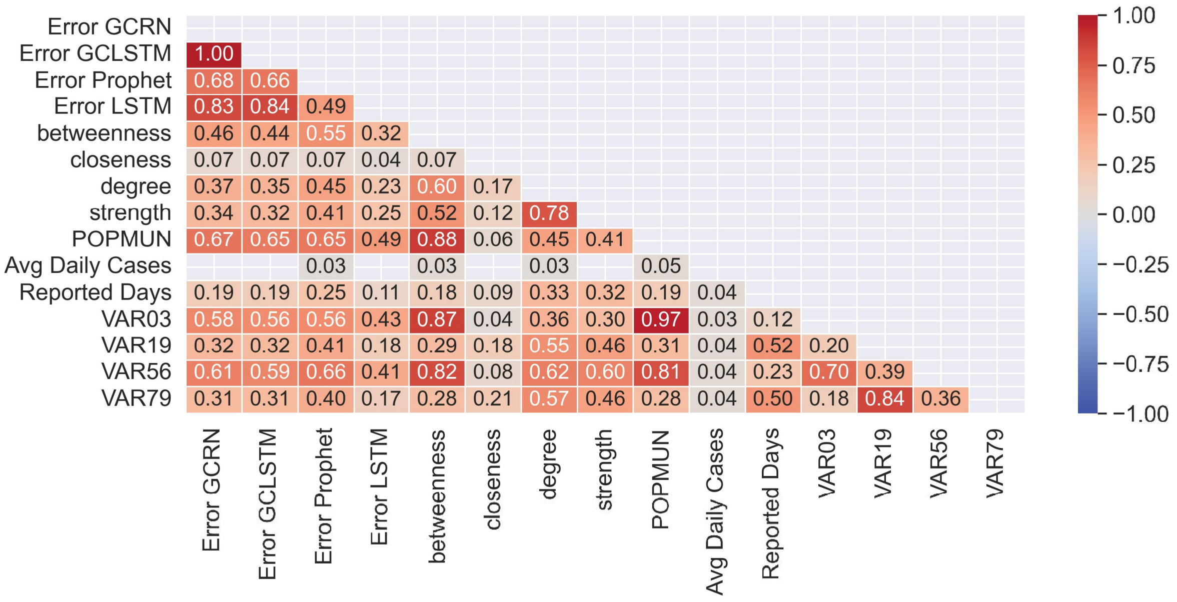


Figure 1. RMSE for COVID-19 predictions across Brazilian municipalities for the LSTM model, depicted on a logarithmic scale.



	Error GCRN	Error GCLSTM	Error Prophet	Error LSTM	betweenness	closeness	degree	strength	POPMUN	Avg Daily Cases	Reported Days	VAR03	VAR19	VAR56	VAR79
Error GCRN															
Error GCLSTM	1.00														
Error Prophet	0.68	0.66													
Error LSTM	0.83	0.84	0.49												
betweenness	0.46	0.44	0.55	0.32											
closeness	0.07	0.07	0.07	0.04	0.07										
degree	0.37	0.35	0.45	0.23	0.60	0.17									
strength	0.34	0.32	0.41	0.25	0.52	0.12	0.78								
POPMUN	0.67	0.65	0.65	0.49	0.88	0.06	0.45	0.41							
Avg Daily Cases			0.03		0.03		0.03		0.05						
Reported Days	0.19	0.19	0.25	0.11	0.18	0.09	0.33	0.32	0.19	0.04					
VAR03	0.58	0.56	0.56	0.43	0.87	0.04	0.36	0.30	0.97	0.03	0.12				
VAR19	0.32	0.32	0.41	0.18	0.29	0.18	0.55	0.46	0.31	0.04	0.52	0.20			
VAR56	0.61	0.59	0.66	0.41	0.82	0.08	0.62	0.60	0.81	0.04	0.23	0.70	0.39		
VAR79	0.31	0.31	0.40	0.17	0.28	0.21	0.57	0.46	0.28	0.04	0.50	0.18	0.84	0.36	

Figure 2. Significant Pearson Correlations (p-value $<$ 0.05) in Brazil.

The Betweenness centrality, along with Degree, Strength, and POPMUN, exhibits a positive correlation with RMSE in prediction models. This implies that centrally located municipalities with numerous connections, high flow between neighbors, and larger populations may experience less accurate predictions.

The variables VAR03 and VAR56 show a strong positive correlation with metrics POPMUN and Betweenness, and a moderate correlation with Degree, Strength and RMSE. Variables VAR19 and VAR79 display a high positive correlation with Degree and Strength, and a lower correlation with POPMUN, Betweenness and RMSE.

Based on the analyzed correlations, we observe that cities characterized by higher population (POPMUN), a significant number of connections (Degree), substantial flow in their connections (Strength), playing a central role or hub in the network (Betweenness), and a more pronounced economic development (VAR03 and VAR56) exhibit higher RMSE values in prediction models. This trend suggests that, potentially, the complexity and dynamics of these municipalities, marked by a combination of socio-economic factors and connectivity, may render less precise predictions. Our hypothesis is that the heterogeneity of these areas, marked by higher population density, a more intricate network of connections, and a more robust economy, could potentially lead to increased noise or disturbances in predictions, especially in locations that are more frequented and densely populated, interpreted as areas of potential aggregation.

4. Conclusions and future work

In conclusion, the analysis reveals correlations among economic indicators (VAR03, VAR19, VAR56, VAR79) and their positive association with centrality metrics. The centrality metrics (Betweenness, Degree, Strength) and POPMUN exhibit positive correlations with RMSE in prediction models, emphasizing their influence on prediction accuracy. Notably, the strong correlation between robust economic indicators and prediction errors suggests that highly developed locales may potentially lead to an unpredictable outcome, causing disturbances in the accuracy of prediction models. This hypothetical interpretation aligns with the notion that areas with higher population density or greater connectivity, whether in terms of quantity or flow, may introduce noise and disturbances, impacting the precision of prediction errors.

For future work, a more in-depth exploration of the intricate relationships between demographic and economic data and the RMSE obtained from forecasting models is warranted, with a focus on elucidating trends, seasonal patterns, and characteristics at macro and micro levels. This entails investigating variations among different regions, including states, capital cities, commercial zones, and others. Such an endeavor would contribute to a more comprehensive understanding of the underlying factors impacting predictive accuracy, thereby providing valuable insights for tailored and context-specific modeling and public health strategies.

Acknowledgements

The authors thank the CNPq, grants 441016/2020-0, 307151/2022-0, 308400/2022-4, FAPEMIG, grants APQ-01518-21, APQ-01647-22, CAPES, grant 88887.506931/2020-00, and Universidade Federal de Ouro Preto (UFOP).

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Chen, J., Wang, X., and Xu, X. (2022). Gc-lstm: graph convolution embedded lstm for dynamic network link prediction. *Applied Intelligence*, 52(7):7513–7528.

Cota, W. (2020). Monitoring the number of covid-19 cases and deaths in brazil at municipal and federative units level. *SciELO Preprints*.

Duarte, F. H. O., Moreira, G. J. P., Luz, E. J. S., Santos, L. B. L., and Freitas, V. L. S. (2023). Time series forecasting of covid-19 cases in brazil with gnn and mobility networks. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 361–375, Cham. Springer Nature Switzerland.

Fanelli, D. and Piazza, F. (2020). Analysis and forecast of covid-19 spreading in china, italy and france. *Chaos, Solitons & Fractals*, 134:109761.

Freitas, V. L., Moreira, G. J., and Santos, L. B. (2020a). Robustness analysis in an intercities mobility network: modeling municipal, state and federal initiatives as failures and attacks toward sars-cov-2 containment. *PeerJ*, 8:e10287.

Freitas, V. L. d. S., Konstantyner, T. C. R. d. O., Mendes, J. F., Sepetauskas, C. S. d. N., and Santos, L. B. L. (2020b). The correspondence between the structure of the terrestrial mobility network and the spreading of covid-19 in brazil. *Cadernos de Saúde Pública*, 36:e00184820.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

IBGE (2017). *Ligações rodoviárias e hidroviárias: 2016*. IBGE, Coordenação de Geografia Rio de Janeiro, Brazil.

IBGE (2020). *Regiões de influência das cidades : 2018*. IBGE, Coordenação de Geografia Rio de Janeiro, Brazil.

Rothan, H. A. and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of autoimmunity*, 109:102433.

Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. (2018). Structured sequence modeling with graph convolutional recurrent networks. In *Neural Information Processing*, pages 362–373.

Smith, D., Moore, L., et al. (2004). The sir model for spread of disease-the differential equation model. *Convergence*.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Vaishya, R., Javaid, M., Khan, I. H., and Haleem, A. (2020). Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339.