Poster: COVID-19 Case Prediction using Cellular Network Traffic

Necati Ayan[1], Sushil Chaskar[1], Anand Seetharam[1], Arti Ramesh[1], Antonio A. de A. Rocha[2]
[1]Computer Science Department, SUNY Binghamton
[2]Institute of Computing, Fluminense Federal University
(nayan1, schaska1, aseethar, artir)@binghamton.edu, arocha@ic.uff.br

***Abstract*—In this paper, our goal is to leverage cellular network traffic data to model and forecast the number of COVID-19 infections in the future. To this end, we partner with one of the main cellular network providers in Brazil, TIM Brazil, and collect and analyze cellular network connections from 973 antennas for all users in the city of Rio de Janeiro and its suburbs. We develop a Markovian model that captures the mobility of individuals across municipalities of the city. The transition probabilities of the Markov chain are determined by analyzing user-level mobility events between antennas from the cellular network connectivity logs. We combine the aggregate mobility characteristics across municipalities as evidenced from the transition probabilities with the number of reported COVID-19 cases in a municipality during a particular week to design mobility-aware COVID-19 case prediction models that predict the number of cases for the following week. Our experiments demonstrate that our mobility-aware models significantly outperform a baseline mobility-agnostic linear regression model in terms of metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).**

I. Introduction

COVID-19 is a global pandemic that has infected human beings in all countries of the world. To help governments combat the pandemic, it is necessary to design mobility-aware case prediction models that accurately predict the number of future infections and assist officials in understanding the connection between human mobility and rising infection rates. These prediction models can also enable officials to design and implement local lockdown measures instead of widely unpopular blanket lockdown measures to contain the spread of COVID-19.

Therefore, in this paper, our goal is to leverage cellular network connectivity data to develop a simple but efficient approach for determining the relationship between human mobility and infection rates in municipalities within a city. To this end, we partner with TIM Brazil, one of the largest cellular network providers in Brazil to collect anonymized cellular network connection logs (i.e., 3G/4G connections, text messages, calls) for all users in the city of Rio de Janeiro. The data consists of individual connections made by users to 973 cellular antennas in and around Rio de Janeiro and its suburbs at 5-minute intervals from April 2020 to July 2020. We also use publicly accessible COVID-19 infection data for Rio de Janeiro's various municipal administrative regions (a total of 27). By analyzing cellular network connections, our methodology uses a data-driven approach to investigate and model the mobility of individuals in a city, and then uses the designed mobility model to forecast the number of COVID-19 infections in the future.

We first identify mobility events (i.e., user movements from one antenna to another) from the connectivity logs. We then use these mobility events to develop a Markovian model that accurately captures the movement of individuals across municipalities in the city. We determine the transition probabilities of the Markov model for each week by considering the mobility events for that week. We design mobility-aware COVID-19 case prediction models by effectively combining the transition probabilities encoding the mobility between source and destination regions with the corresponding number of infections in the source to predict the number of infections in the destination regions for the next week. We observe that our models significantly outperform a baseline mobility-agnostic linear regression model in terms of metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error (RE) when predicting the number of future cases.

Related Work: We discuss some pertinent related work before presenting our research. To understand the relationship between mobility and the spread of COVID-19, Gao et. al map county-level mobility pattern changes in the US in response to COVID-19 [1]. Similarly, Huang et. al analyze the impact of the COVID-19 pandemic on transportation-related behaviors using human mobility data [2]. Human mobility modeling has received significant attention in the last decade, and few recent examples include modeling semantic-rich human mobility using hidden Markov models [3] and analyzing user mobility in cellular networks [4].

II. Data and Methods

In this section, we present an overview of two datasets we analyze in this study. We work with TIM Brazil, one of Brazil's largest cellular network providers, to collect cellular network connection logs for all users in the city of Rio de Janeiro. Additionally, we use publicly accessible COVID-19 infection data for Rio's various municipalities. Our goal is to use the cellular network connectivity dataset to first understand the aggregate mobility of people during COVID-19, and then to leverage the two datasets to design mobility-aware COVID-19 prediction models.

Annex to ISBN 978-3-903176-39-3

A. Cellular Network Connectivity Data

This dataset consists of approximately 10 billion anonymized cellular network logs (i.e., phone calls, text messages, 3G/4G data connections) of users along with the information of the specific antenna (there are a total of 973 antennas in our dataset) through which the connections are established from April 5^{th} to July 2^{nd}. A couple of example entries in the dataset is shown in Table I. We first identify mobility events from the network logs. If a user moves from one antenna to another antenna with different timestamps, we consider this to be a mobility event.

TABLE I: Cellular Network Connectivity Dataset

Timestamp	User ID	Latitude	Longitude
timestamp-1	hash-1	-23.003431	-43.342206
timestamp-2	hash-2	-22.8415	-43.278389

Figure 1 depicts the change in the total number of connections and the number of mobility events over weeks. The number of connections (i.e., the blue line), is shown on the left y-axis and the number of mobility events (i.e., the red line) is shown on the right y-axis. The vertical dotted line in the figure is the day when Brazil eased its lockdown (June 1), which corresponds to the beginning of week 9 in our analysis. Even before the lockdown restrictions are lifted, we note a significant rise in the number of connections and mobility events per week.

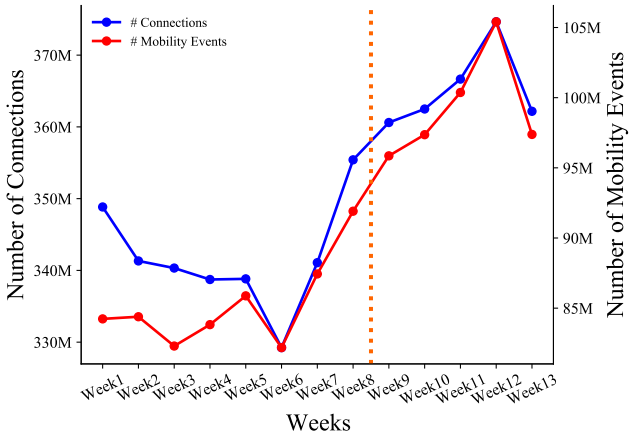

Fig. 1: The number of connections and mobility events per week from April 5^{th}, 2020 to July 2^{nd}, 2020

B. COVID-19 Case Data

Our second dataset consists of the daily number of positive COVID-19 cases for each municipality in Rio. We show some example instances from this dataset in Table II. Each row in the table represents a single COVID-19 positive case. Though we have COVID-19 case data from March, we focus on data from April 5^{th} to July 2^{nd} because it corresponds to our cellular network connections data.

TABLE II: COVID-19 Case Dataset

Timestamp	Latitude	Longitude	Region Code
timestamp-1	-22.888272	-43.552508	144
timestamp-2	-22.898441	-43.223156	10

III. Mobility-Aware COVID-19 Case Prediction Model

Our objective in this paper is to develop mobility-aware COVID-19 case prediction models in order to forecast the number of future infections and to better understand the connection between human mobility and the spread of the disease. To achieve this, we first use cellular network connection logs to model aggregate human mobility patterns across Rio's municipalities, and then combine this aggregate mobility information with the case data from the various regions to make informed future predictions. We first describe our Markovian models for modeling the mobility across municipalities and then discuss our mobility-aware COVID-19 case prediction models.

A. Markov Models for Human Mobility

We construct a Markov model at the municipality level to elegantly capture and model aggregate human mobility patterns in Rio. Each state in our Markov model corresponds to a municipality and transitions between states encode movement between municipalities. We first peruse the cellular network connectivity logs to identify mobility events between the different antennas for all the users to determine the transition matrix of the Markov model. As one municipality can have multiple antennas, a mobility event between two antennas in the same municipality corresponds to a same state transition. In comparison, mobility events between two antennas in different municipalities lead to transitions between the corresponding states of the Markov chain. As our aim is to predict the number of COVID-19 cases for the coming week using data from the previous week, we calculate the Markovian model's transition probabilities on a weekly basis.

B. COVID-19 Case Prediction Model

In this subsection, we discuss how we combine the mobility model with the current active COVID-19 cases to predict the number of future cases. We design two versions of our mobility-aware prediction model — *i)* a linear model, and *ii)* a polynomial model.

1) Linear Model: Our linear mobility-aware prediction model determines the number of COVID-19 cases for the next week in a municipality (i.e., destination) by considering the linear weighted sum of the current COVID-19 infections in the different municipalities (sources) multiplied by the one step transition probability from the different sources to the destination (Eqn. (1)).

$$c_j(t+1) = m_{0j} + m_{1j}(\sum_i p_{ij}(t)c_i(t)) \quad (1)$$

where $c_j(t+1)$ denotes the number of cases in week $(t+1)$ in state j, $c_i(t)$ denotes the cases in week t in state i, $p_{ij}(t)$ is the probability of moving from state i to state j (1-step

transition probability from the Markov chain). m_{0j} and m_{1j} are the intercept and slope of the line.

2) Polynomial Model: In addition to the linear model, we also design higher-order polynomial models to better capture nuances in the underlying data. Unlike the linear model, the higher-order polynomial model in Eqn (2) fits the best curve.

$$c_j(t+1) = m_{0j} + m_{1j}(\sum_i p_{ij}(t)c_i(t)) + m_{2j}(\sum_i p_{ij}(t)c_i(t))^2 + ... + m_{nj}(\sum_i p_{ij}(t)c_i(t))^n \quad (2)$$

where $m = (m_{0j}, m_{1j}, ..., m_{nj})$ are the coefficients of the polynomial terms. We experiment with primarily polynomials of orders 2 and 3 to keep the number of parameters to a minimum and to avoid overfitting the model to the data. The logic behind this approach is to account for the number of active infections at a source and then use the mobility metric (i.e., transition probability from source to destination) as a measure of infection spread from source to destination.

IV. Experiments

In this section, we present experimental results to demonstrate the superior prediction performance of our mobility-aware COVID-19 case prediction models when compared to a baseline mobility-agnostic linear prediction model. The baseline linear regression model considers only the past COVID-19 cases in a region and produces the best fit straight line for the data. We evaluate the models with respect to 3 different error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Relative Error (RE).

$$\text{MAE}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1}|y_i - \hat{y}_i| \quad (3)$$

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2} \quad (4)$$

$$\text{RE}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1}\frac{|y_i - \hat{y}_i|}{y_i} \quad (5)$$

where y_i and $\hat{y}_i$ are the i^{th} actual and predicted values, and N denotes the number of samples (in our case there are 12 samples, one corresponding to each week).

Figure 2 shows the prediction performance for one of Rio's municipalities, Maduereira (MA). The green and black lines correspond to the mobility-aware linear and polynomial (order 3) models. The mobility-agnostic linear regression fits a straight line (i.e., blue line) with respect to the actual case numbers (i.e., red points). As our mobility-aware linear and polynomial models fit linear and higher order polynomials between past and future cases while taking the one-step transition probabilities into account (Eqns 1 and 2), we observe qualitatively from Figure 2 that our mobility-aware models provide better performance than the baseline mobility-agnostic model. To support this finding, we present the RMSE, MAE, and RE for two regions, Maduereira (MA) and Meier (ME), in Table III. We observe from the table that the mobility-aware models overall outperform the baseline for all the error metrics. As expected, higher-order polynomial models have better prediction performance because they can better model the underlying patterns in the data.

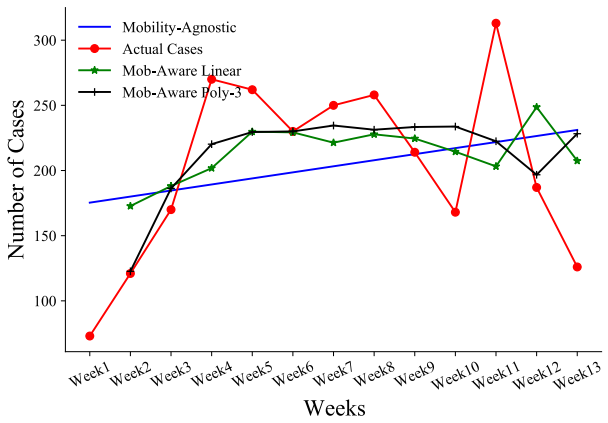

Fig. 2: COVID-19 Case Prediction Performance

TABLE III: Error rates region PA and CE

Regions	Models	Errors		
		MAE	RMSE	RE
MA	Mobility-Agnostic	53.094	60.463	0.273
	Mob-Aware Linear	44.968	54.146	0.233
	Mob-Aware Poly-2	35.755	48.488	0.160
	Mob-Aware Poly-3	35.832	48.472	0.159
ME	Mobility-Agnostic	90.728	104.250	0.294
	Mob-Aware Linear	76.774	95.450	0.246
	Mob-Aware Poly-2	64.302	90.427	0.185
	Mob-Aware Poly-3	65.776	90.159	0.187

V. Conclusion

In this paper, we designed a mobility-aware COVID-19 case prediction model that predicts the number of future infections. Via experiments on large scale real-world cellular network traffic data from Rio, we demonstrated that our models outperformed a baseline mobility-agnostic model. Our method can be easily extended to other cities, states, and countries around the world and can help government officials better understand the spread of the disease and enact targeted local lockdowns instead of widely unpopular blanket lockdowns.

References

[1] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. Mapping county-level mobility pattern changes in the united states in response to covid-19. *SIGSPATIAL Special*, 12(1):16–26, 2020.

[2] Jizhou Huang, Haifeng Wang, Miao Fan, An Zhuo, Yibo Sun, and Ying Li. Understanding the impact of the covid-19 pandemic on transportation-related behaviors with human mobility data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3443–3450, 2020.

[3] Wanzheng Zhu, Chao Zhang, Shuochao Yao, Xiaobin Gao, and Jiawei Han. A spherical hidden Markov model for semantics-rich human mobility modeling. In *AAAI Conference on Artificial Intelligence*, 2020.

[4] Shamma Nikhat and Mustafa Mehmet-Ali. An analysis of user mobility in cellular networks. In *Proceedings of the International Symposium on Mobility Management and Wireless Access*, 2018.