Modeling Temporal-Spatial Correlations to predict New COVID-19 Cases

Group 13
Ariel Liang (s2614693) and Luuk Nolden (s1370898)

10th April 2020

1 Introduction and Summary of the Selected Paper

- The selected paper [1] aims at capturing the dynamics of urban crime by integrating fine-grained urban, mobile and public service data in order to predict criminal activities more accurately than demographic data can.
- Crime counts are organised as a matrix of N regions within a city and K time slots: each column represents a specific time slot and is associated with a feature matrix describing M features in all regions (Figure 1). This three-dimensional data is then utilised to predict the number of crimes in the $K + 1^{th}$ time slot.
- The authors focus on 133 disjointed regions of the New York City, with collected crime data from July 1^{st}, 2012 to June 30^{th}, 2013. They extracted about 50 features from sources of public security, meteorology, human mobility and public-service complaints. First, a mapping vector that projected each feature matrix onto the corresponding column of each time slot was learned, followed by a new mapping vector estimating the previous K ones and projecting the $K + 1^{th}$ feature matrix onto the prediction. Intuitively speaking, temporal-spatial correlations were modelled by similar projection rules.

2 Problem Statement

The drastic spread of COVID-19 has struck society since the beginning of 2020 and shortage of medical service is becoming an emergency in many places. We intend to employ the TCP framework proposed in [1] to capture the dynamics of the infected population in time and space, seeking to improve the efficiency of resource allocation in advance.

Although the original method overlooked the possibly stronger auto-correlation of crimes as a time series (e.g revenge has much heavier impact in street violence than similar weather and taxi trajectories), this weakness

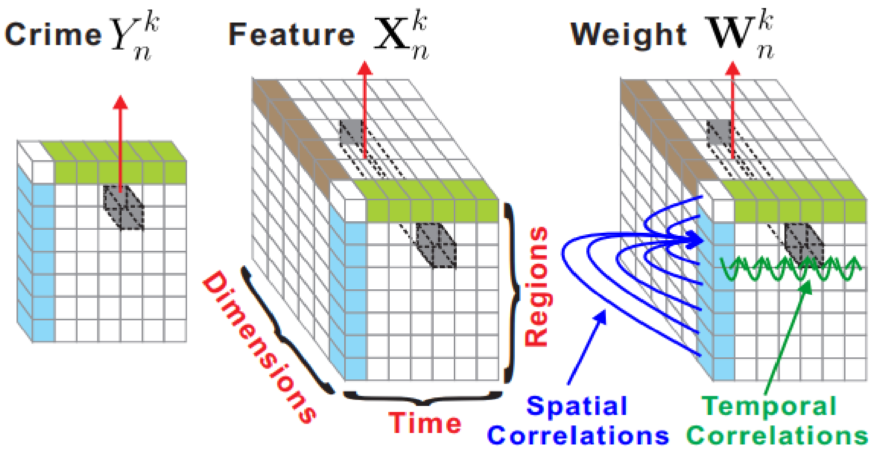

Figure 1: Illustration of data organisation

should be significantly mitigated when we consider the increase in the number of confirmed cases, rather than the aggregate count.

3 Research Questions

- Do temporal-spatial correlations exist between the new confirmed COVID-19 cases everyday in Los Angeles Metropolitan Area (16 urban regions)? Is there a pattern?
- Given relevant data until today, how many individuals will be newly confirmed as infected in each urban region tomorrow?

4 Methodology

The plan is to start with a preliminary study of the daily increases in the numbers of confirmed COVID-19 cases in the 16 counties of Los Angeles Metropolitan Area, dating from January 26, 2020, when the very first patient there was diagnosed. The results will be presented as two plots of the average differences in new confirmed cases to explore the intra-region temporal correlation and the inter-region spatial correlation respectively.

If the presence of temporal-spatial correlations are validated, we would like to apply the novel TCP framework to the COVID-19 data. A feature matrix integrates social-demographic (e.g unemployment, GDP per capita, and crime), weather (e.g average temperature, precipitation, humidity, and pressure), and check-ins from the POIs, as they have all been found relevant to the spread of a pandemic [2][3]. In addition, [4] makes it feasible to detect violations of social distancing by crawling Tweets locations and contents of gatherings, such as parties and concerts; according to WHO, this could be the most powerful factor of Corona-virus transmission [5].

The model will be trained by minimising the following objective function, using the ADMM optimisation algorithm. This procedure simultaneously pushes three pairs of values as "close" as possible, namely the actual number of new confirmed cases and the prediction from the features, the mapping vectors for subsequent time slots, and the mapping vectors for adjacent regions.

$$\min \mathcal{L}(\mathbf{W}) = \sum_{k=1}^{K} \Big(\sum_{n=1}^{N} (\mathbf{X}_n^k \mathbf{W}_n^k - Y_n^k)^2 + \frac{1}{2} \|\mathbf{W}^k \mathbf{P}\|_1 \Big) + \sum_{i=n}^{N} \|\mathbf{W}_n \mathbf{Q}\|_1$$

The solutions, $\mathbf{W}_n = [\mathbf{W}_n^1, ..., \mathbf{W}_n^K]$, $n = 1, ..., N$, will then be iteratively regressed upon to compute $\mathbf{W}_n^{K+1}$, the mapping vector for the n^{th} region looked one day forward. Finally, we predict the number of newly confirmed cases there in the next day as $Y_n^{K+1} = \mathbf{X}_n^{K+1} \mathbf{W}_n^{K+1}$.

Challenges are expected. We prepare to experiment with several definitions of distance between non-regular regions. Missing data will be handled by imputation. Furthermore, if the performance is unsatisfactory, we will try to model the number of new deaths every day.

5 Evaluation approach

- **Metrics:** The squared differences between the actual counts of daily new confirmed cases and the prediction in each urban region of Los Angeles Metropolitan Area measure model performance properly as extreme deviations from true values should be penalized.
- **Baselines:** If our results are less accurate than the Spatio-Temporal Auto-Regression, most of the predictive information has been included in history of confirmed cases already, making the integration of massive feature data unnecessary.

6 Data Sources

Funded by the U.S. National Science Foundation, [1] provided neither the dataset nor the code. It is also unpractical to work on a fine-grained grid of NYC because the coordinates of each infected patient remain unavailable. Alternatively, we decide to analyze the county-level dataset of confirmed cases released by New York Times [6], and crawl POIs from FourSquare and social activities from Twitter. Moreover, data of unemployment, temperatures, precipitation etc. in each county have always been accessible online, but we are still trying to find (or construct by ourselves) integrated datasets from the U.S. Bureau of Labour Statistics [7], Los Angeles Police Department [8], and the Weather Channel [9].

References

[1] *Proc. 17th The Conference on Information and Knowledge Management (Singapore, November 2017)*. Session 3A: Spatiotemporal. Singapore, Singapore: Association for Computing Machinery, 2017.

[2] ROCHE B. COHEN J. RENAUD F. THOMAS F. GAUTHIER-CLERC M. VITTECOQ M. “Does the weather play a role in the spread of pandemic influenza? A study of H1N1pdm09 infections in France during 2009–2010”. In: ().

[3] Amy Maxmen. *How poorer countries are scrambling to prevent a coronavirus disaster*. 2020. URL: https://www.nature.com/articles/d41586-020-00983-9 (visited on 02/04/2020).

[4] Michael W. Kearney. “rtweet: Collecting and analyzing Twitter data”. In: *Journal of Open Source Software* 4.42 (2019). R package version 0.7.0, p. 1829. DOI: 10.21105/joss.01829. URL: https://joss.theoj.org/papers/10.21105/joss.01829.

[5] World Health Organization. *Basic protective measures against the new coronavirus*. 2020. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public (visited on 18/03/2020).

[6] Data World. *US Covid-19 data from NYTimes*. 2020. URL: https://data.world/liz-friedman/us-covid-19-data-from-nytimes (visited on 01/04/2020).

[7] U.S. Bureau of Labor Statistics. *Local Area Unemployment Statistics*. 2020. URL: https://www.bls.gov/lau/ (visited on 27/03/2020).

[8] Los Angeles Police Department. *Statistical Data*. 2020. URL: http://www.lapdonline.org/statistical_data (visited on 04/04/2020).

[9] the Weather Channel. *Los Angeles, CA 10 Day Weather*. 2020. URL: https://weather.com/weather/tenday/l/a4bf563aa6c1d3b3daffff43f51e3d7f765f43968cddc0475b9f340601b8cc26 (visited on 10/04/2020).