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

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

21st April 2020

1 Introduction and Summary of the Selected Paper

- The selected paper [1] aimed 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 were organised as, $\mathbf{Y}$, a matrix of N regions within a city and K time slots; the k^{th} column represented the k^{th} time slot and was associated with an N-by-M feature matrix, $\mathbf{X}^k$, describing M features in all regions (see Figure 1). This three-dimensional data were then utilised to predict $\mathbf{Y}_n^{k+1}$, the number of crimes in the n^{th} region in the $K+1^{th}$ time slot.
- The authors focused on 133 disjointed grids of the New York City, with collected crime data from July 1, 2012 to June 30, 2013. They extracted about 50 features from sources of public security, meteorology, human mobility, and public-service complaints. First, a mapping vector, $\mathbf{W}_n^k$, that projected features $\mathbf{X}_n^k$ onto Y_n^k was learned for each n and k. Then a new mapping vector, $\mathbf{W}_n^{K+1}$, was estimated from the previous K ones and projected $\mathbf{X}_n^{K+1}$ onto the prediction, $\hat{Y}_n^K$. 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 become 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.

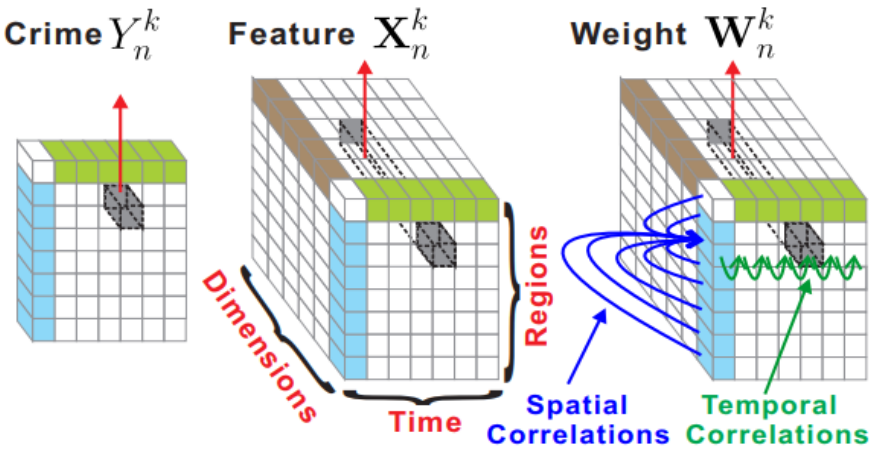

Figure 1: Illustration of data organisation

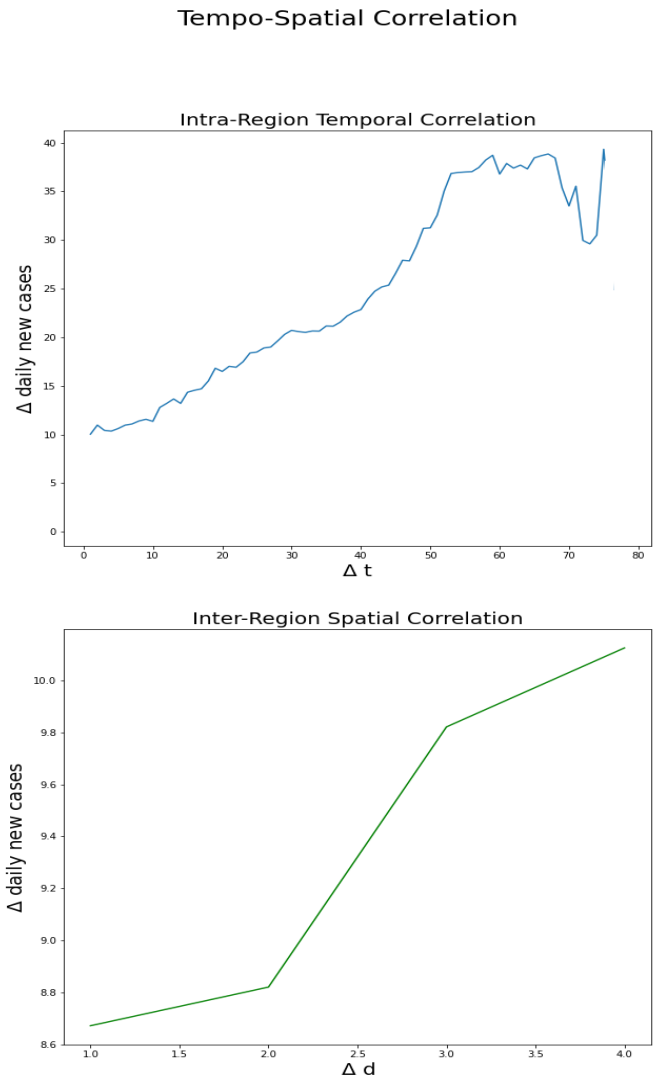

Figure 2: Tempo-Spatial Correlation among SFBA Daily New Confirmed Cases[1]

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 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 San Francisco Bay Area (9 urban counties)? Is there a pattern?
- Given relevant data until today, how many individuals will be newly confirmed as infected in each county tomorrow?

[1] https://github.com/PawinData/UC/tree/SFBA

4 Methodology

4.1 Validation of Temporal-Spatial Correlation

We have validated the presence of **intra-region temporal correlation** and **inter-region spatial correlation** among the numbers of daily new confirmed COVID-19 cases in the nine counties of San Francisco Bay Area (**SFBA**) from January 31 to April 13, 2020, defining as distance how many times one has to cross county borders to arrive one county from another. With either the increase of time lag or distance between two counties, the difference in daily new cases tends to climb up in general (see Figure 2). SFBA is chosen because the arrangement of its counties is most grid-like (for example, the counties in New York Metropolitan Area are arranged in a radial pattern), and the approach proposed in [1] only works for grid data.

4.2 Prediction

In order to model temporal-spatial correlations quantitatively, we would like to apply the novel TCP framework to the COVID-19 data. Feature matrices $\{\mathbf{X}^k\}_{k=1}^{k=K}$ integrate 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 violation of social distancing by crawling Tweets locations and contents of gatherings, such as parties and concerts; according to WHO, this can be the most powerful factor of Coronavirus 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, (i) the actual number of new confirmed cases and the prediction from the features, (ii) the mapping vectors for subsequent time slots, and (iii) the mapping vectors for adjacent regions, represented by the three terms in the objective function respectively. $\mathbf{P}$ and $\mathbf{Q}$ are simply sparse matrices that generate difference terms for subsequent time slots or adjacent regions.

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

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 $Y_n^{K+1} = \mathbf{X}_n^{K+1}\mathbf{W}_n^{K+1}$. If the performance is unsatisfactory, we will try to model the number of new deaths every day.

5 Evaluation approach

- **Metrics:** The mean squared error (MSE) 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 (**STAR**)[2], 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], San Francisco Police Department [8], and the Weather Channel [9].

[2]We will write our own program for STAR: $\mathbf{Y}^k = c + \sum_{\tau=1}^{\tau=T}\Phi_\tau\mathbf{D}_N\mathbf{Y}^{k-\tau} + \epsilon^k$, $(k = 1, ..., K)$, where $\mathbf{Y}^k$ and ϵ^k are N-element vector, and experiment with different T and $\mathbf{D}_N$ (e.g inverse-distance matrix).

References

[1] Tang J. Zhao X. "Modeling Temporal-Spatial Correlations for Crime Prediction". In: *Proc. 17th The Conference on Information and Knowledge Management (CIKM'17)*. Session 3A: Spatiotemporal. Association for Computing Machinery. Singapore, Singapore, 2017. DOI: https://doi.org/10.1145/3132847.3133024.

[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: https://www.sanfranciscopolice.org/stay-safe/crime-data/crime-dashboard (visited on 20/04/2020).

[9] the Weather Channel. *San Francisco, CA 10 Day Weather*. 2020. URL: https://weather.com/weather/tenday/l/69bedc6a5b6e977993fb3e5344e3c06d8bc36a1fb6754c3ddfb5310a3c6d6c87 (visited on 20/04/2020).