EXPLORATION OF POTENTIAL FACTORS AFFECTING THE SPREAD OF COVID-19 USING SPATIOTEMPORAL ANALYSIS

A Thesis

by

ZIYI ZHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Nicholas Duffield
Committee Members,	Zhe Zhang
	Robert Hardin
	Kevin Nowka
	Xiaoning Qian
Head of Department,	Aniruddha Datta

May 2021

Major Subject: Engineering

ABSTRACT

The coronavirus (COVID-19) has caused a huge negative influence on the global. The virus continues to spread and has sickened more than 90,201,652 people until January 2021 and caused 1,937,091 deaths in the world. So far, social distancing was given as an effective way to control the coronavirus. Governments issue restrictions on traveling, institutions cancel gatherings, and citizens socially distance themselves to limit the spread of the virus. The thesis's main focus is to explore the spatiotemporal patterns under the COVID-19 pandemic. Additionally, we conducted both spatiotemporal modeling analysis and spatiotemporal clustering analysis in Texas to identify some potential variables which are associated with the increase of confirmed cases. My thesis will help local governments locate the medical facilities and improve the social distancing recommendations regarding the COVID-19 outbreak. For instance, governments can use our results to locate risk areas and enforce guidelines to limit interaction in those areas and provide additional medical facilities.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Duffield, and my committee members, Dr. Zhang, Dr. Qian, Dr. Hardin, and Dr. Nowka, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother and father for their encouragement.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Nicholas Duffield, Xiaoning Qian, and Kevin Nowka of the Department of Electrical and Computer Engineering, Professor Zhe Zhang of the Department of Geography, and Professor Robert Hardin of the Department of Biological and Agricultural Engineering.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by Institute of Data Science of Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

1. INTRODUCTION

The coronavirus (COVID-19) has caused a huge negative influence on the global. The virus continues to spread and has sickened more than 90,201,652 people up until January 2021 and caused 1,937,091 deaths worldwide. Governments issued restrictions on traveling, institutions cancelled gatherings, and citizens socially distanced themselves to limit the increase of confirmed cases. COVID-19 has dragged down the global economy and travel around the world is restricted. Research has indicated that some spatial factors and locations are determined as a vital role in the early outbreak [1]. Although some timely efforts have been done such as restrictions on food, drink, non-essential travel, and social distancing, and some studies have been conducted on reducing the increase of confirmed cases, we found few works focused on the spatial and temporal variations of explanatory factors simultaneously. To analyze the relative importance of different factors in their influence of the evolution of the ongoing pandemic, researchers need to adopt an interdisciplinary approach. Spatiotemporal analysis should be firstly considered to handle COVID-19 related variables since these factors are changing over both spatially and temporally[2].

Many of unknowns affecting the spread of COVID-19 are spatially autocorrelated and thus the research should have ability to handle some spatial variables from different fields to interpret the outbreak and spread of COVID-19[3]. Researches across the world have shown that factors such as weather[4], human mobility[5], social vulnerability[6], and environmental conditions[7] may influence the severity of COVID-19. Until now, research concerning spatial analysis of potential factors affecting the spread of COVID-

19 has been conducted using Geographically Weighted Regression[8][9] and Multiscale Geographically Weighted Regression. However, these methods did not incorporate temporal analysis, which may lead to erroneous estimates since they cannot capture temporal variation as the epidemic progresses. The goal of our study is to explore the associations between time-dependent variables such as mobility, weather, demographic variables and the outbreak of COVID-19.

Our study can be separated into two main parts: exploration using extended GWR models and spatiotemporal clustering analysis.

1.1. Exploration using Extended GWR Models

Studies have explored demographic factors by using Geographically Weighted Regression [8][9]. For example, Karaye et al. (2020) [6] explored how social vulnerability influenced the COVID-19, and Pierre et al. (2021) explored the reduction in mobility and COVID-19 transmission. However, these studies only considered the spatial variations or added temporal information after finishing exploring the spatial variations[2], few studies handled spatial and temporal variations simultaneously when exploring COVID-19 related variables. Our experiment compared the performance of Geographically Temporally Weighted Regression[11] and Spatiotemporal Weighted Regression[12], and then applied the better model to explore the spatiotemporal variations of explanatory simultaneously. The result of the first part indicates Social Vulnerability Index (SVI), Population density, and Climate are associated with the increase of confirmed cases, and population density is the most important factor included in the model to drive the spread of COVID-19. With the advantage of extended GWR including GTWR and STWR on handling spatial and temporal variations simultaneously, we took Texas as a study area and analyzed the effect of different variables on county level. We found population density is the main driving factor in the majority of counties in Texas, especially for the middle and northeast of Texas; Meanwhile, our results indicate that people are easily infected in a low wind speed, low temperature, and high SVI environment in the majority of counties in Texas.

1.2. Spatiotemporal Clustering Analysis

In this section, our focus switched to spatiotemporal clustering analysis. Studies have explored the pattern of COVID-19 related data by using time series clustering. For example, Huang et al.[13] applied K-means time series clustering on home dwell time records from Safe Graph [14], which reflects the pattern of length of time people stay in their home. However, clustering time series by K-means with Euclidean distance may cause misalignment because the time lag for each time series is different. K-means with dynamic time warping (DTW) on clustering time series can be treated as a good way to eliminate this misalignment, but clustering with K-means with DTW may cause a random result[15] (Details in Section 3). In order to eliminate the misalignment, Bathwal et al. [15] proposed a novel time series clustering method based on magnitude-differences (*dm*) over the available time-differences (*dt*)[16], and applied it to COVID-19 death time series. However, we found that the clustering result is easily affected by the size of the bin, and the random choice of the bin range may cause a misleading result. Another drawback for this clustering method is that the final result is hard to be interpretable. Therefore, inspired by a geo-referenced time series co-clustering method

[17] which can map locations and timestamps to the location and time step cluster simultaneously, we proposed a novel feature-based deep geo-referenced time series clustering method (DGTSC). The objective of our method is not only clustering the location, but also clustering the time period globally. The result from DGTSC will indicate which counties are clustered during which time period simultaneously. We applied DGTSC on the mobility data of Texas from the Bureau of Transportation Statistics[18] during COVID-19, and combined the policy changes and holidays to make the results interpretable.

2. EXPLORATION USING EXTENDED GWR MODELS

Recent studies have found that spatial regression is an effective strategy to model phenomena such as the outbreaks of COVID-19, and to better understand it [1]. Therefore, we firstly choose spatial regression as our basic methodology to analyze the relationship between COVID-19 and some potential driving factors, and then we improve the methodology and performance step by step based on the simple global spatial regression in single spatial analysis. In general, two strategies are commonly used: one is global regression, and the other is local regression. The famous and popular model under the global regression strategy is ordinary least square (OLS) regression, which has the general form of:

$$y_i = \beta_0 + x_i\beta + \varepsilon_i$$

where y_i is the regression point at location i, β_0 is the intercept, x_i is the observation at location i, β accounts for the influence and importance of observation x_i on regression points y_i, and ε_i is the residual.

However, when doing geographic spatial analysis, geographic features cannot meet the assumptions and requirements of OLS regression due to geographic features being spatially autocorrelated, and the modeling process is non-stationary. Thus, geographically weighted regression (GWR) [8][9] was proposed to handle geographic features. GWR has achieved a great success in COVID-19 spatial analysis with its excellent performance on handling spatial variation, Maiti et al. (2021) [2] explored some COVID-19 related driven factors in the United States, but along with the arise of time-dependent data, the limit of GWR on handling temporal information is gradually being discovered, for example, GWR can account for the population variation of some counties on spatial in a specific year, but it cannot well-handled the variation of the data on temporal from a year to another year. Considering some potential factors affecting the spread of COVID-19 are time-sensitive variables such as weather and mobility, a novel method should be applied to handle the spatial and temporal information simultaneously. So, we switched from GWR which can only handle spatial variation to GTWR and STWR, which are proposed by Huang et al.[11], Fotheringham el al. [21]and Xiang Que et al.[12], and compared their performance on handling COVID-19 related data.

2.1. Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a spatial regression strategy[8][9], which can estimate parameters locally. the form of GWR model is:

$$Y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)\, X_{ik} + \varepsilon_i$$

where (u_i, v_i) is defined as the coordinates of location i, $\beta_0(u_i, v_i)$ is defined as intercept of GWR, and $\beta_k(u_i, v_i)$ is the coefficient estimated at i, X_{ik} is the k^{th} observation at i. Unlike the global regression, GWR allows the parameter to be estimated locally on a spatial basis and thus handles local spatial variation well. In the GWR model, $\beta_k(\cdot)$ is estimated by using least square estimation:

$$\hat{\beta}(\cdot) = [X^T W(\cdot) X]^{-1} X^T W(\cdot) Y$$

where $W(\cdot)$ is a $n \times n$ weight matrix for location i, whose diagonal elements represents the level of influence of observation points on location i. According to Tobler's first law[19], the observation close to location i will influence much more on location i than the observation which is farther from i. Therefore, if the observation is close to location i, the weight will be greater. Two strategies are commonly used to construct the weighted matrix: one is the fixed kernel and the other is the adaptive kernel. The difference between fixed kernel and adaptive kernel is: distance is fixed in fixed kernel but an adaptive kernel means the number of neighbors is fixed but distance varies. GWR[8] used Gaussian decay-based function to construct the weight matrix:

$$W_{ij} = \exp\,(-\frac{d_{ij}^2}{h^2})$$

where h denotes as bandwidth, and d_{ij} represents the spatial Euclidean distance between location i (u_i, v_i) to location j (u_j, v_j), which can be expressed as:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

from the above equation, we can see if $i = j$, the distance will be 0, and the weight will be unity. If the spatial distance increases, the weight will meet a decrease. Another popular and commonly used function in GWR is Bi-square kernel function, given a specific bandwidth h, and the weight matrix is:

$$w_i(u_i, v_i) = \begin{cases} [1 - (\frac{d_{ij}^2}{h^2})]^2, & |d_{ij}| < h \\ 0, & |d_{ij}| > h \end{cases}$$

where d_{ij} is defined as the distance from location i (u_i, v_i) to the location j (u_j, v_j), and h is the optimal bandwidth. The choice of the optimal bandwidth is also important after the determination of weight matrix, there are three commonly used criterions in GWR model:

1. Cross-validation (CV) criterion:
   Given a specific bandwidth h and remove the i^{th} observation, then estimate the parameter using the rest $(n-1)$ observations and get the result $\hat{Y}_{(-i)}(h)$ on X_i.

$$CV(h) = \frac{1}{h}\sum_{i=1}^{n}(Y_i - \hat{Y}_{(-i)}(h))^2$$

   The optimal bandwidth will be:

$$h_0 = arg \min_{h>0} CV(h)$$

2. Generalized cross-validation (GCV) criterion:
   Assume $Y(h) = (Y_1(h),\, Y_2(h), \dots, Y_N(h)) = L(h)Y$, and GCV can be expressed as:

$$GCV(h) = \frac{n}{(n - tr(L(h)))^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i(h))^2$$

   The optimal bandwidth h_0 will be:

$$h_0 = \ arg \min_{h>0} GCV(h)$$

3. Corrected Akaike Information (AIC) Criterion:
   Let $\hat{Y}(h) = L(h)Y$, and $\hat{\varepsilon} = Y^T\left(I_n - L(h)\right)^T\left(I_n - L(h)\right)Y$, AIC can be expressed as:

$$AICc(h) = \log\left(\frac{1}{n}\hat{\varepsilon}^T\hat{\varepsilon}\right) + \frac{n + tr(L(h))}{n - 2 - tr(L(h))}$$

   and the optimal h_0 will be:

$$h_0 = \ arg \min_{h>0} AICc(h)$$

2.2. Geographically Temporally Weighted Regression

GWR can be used as a local variation modeling technique to explore the potential factors affecting spread of COVID-19 to some extent. However, for some COVID-19 related data such as confirmed cases, deaths and weather, the temporal effect is also important, and thus the limit of GWR on handling temporal effects is obvious. In order to not only account for spatial variations, but also account for temporal variations. We explored the existing spatiotemporal methodologies, and found some works [11] [21] extended the traditional GWR into a Geographically Temporal Weighted Regression model which captures both spatial and temporal variations. In the case study given in the GTWR, its performance on modeling spatiotemporal house price exceeded that of GWR with a higher R^2 with an optimal balanced parameter. The main difference between GWR and GTWR is: the weight matrix in GTWR is determined not only by spatial distance but also by temporal distance between two locations. Unlike the format of input in GWR (x, y), which only includes the coordinate of location i, GTWR added the timestep t and changed the form as (x, y, t), where (x, y) represents the latitude and longitude of location i, and t represents the time stage. Therefore, the GTWR extended from GWR can be expressed as:

$$Y_i = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i)X_{ik} + \varepsilon_i$$

where (u_i, v_i) represents the latitude and longitude of location i, k represents the k^{th} observations in location i, and t_i represents the time stage. Similarly, β in GTWR can be estimated by using ordinary least square which can be expressed as:

$$\hat{\beta}(u_i, v_i, t_i) = [X^T W(u_i, v_i, t_i)X]^{-1} X^T W(u_i, v_i, t_i)Y$$

In order to balance the temporal and spatial weights, μ and λ are introduced, and the spatiotemporal distance in GTWR can be expressed as:

$$(d_{ij}^{ST})^2 = \lambda\left[\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2\right] + \mu\left(t_i - t_j\right)^2$$

where (u_i, v_i) (u_j, v_j) represents the coordinate of location i and j, t_i and t_j are observed time stages at location i and j, λ and μ are the parameter to balance the distance between spatial distance and temporal distance. If we define: $\tau = \frac{\mu}{\lambda}$, the above equation can be rewrite as:

$$\frac{(d_{ij}^{ST})^2}{\lambda} = \left[\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2\right] + \tau\left(t_i - t_j\right)^2$$

λ and μ can be determined by cross-validation in terms of R^2.

2.3. Spatiotemporal Weighted Regression

Although GTWR can handle the temporal and spatial variations to some extent, another existing spatiotemporal regression methodology called spatiotemporal weighted regression was first proposed by Xiang Que et al. (2020), points out the concept in GTWR on treating temporal distance as time interval is inappropriate[12], and STWR proposed a novel method to balance the temporal and spatial variation, besides, STWR designed a novel temporal kernel function. In GTWR models, which proposed by Huang et al.[11] and improved by Fotheringham. [21], time interval $t_1 - t_2$ is used to account for temporal variation, and Gaussian and Bi-square kernel functions from GWR are used directly to construct the weight matrix in GTWR. However, STWR points that these two strategies are not reasonable. Firstly, instead of using time interval $t_1 - t_2$ in GTWR, STWR used the value variation during a time interval in order to make the temporal distance be more reasonable, which can be expressed as $\Delta t = \frac{(y_1 - y_2)/y_2}{t_1 - t_2}$. Take the confirmed case from COVID-19 as an example, if the confirmed case variation is too small, the observation at this time may have a relative low influence. Secondly, unlike the temporal kernel function in GTWR which still used Gaussian function, STWR designed a novel temporal kernel function:

$$w_{ij\Delta t}^{t} = \begin{cases} \left[\dfrac{2}{1 + \exp\left(-\dfrac{\left|(y_{j(t)} - y_{j(t-q)})/y_{j(t-q)}\right|}{\Delta t/b_T}\right)}\right] - 1, if\ 0 < \Delta t < bT \\ 0, \qquad otherwise \end{cases}$$

where $y_{j(t)} - y_{j(t-q)}$ is the difference of value between observation and regression location i. $y_{j(t)} - y_{j(t-q)}/y_{j(t-q)}$ is the change rate. The faster value change, the bigger weight is and the larger impact is. The spatial kernel function is the same as that in GTWR: Gaussian kernel function and Bi-square kernel function. Finally, the novel spatiotemporal kernel function can be expressed as:

$$w_{ijST}^{t} = (1-\alpha)k_s\left(d_{sij}b_{ST}\right) + \alpha k_T\left(d_{ij}b_T\right), 0 \leq \alpha \leq 1$$

where k_s and k_T are is the spatial and temporal kernel function, b is the bandwidth and α is a parameter to adjust the spatial and temporal weight. The bandwidth and parameter estimation can be obtained by using Cross Validation and AIC. Therefore, Xiang Que et al. (2020) purports to develop a novel strategy in STWR which aims to account for temporal variations more perfectly and construct the kernel function to optimize spatial and temporal effects simultaneously.

2.4. Datasets

2.4.1. Data Collection, Selection and Preprocessing

We collected the weather, health, demographic which contains 213 variables, and COVID-19 confirmed case data. We matched these variables by their corresponding Federal Information Processing Standard (FIPS). In the process of selection of variables, we did the collinearity test through computing Variance Inflation Factors (VIF) of 213 variables, and assume a high collinearity will be met if VIF value is higher than 10. For example, we found the VIF value of Social Vulnerability Index and some demographic variables such as unemployment and poverty are higher than 10, which means they have a high collinearity, this is caused by the calculation of SVI which considers socioeconomic status such as income and unemployed rate. We filtered the redundant variables step by step in order to ensure the multicollinearity was entirely eliminated. Finally, we chose population density, wind speed, temperature, humidity and SVI as our independent variables (Table 1: VIF Value).

	VIF
Confirmed Case	3.981
Population Density	4.043
Wind Speed	1.554
Temperature	3.090
Humidity	2.398
SVI	1.065

Table 1: VIF Value

In order to improve the efficiency of computation in models, we did a Z-score normalization on filtered variables, whose result is standardized by mean and standard deviation of input variables. The processes of variables were conducted in Python 3.6.

2.4.2. COVID-19

COVID-19 U.S. confirmed case data from New York Times are the cumulative number of confirmed cases and daily confirmed cases in the form of time series.

2.4.3. Demographic Variables

Demographic Variable contains population density data, we used the population density in each county as one of our potential factors which helps the spread of COVID-19.

2.4.4. Weather

Weather includes temperature, wind speed and humidity data. Over 9000 stations' data are available. We used temperature, wind speed, and humidity. Each county is paired with the nearest weather station. Most stations are within 50 km of the county center, and virtually all are within 100 km of the county center.

2.4.5. Social Vulnerability Index (SVI)

The SVI is calculated by 14 social factors such as socioeconomic factors, household and transportation factors. SVI can help governments to understand which county is easily affected by some disasters such as flood or pandemic. For example, if the SVI of a county is higher, which means this county may easily be affected by some disasters.

2.5. Comparison and Result Analysis

We compared the performance of OLS, GTWR, STWR on handling Texas weather, population, and SVI data respectively, and the result indicates that STWR outperforms when compared with OLS and GTWR, with a higher $R^2 = 0.865$, lower $Root\ Sum\ Square\ (RSS) = 16.63$, lower $AIC = 22.535$, and $AICc = 32.004$.

Comparison:

Model	***R2***	***Adj R2***	***RSS***	***AIC***	***AICc***
OLS	*0.454*	*0.443*	*288.86*	*1688.734*	*1688.832*
GTWR	*0.812*	*0.805*	*94.74*	*528.998*	*598.933*
STWR	***0.865***	***0.861***	***16.63***	***22.535***	***32.004***

Table 2: Result of OLS, GTWR& STWR

2.5.1. GTWR Coefficient Estimation and Optimal Parameter

Optimal Parameter:

Parameter	**Value**
Optimal Bandwidth	3.56
Lambda	0.80
Adaptive	FALSE
Kernel	bi-square

Table 3: Optimal Parameter in GTWR

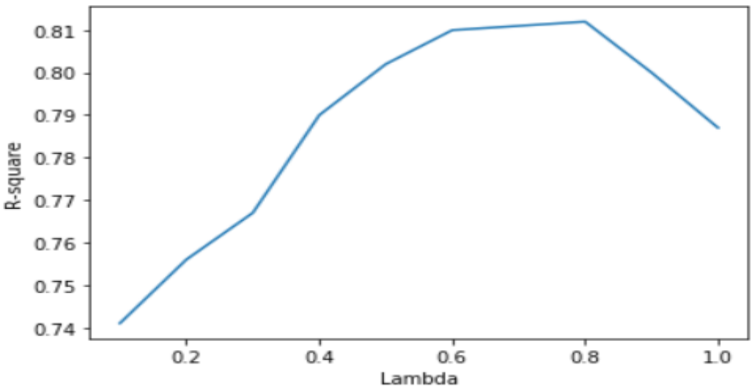

Figure 1: Balanced Parameter

In the GTWR, we chose Cross-Validation as the criterion and Bi-square as kernel function. The optimal spatiotemporal bandwidth is 3.56, and the optimal balance parameter between spatial and temporal effects is $\lambda = 0.8$ with the highest $R^2 = 0.812$.(Figure 1: Balanced Parameter)

Coefficient Estimation:

	Min	**Median**	**Max**
Intercept	-6.820	-0.043	1.149
Wind Speed	-2.136	0.002	0.450
Humidity	-2.319	-0.017	0.451
Temperature	-2.239	-0.011	1.195
SVI	-0.154	0.081	2.142
Population Density	-0.796	0.846	4.621

Table 4: Coefficient Estimation in GTWR

The parameters in (Table 4: Coefficient Estimation in GTWR) are generated from GTWR on modeling weather, population density and SVI with optimal parameters mentioned above. From the estimation of coefficients in GTWR, we found although these variables have been proved to be the driven factors affecting the spread of COVID-19, when compared with each other, population density will be the dominant factor which caused the increase of confirmed cases. The work of GTWR was conducted in R 4.0.2.

2.5.2. STWR Coefficient Estimation and Optimal Parameter

Optimal Parameter:

Parameter	**Value**
Optimal Bandwidth	8.90
Alpha	0.20
Adaptive	FALSE
Kernel	bi-square

Table 5: Optimal Parameter in STWR

In the STWR, the optimal bandwidth is 8.90 which calculated by Cross-Validation, and the balance factor $\alpha = 0.2$. We set adaptive be false, and used the same spatial kernel function as that in GTWR.

Coefficient Estimation:

	Min	**Median**	**Max**
Intercept	-0.498	-0.136	0.832
Wind Speed	-0.113	0.007	0.124
Humidity	-0.324	0.019	0.149
Temperature	-0.406	-0.087	0.159
SVI	-0.003	0.066	0.290
Population Density	0.637	0.929	4.564

Table 6: Coefficient Estimation in STWR

From the (Table 6: Coefficient Estimation in STWR), we can see that weather and SVI can affect the spread of COVID-19 to some extent, but population density is still the dominant factor affecting the spread of COVID-19 in STWR. This work was conducted in Python 3.6.

2.5.3. Visualization on the Map

We finally chose STWR to model the spatial and temporal variations simultaneously since it has a better performance when compared with OLS and GTWR even if STWR cost much time than GTWR.

Firstly, we generated the local R^2 (Figure 2: Local R2) by using STWR, and found the R^2 of the majority of counties are higher than 0.8, which means the confirmed case can be explained by weather, SVI, and population density very well in most counties of Texas. Then, we displayed the estimation of coefficients of each county on the map (Figure 3: Coefficient of Intercept to Figure 8: Coefficient of SVI) so that we can see the influence of these variables on affecting the spread of COVID-19 on different locations.

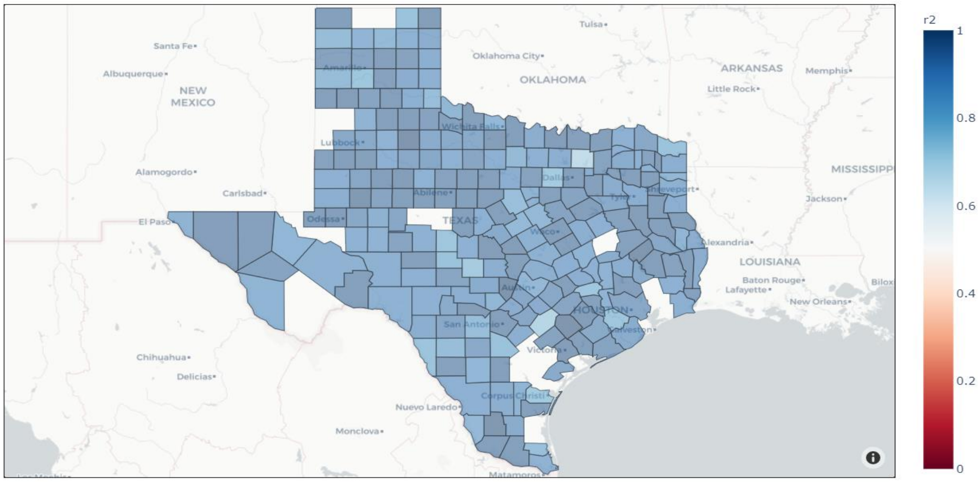

Figure 2: Local R2

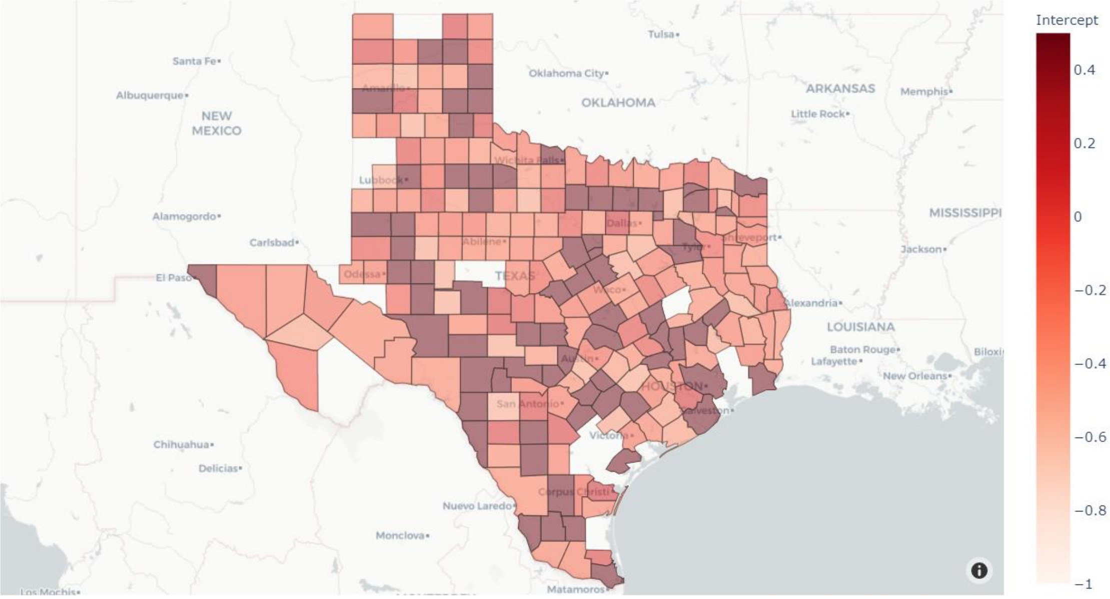

Figure 3: Coefficient of Intercept

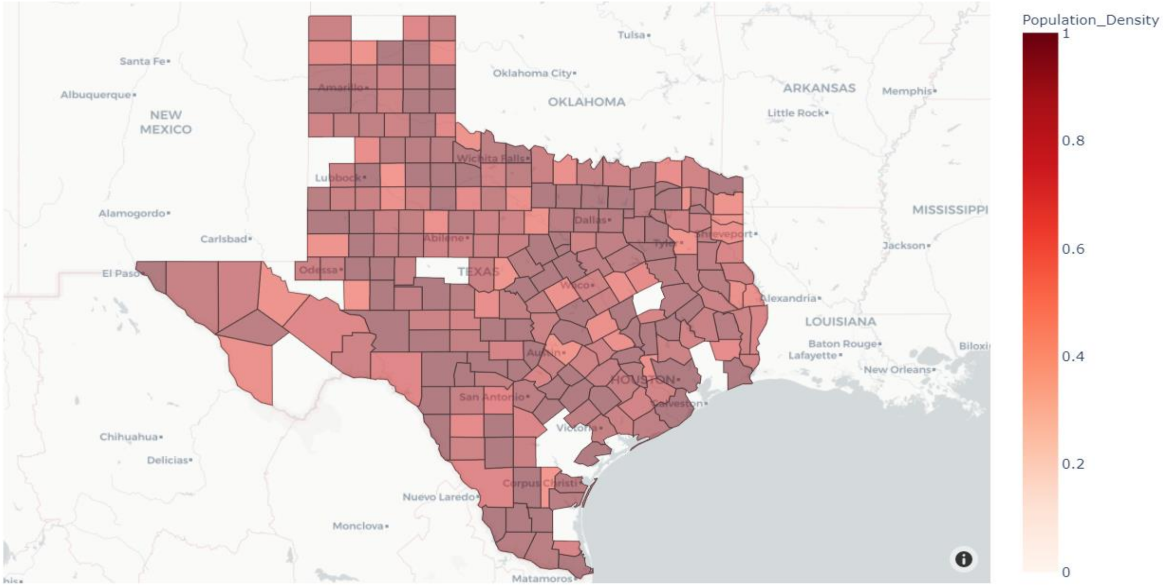

Figure 4: Coefficient of Population Density

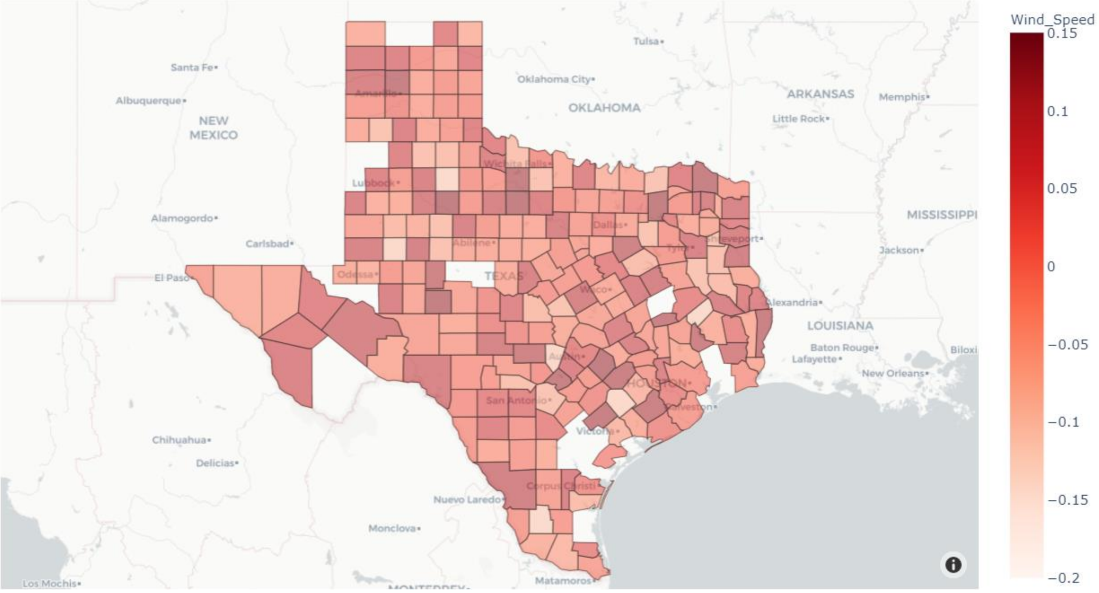

Figure 5: Coefficient of Wind Speed

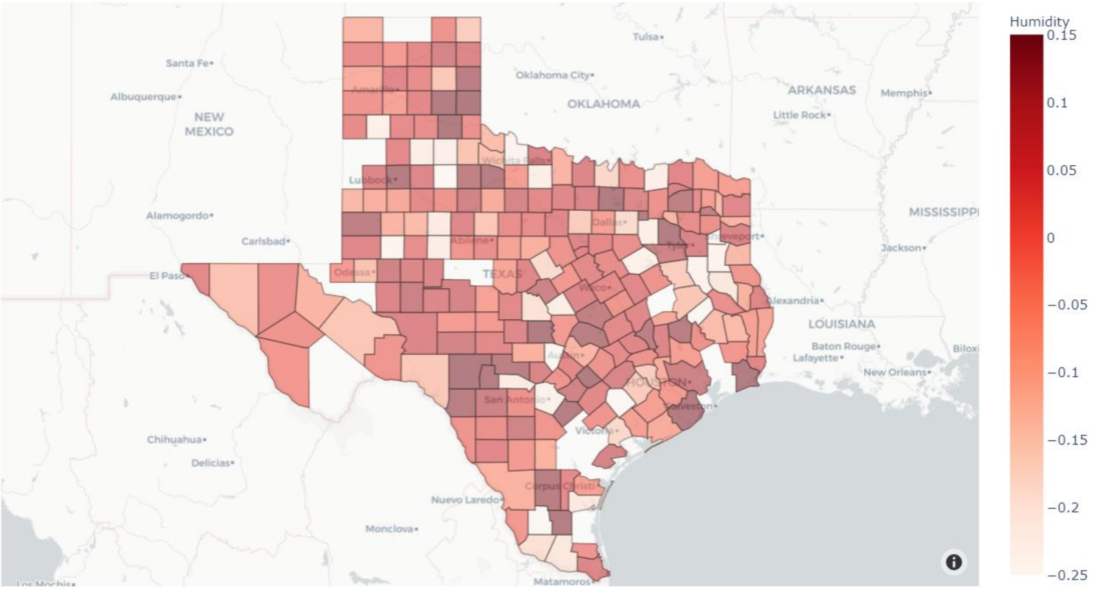

Figure 6: Coefficient of Humidity

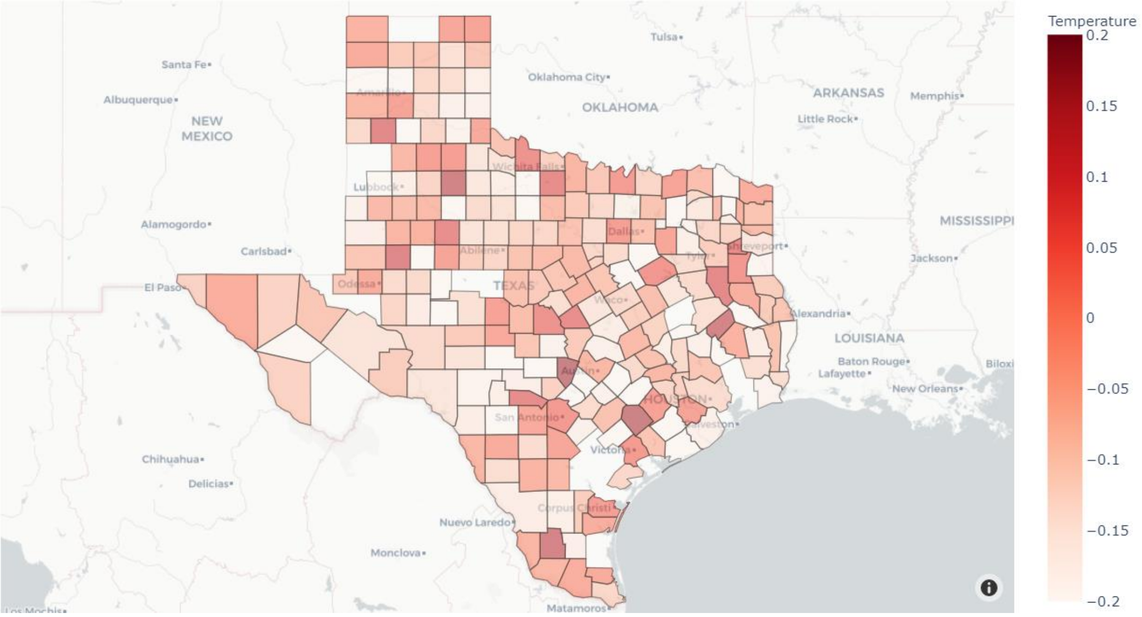

Figure 7: Coefficient of Temperature

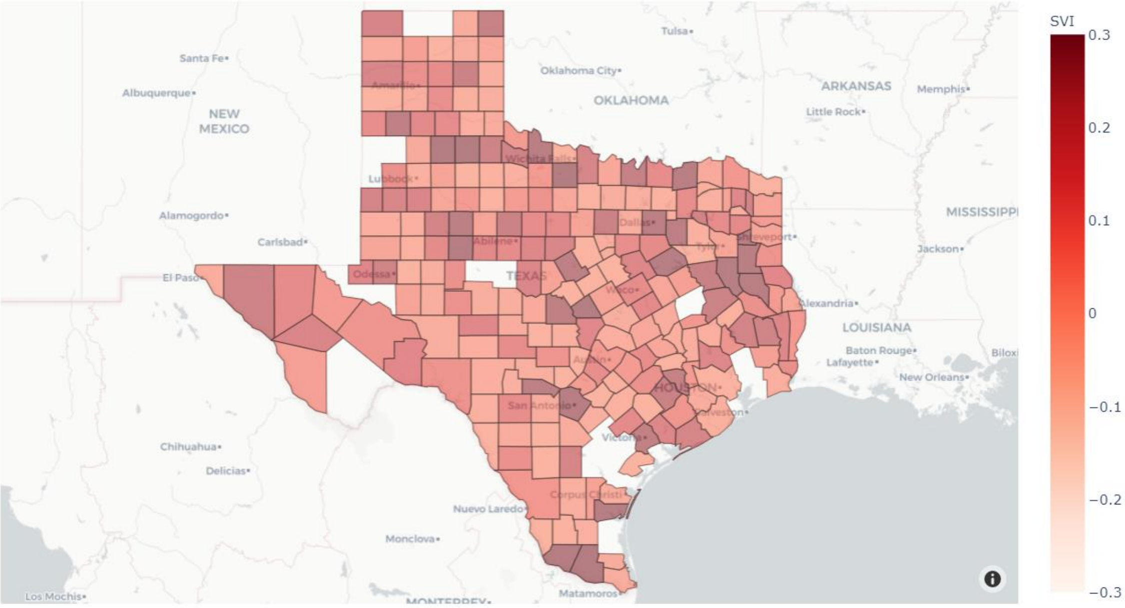

Figure 8: Coefficient of SVI

2.6. Conclusion

(Figure 3: Coefficient of Intercept to Figure 8: Coefficient of SVI) shows us the estimation of coefficients in Texas on county level. Although some studies have been done which provide the foundation that increase of confirmed cases is associated with the temperature, wind speed, humidity[22], SVI[23], and population density[24]. Few studies explored spatial and temporal variations simultaneously and revealed which factor is the most influential to affect the spread of COVID-19. In the first part of our study, we explored these factors and found population density is the main driving factor in the majority of counties in Texas, especially in the middle and northeast of Texas, to help the increase of confirmed cases. Besides, we found wind speed and humidity will also be associated with the increase of confirmed cases in some scatter counties. Although the spread of COVID-19 is fastened by low temperature in majority of counties, high temperature in some counties is evident to help the spread of COVID-19 such as Blanco County (FIPS: 48031), Lavaca County (FIPS: 45285) and Jim Hogg County (FIPS: 48247). Meanwhile, our results indicated that people are easily infected in a low wind speed, high humidity, low temperature and high SVI environment in most counties of Texas.

3. SPATIOTEMPORAL CLUSTERING ANALYSIS

3.1. Introduction to Time Series Clustering

Time-series data are increasing with the improvement of data storage[25]. Extracting and analyzing patterns from time series data is important[26][27] since we can learn some useful knowledge which is hidden in the data[28]. Clustering has been widely used in spatiotemporal data mining which aims to group similar data [29]. Therefore, it is necessary to cluster spatiotemporal COVID-19 related data so that we can learn some hidden knowledge. Time series clustering is to find patterns in time series. In general, there are three common ways to cluster time series[25]: Whole time series clustering means clustering on the whole raw time series without doing any transformation on original time series, subsequence clustering will cluster the subsequences generated from original time series, and point clustering is based on the time-point values. [25] Furthermore, there are two common approaches for clustering time series: shaped-based and feature-based. In a shape-based approach, time clustering algorithm will cluster the time series if their shapes are similar. With a feature-based approach, we need to extract features from original time series and then apply K-means, DBSCAN or some clustering algorithms on these features.

3.2. Literature Review

Recent studies have explored the COVID-19 related data by using time series clustering, for example, Huang et al.[13] applied K-means on home dwell time records from Safe Graph [14]. However, we found it may generate a misleading result if the time lag between two times-series is different when we applied this strategy on mobility datasets from BTS. Dynamic Time Warping (DTW) is an algorithm to match two time series as possible and thus eliminate the time lags between two time series (Figure 9: Dynamic Time Warping), which can be treated as a good way to eliminate this misalignment. domain[30].

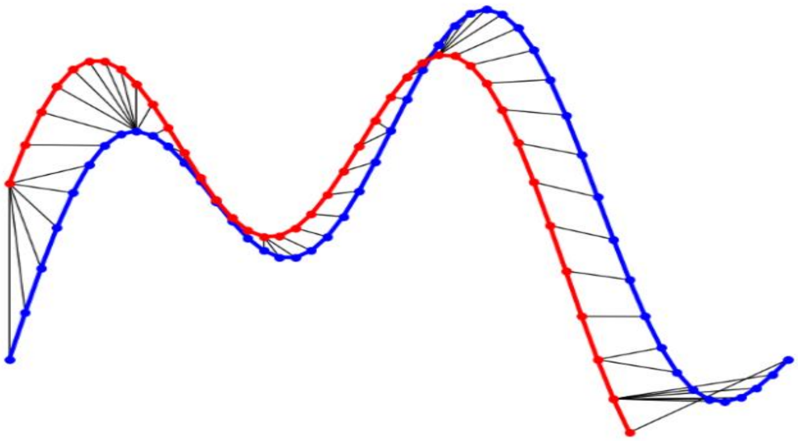

Figure 9: Dynamic Time Warping (Image by Esmaeil Alizadeh[31])

However, the result of clustering geo-referenced time series by using K-means + Euclidean distance or DTW distance seemed random. (Figure 10: Clustered by using K-means + DTW).

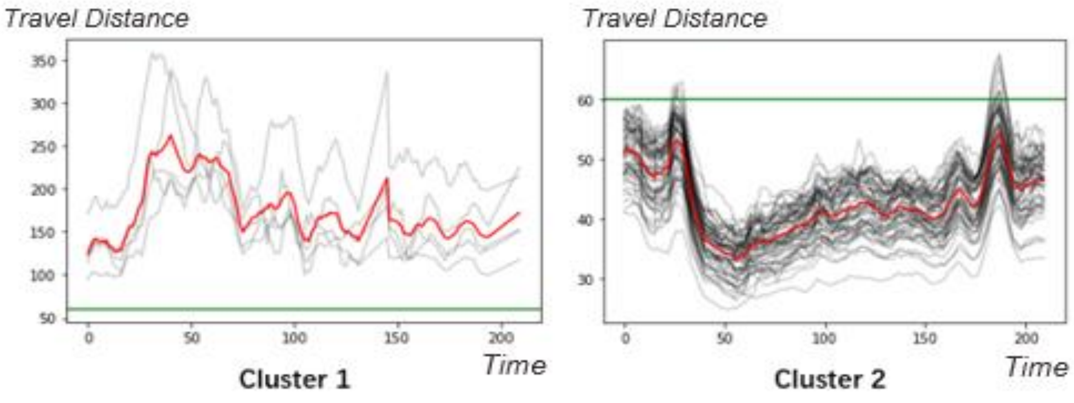

Figure 10: Clustered by using K-means + DTW

Bathwal et al. (2020) also mentioned this problem in their work and proposed a novel method based on *dmdt* [16], which is a concept of transferring from time series into image. (Figure 11: The novel method [15] *based* **on *dmdt***). Take the time-series generated by death during COVID-19 in a county as an example, *dmdt* will generate the difference of magnitude and the difference of time firstly in the form of $[dm, dt]$, and then the bin range will be set manually. Meanwhile, an image with size 8×8 will be created according to the bin size to receive the patterns generated by *dmdt*, thus the color for each bin represents the number of features located in that bin (pixel). In the last step, a new 4×4 image will be transformed after a max-pooling, and this 4×4 image will be flattened into a 1×16 vector to do clustering.

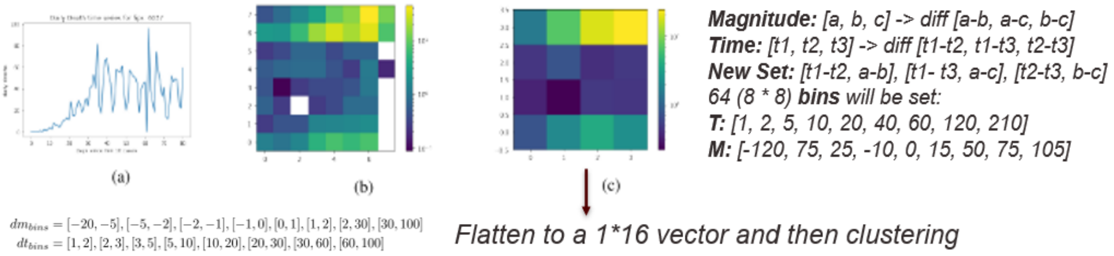
(a)
(b)
(c)
Magnitude: [a, b, c] -> diff [a-b, a-c, b-c]
Time: [t1, t2, t3] -> diff [t1-t2, t1-t3, t2-t3]
New Set: [t1-t2, a-b], [t1- t3, a-c], [t2-t3, b-c]
64 (8 * 8) bins will be set:
T: [1, 2, 5, 10, 20, 40, 60, 120, 210]
M: [-120, 75, 25, -10, 0, 15, 50, 75, 105]
$dm_{bins} = [-20, -5], [-5, -2], [-2, -1], [-1, 0], [0, 1], [1, 2], [2, 30], [30, 100]$
$dt_{bins} = [1, 2], [2, 3], [3, 5], [5, 10], [10, 20], [20, 30], [30, 60], [60, 100]$
Flatten to a 1*16 vector and then clustering

Figure 11: The novel method [15] based on *dmdt* [16]

We applied the method based on *dmdt* on mobility datasets, and we tried both max-pooling and average-pooling when doing image transformation. Although our result (Figure 12: Max-pooling *dmdt* and Figure 13: Average Pooling *dmdt*) seems good, there are some drawbacks for this method: 1. It is hard to find the optimal bin, if we change the range of the bin, the result will be different. 2. The result of the cluster only tells us how many features are located in each bin, which is hard to make results interpretable.

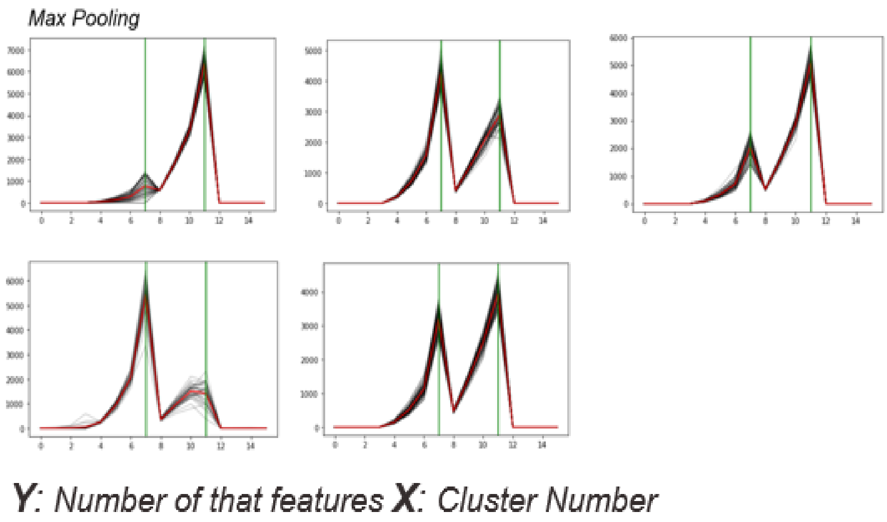

Figure 12: Max-pooling *dmdt*

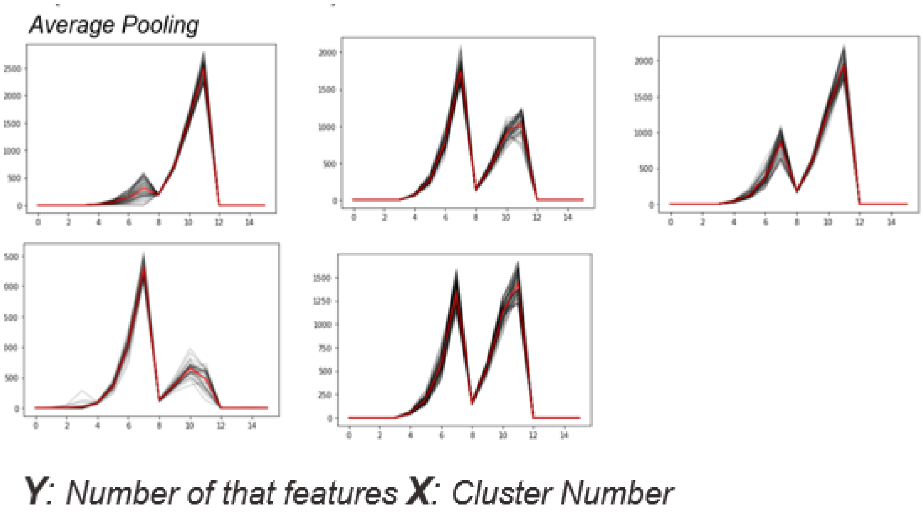

Figure 13: Average Pooling *dmdt*

3.3. Deep Geo-referenced Time Series Clustering

Based on the previous research and literature review, we designed a novel feature-based time series clustering method (Figure 14: Deep Geo-referenced Time Series Clustering). We designed three stages in this network. The first stage denoises and splits the original time series into subsequences, the second stage finds the effective latent representation for temporal clustering by minimizing mean square error, and the last stage does temporal clustering. To be more specific, in the first stage, consider a temporal sequence $T: x_1, x_2, \dots, x_n$, we firstly did a 7-day moving average to reduce the impact of some outliers and noise in original time series T, and another reason we choose 7 days is because we want to explore weekly pattern, the result is a new transformed time series and we defined it as T_{MA}. Secondly, we added a sliding window with adjusting parameters $width$ and $steps$, the result is a set of subsequences which extracted from T_{MA}. The set of subsequences can be expressed as: $T_{MA} = \{T_{MA}^1, T_{MA}^2, \dots, T_{MA}^N\}$, where N is the number of subsequences. For example, if we set the timesteps be 128, and we set $width\ =\ 8$ and $steps\ =\ 8$, there will be $\frac{128}{8} = 16$ subsequences. In the second stage, our goal is to find an informative latent representation by making use of a temporal autoencoder. The first two layers will be a 1D convolutional layer with max pooling layer, which can capture the short-term features. The next two layers will be Bidirectional-Long Short-Term Memory to learn temporal changes in two directions of time, which casts the original subsequence into a much smaller latent space. Finally, the features in latent space which retains most of the relevant information will be assigned to a clustering algorithm. In this process, our objective is to minimize loss so that latent representations can retain more useful information.

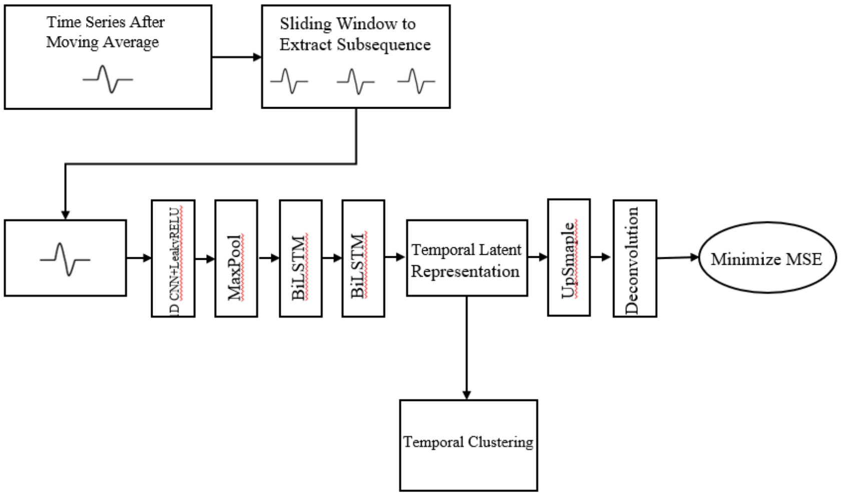

Figure 14: Deep Geo-referenced Time Series Clustering

3.4. Result and Analysis

We chose K-means as a clustering algorithm and applied DGTSC on the mobility time series for 254 counties in Texas. Mobility datasets are from the Bureau of Transportation Statistics[18], which provides trips by distance for each county. The mobility data can reflect the travel distance of the people in a specific county. The parameter of the network is: $input\ size\ =\ 8$ and output will be 2-dimensional latent representations. Then, we chose $K = 4$ by applying elbow method (Figure 15: Elbow Method to find best K=4), and (Figure 16: Cluster of Latent Representations)is the plot of the result of clustering 2-dimensional latent representations. In this plot, X and Y are both represent the mobility level, if X is greater, which means the mobility level will be at a higher level. So, from this plot, we can see there are four mobility levels: yellow cluster is the highest mobility level in the right top corner and the dark blue cluster is the second mobility level, green is the third mobility level and purple cluster is the lowest mobility level in the left bottom corner.

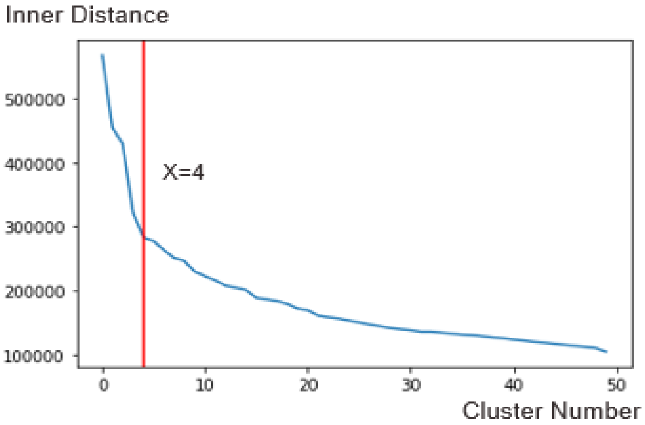

Figure 15: Elbow Method to find best K=4

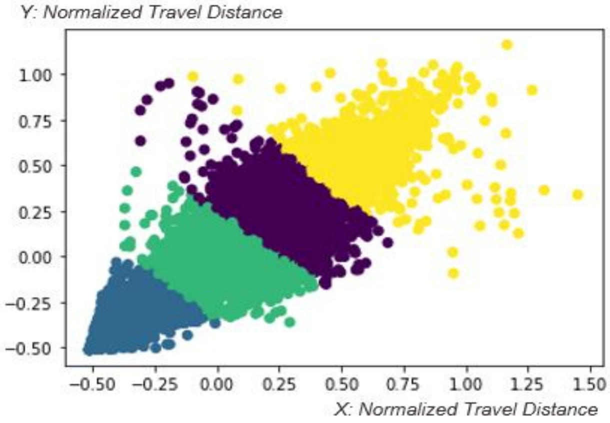

Figure 16: Cluster of Latent Representations

The result of deep geo-referenced time series clustering can also be displayed as a heatmap (Figure 17: Cluster Heatmap), where *X* represents 254 counties in Texas and *Y* represents 33 weeks we explored from Jan 1st, 2020 to the beginning of September 2020.

In the heatmap, orange is the highest mobility level and we define it as M_1, yellow is the second mobility level and we define it as M_2, blue is the third mobility level and we define it as M_3, and green is the lowest mobility level and we define it as M_4. From the heat map we can see a clear mobility pattern change that Texas had suffered, from initial mobility level M_2 to the last mobility level M_3. We plot the heatmap on the map, and the change of clusters from Jan 1st 2020 to September 2020 from Figure 18: Mar 12nd, Mobility Level to Figure 31: Sep 17th, Mobility Level can be obtained. In these figures, yellow represents the highest mobility level and we defined it as M_1, purple is the second mobility level M_2, green is the third mobility level M_3, and grey green is the lowest mobility level M_4, and we combined the policy and holiday to make the cluster be interpretable.

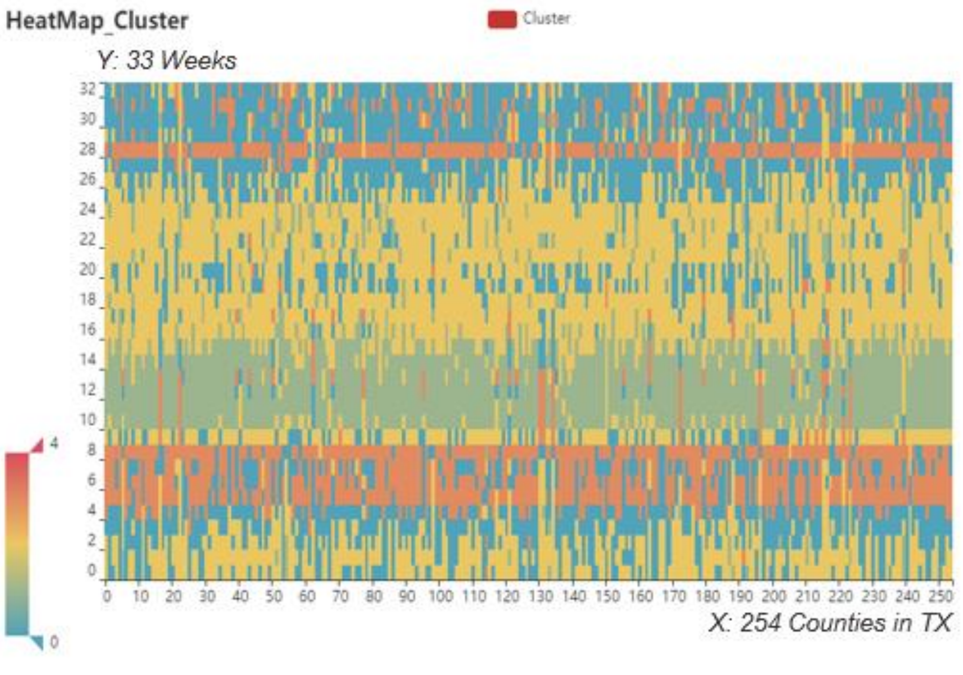

Figure 17: Cluster Heatmap

Date: 3.12 ~ 3.20

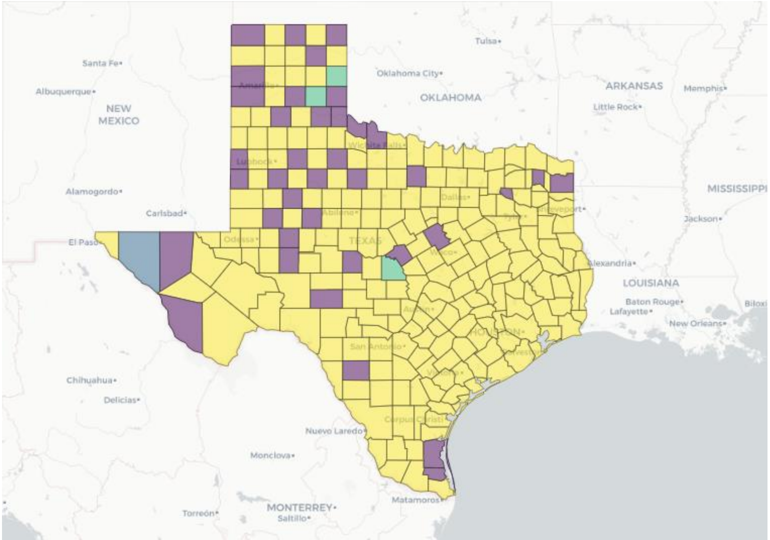

Figure 18: Mar 12nd, Mobility Level

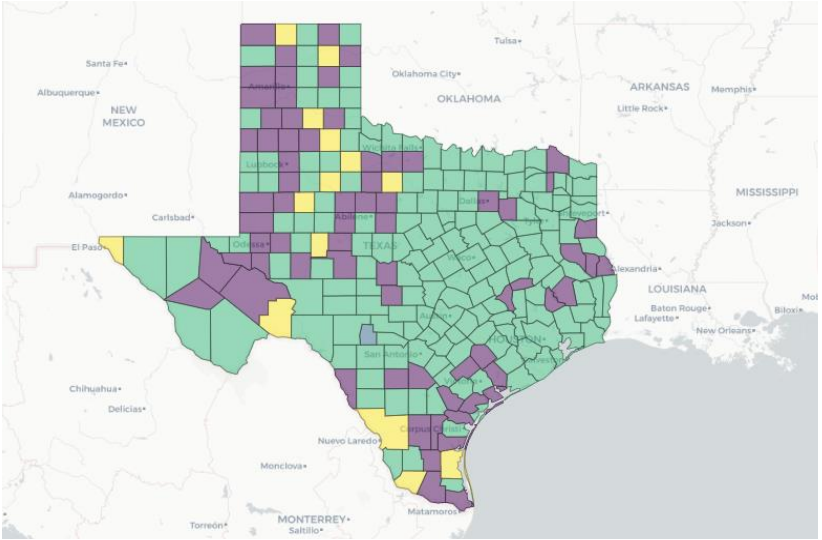

Figure 19: Mar 20th, Mobility Level

From Mar 12th to Mar 20th, we can see a jump on mobility level from M_1 to M_3 in the majority of counties in Texas, which means the mobility level meets a decrease.

According to the policy, we found the policy that state of emergency was announced on Mar 13th appears associated with this decrease.

Date: 3.20 ~ 3.28

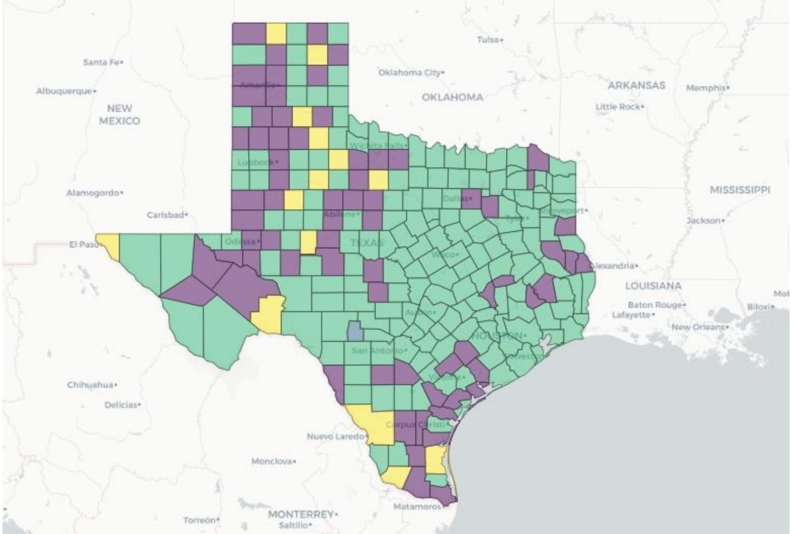

Figure 20: Mar 20th, Mobility Level

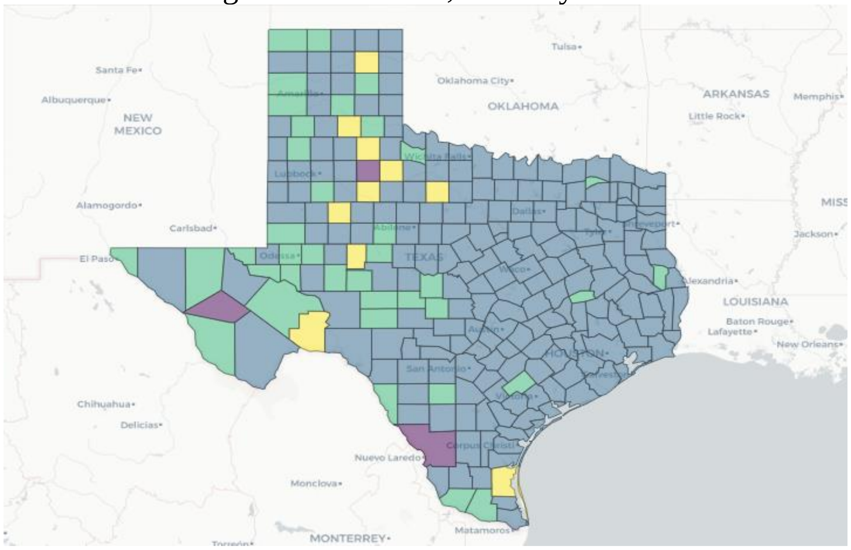

Figure 21: Mar 28th, Mobility Level

From Mar 20th to Mar 28th, we can see a decrease on mobility level from M_3 to M_4 in the majority of counties in Texas. According to the policy, we found the policy that food, drink and non-essential travel restrictions was announced on Mar 16th and appears associated with this decrease.

Date: 3.28 ~ 4.5

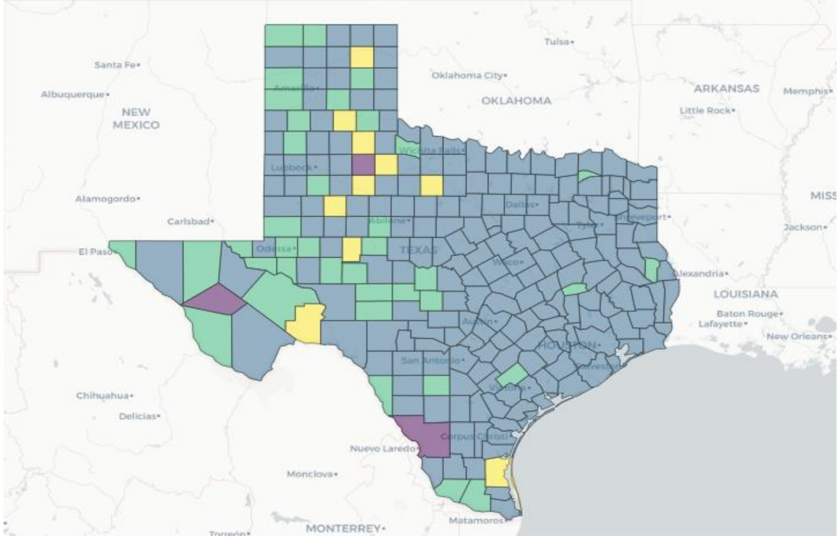

Figure 22: Mar 28th, Mobility Level

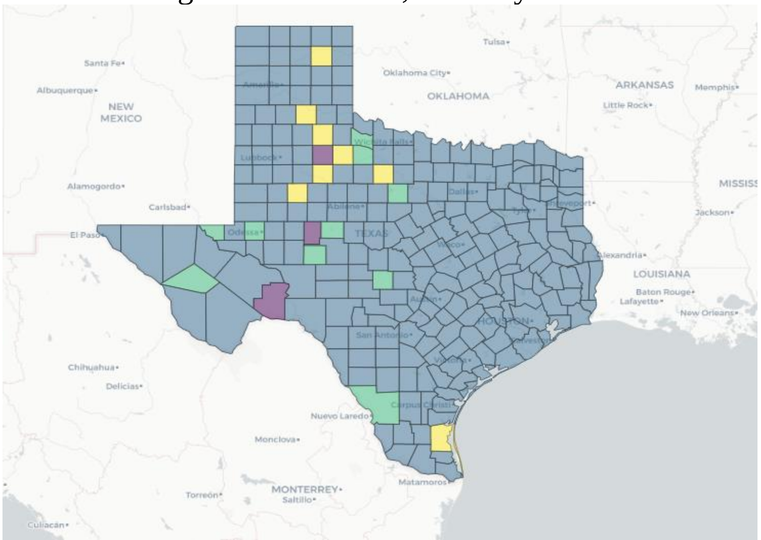

Figure 23: Apr 5th, Mobility Level

From Mar 28th to Apr 5th, we can see a continuous decrease on mobility level from M_3 to M_4 in some counties located in the west of Texas . According to the policy, we found the policy that shelter in place was announced on Mar 24th appears associated with this decrease.

Date: 4.28 ~ 5.14

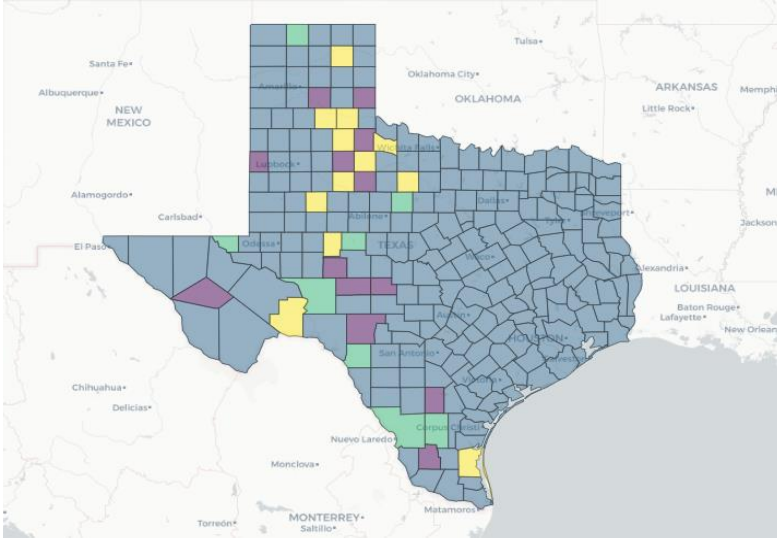

Figure 24: Apr 28th, Mobility Level

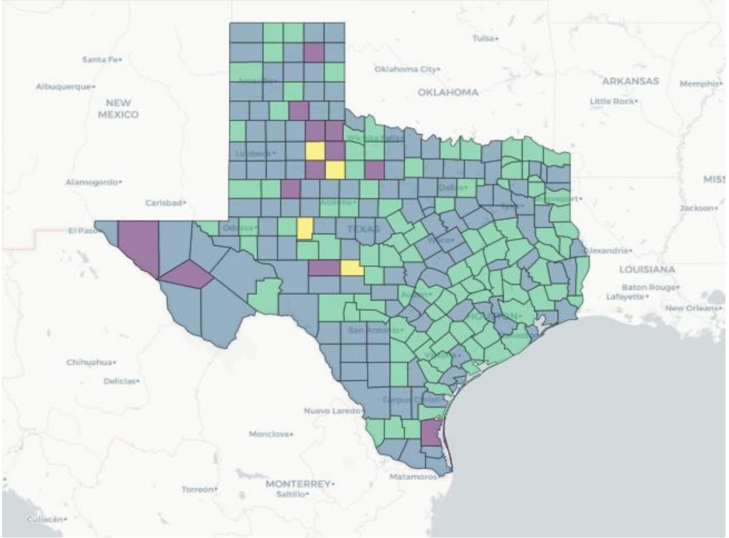

Figure 25: May 14th, Mobility Level

From Apr 28th to May 14th, we can see a sudden increase on mobility level from M_4 to M_3 in some counties located in the middle and east of Texas . According to the policy, we found the policy that a stop shelter in place was announced on Apr 30th appears associated with this decrease.

Date: 5.6 ~ 5.24

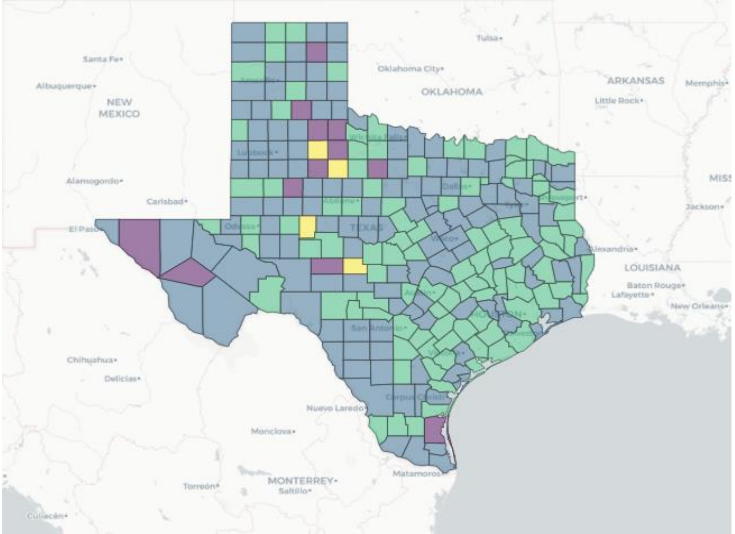

Figure 26: May 6th, Mobility Level

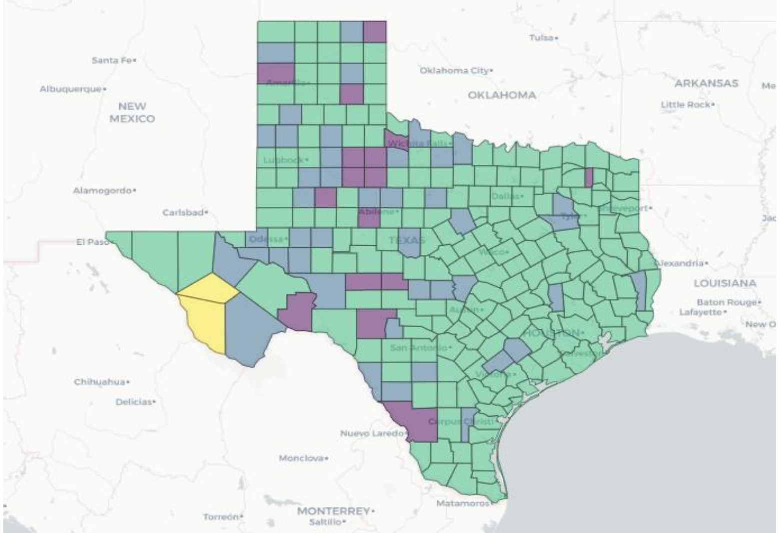

Figure 27: Mar 24th, Mobility Level

From May 5th to May 24th, we can see a continuous increase on mobility level from M_4 to M_3 in some counties located in the west of Texas . According to the policy, we found the policy that mandate face mask use by all individuals in public was announced on May 8th appears associated with this decrease.

Date: 6.23 ~ 7.1

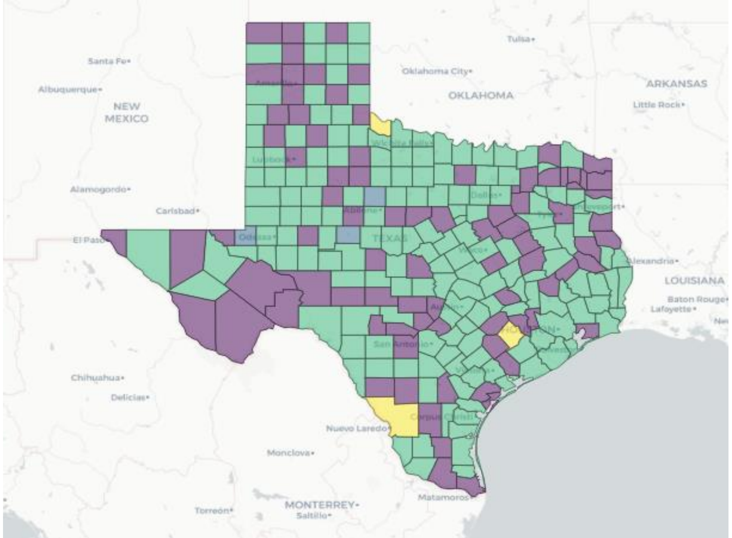

Figure 28: Jun 23rd, Mobility Level

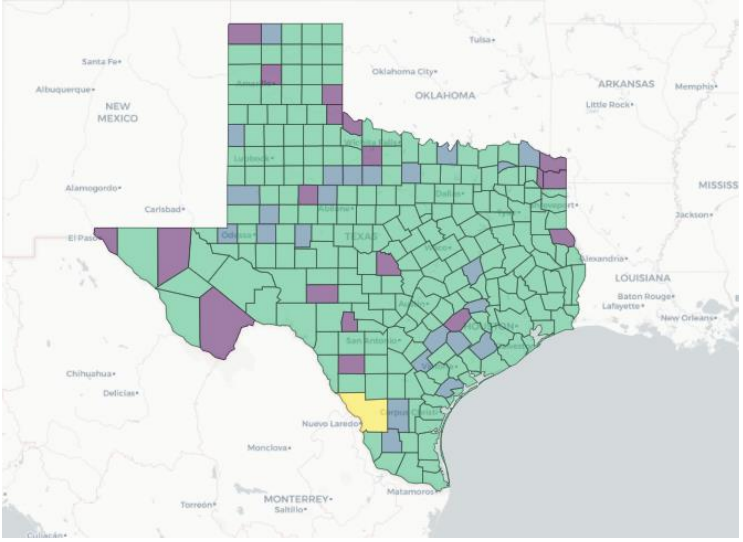

Figure 29: Jul 1st, Mobility Level

From Jun 23th to Jul 1st, we can see a decrease on mobility level from M_2 to M_3 in some counties of Texas . According to the policy, we found the policy that stops reopen in Texas and Florida was announced on Jun 26th appears associated with this decrease.

Date: 9.1~9.17

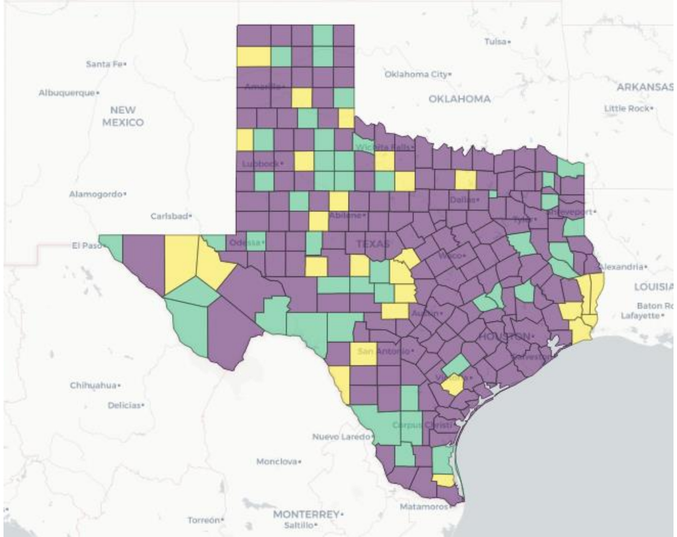

Figure 30: Sep 1st, Mobility Level

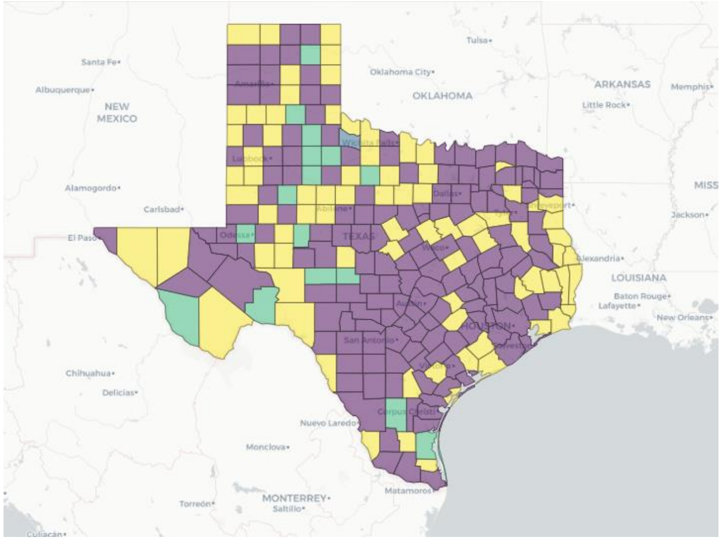

Figure 31: Sep 17th, Mobility Level

From Sep 1st to Sep 17th , we can see a increase on mobility level from M_2 to M_1 in some counties located in the west of Texas . According to the holiday, we found Sep 6th is labor day appears associated with this increase.

4. CONCLUSIONS AND FUTURE WORK

4.1. Conclusion

We make use of two strategies: modeling analysis and clustering analysis to explore the potential driven factors affecting the increase of confirmed cases. The first part of our work indicates that population density, weather, and SVI are associated with the increase of confirmed cases, and population density was the most influential factor included in the model affecting the increase of confirmed cases. We take Texas as a study area and found the spread of COVID-19 is mainly affected by population density in the majority of counties in Texas, especially for the counties located in the northeast of Texas. Meanwhile, a county with high SVI, low temperature, high wind speed will also help the spread of COVID-19. In the second part, our result indicates that the mobility level is affected by the announced policy and holidays to some extent. According to the pattern of mobility, we found that the mobility response to a policy will first show in the middle and east of Texas, and then spread to some counties located in west Texas. Besides, the mobility response of some counties located in northwest Texas was not affected like that in the middle and east of Texas.

4.2. Future Work

Although we have explored some potential driving factors and some hidden information in mobility by GTWR, STWR and time series clustering, we still have some work to do in the future:

1. For modeling analysis, we make use of Extended GWR models, but the pattern of these variables for Texas is not enough, we will continue to extend from Texas to the

whole United States in the future work.

2. For time series clustering, we also need to explore the mobility pattern of the whole United states. Another work needs to be done is we should try some different datasets and to see whether the patterns are similar so that make the result be more reliable. Besides, our network still needs to be improved, we should combine the autoencoder part with the clustering part and optimize them simultaneously, which may improve the accuracy of clustering.
3. Although many time series clustering strategies have been proposed, geo-referenced time series is different from the common time series such audio time series since Geo-referenced time series have a strong autocorrelation that we cannot ignore[19]. So, another future work we should do is to incorporate spatial autocorrelation into time series clustering networks.

5. REFERENCE

[1] Desmet, K., & Wacziarg, R. (2020). Understanding Spatial Variation in COVID-19 across the United States. https://doi.org/10.3386/w27329

[2] Maiti, A., Zhang, Q., Sannigrahi, S., Pramanik, S., Chakraborti, S., & Pilla, F. (2020). Spatiotemporal effects of the causal factors on COVID-19 incidences in the contiguous United States. In arXiv (Vol. 68, p. 102784). arXiv. https://doi.org/10.1016/j.scs.2021.102784

[3] Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of the Total Environment, 739, 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

[4] Tosepu, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. A. I., Lestari, H., Bahar, H., & Asfian, P. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of the Total Environment, 725, 138436. https://doi.org/10.1016/j.scitotenv.2020.138436

[5] Warren, M. S., & Skillman, S. W. (2020). Mobility Changes in Response to COVID-19. In arXiv. arXiv. https://arxiv.org/abs/2003.14228v1

[6] Karaye, I. M., & Horney, J. A. (2020). The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships. American Journal of Preventive Medicine, 59(3), 317–325. https://doi.org/10.1016/j.amepre.2020.06.006

[7] Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Science of the Total Environment, 728, 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

[8] Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

[9] Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically Weighted Regression-Modelling Spatial Non-Stationarity. Journal of the Royal Statistical Society. Series D (The Statistician), 47(3), 431-443. http://www.jstor.org/stable/2988625

[10] Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N., Cattarino, L., Cooper, L., Coupland, H., Cucunuba Perez, Z., Cuomo-Dannenburg, G., Dighe, A., Djaafara, A., Dorigatti, I., Eales, O., Van Elsland, S., NASCIMENTO, F., … Donnelly, C. (2021). Reduction in mobility and COVID-19 transmission. Nature Communications 2021 12:1, 12(1), 1–9. https://doi.org/10.1038/s41467-021-21358-2

[11] Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. International Journal of Geographical Information Science, 24(3), 383–401. https://doi.org/10.1080/13658810802672469

[12] Que, X., Ma, X., Ma, C., & Chen, Q. (2020). A spatiotemporal weighted regression model (STWR v1.0) for analyzing local nonstationarity in space and time. Geoscientific Model Development, 13(12), 6149–6164.

https://doi.org/10.5194/gmd-13-6149-2020

[13] Huang, X., Li, Z., Lu, J., Wang, S., Wei, H., & Chen, B. (2020). Time-series clustering for home dwell time during COVID-19: What can we learn from it? ISPRS International Journal of Geo-Information, 9(11), 675. https://doi.org/10.3390/ijgi9110675

[14] Home. (n.d.). Retrieved March 20, 2021, from https://www.safegraph.com/

[15] Bathwal, R., Chitta, P., Tirumala, K., & Varadarajan, V. (2020). Ensemble Machine Learning Methods for Modeling COVID19 Deaths. 1–10. http://arxiv.org/abs/2010.04052

[16] Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S. G., Drake, A. J., & Graham, M. J. (2018). Deep-learnt classification of light curves. 2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings, 2018-Janua, 1–8. https://doi.org/10.1109/SSCI.2017.8280984

[17] Wu, X., Zurita-Milla, R., & Kraak, M. J. (2015). Co-clustering geo-referenced time series: exploring spatio-temporal patterns in Dutch temperature data. International Journal of Geographical Information Science, 29(4), 624–642. https://doi.org/10.1080/13658816.2014.994520

[18] Trips by Distance | Open Data | Socrata. (n.d.). Retrieved March 8, 2021, from https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

[19] Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, 46, 234. https://doi.org/10.2307/143141

[20] Wang, J., Tang, K., Feng, K., Lin, X., Lv, W., Chen, K., & Wang, F. (n.d.). High Temperature and High Humidity Reduce the Transmission of COVID-19. https://arxiv.org/ftp/arxiv/papers/2003/2003.05003.pdf

[21] Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Geographical and Temporal Weighted Regression (GTWR). Geographical Analysis, 47(4), 431–452. https://doi.org/10.1111/gean.12071

[22] Islam, N., Shabnam, S., & Erzurumluoglu, A. M. (2020). Temperature, humidity, and wind speed are associated with lower Covid-19 incidence. In medRxiv (p. 2020.03.27.20045658). medRxiv. https://doi.org/10.1101/2020.03.27.20045658

[23] Nayak, A., Islam, S. J., Mehta, A., Ko, Y. A., Patel, S. A., Goyal, A., Sullivan, S., Lewis, T. T., Vaccarino, V., Morris, A. A., & Quyyumi, A. A. (2020). Impact of social vulnerability on COVID-19 incidence and outcomes in the United States. In medRxiv(p.2020.04.10.20060962). https://doi.org/10.1101/2020.04.10.20060962

[24] Feehan, D. M., & Mahmud, A. S. (2021). Quantifying population contact patterns in the United States during the COVID-19 pandemic. Nature Communications, 12(1), 1–9. https://doi.org/10.1038/s41467-021-20990-2

[25] Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. Information Systems, 53, 16–38. https://doi.org/10.1016/j.is.2015.04.007

[26] Li, X., et al., 2012. Explore multivariable spatio-temporal data with the time wave: case study on meteorological data. In: A.G.O. Yeh et al., ed. Advances in spatial data handling and GIS. Lecture notes in geoinformation and cartography, part 3. Berlin: Springer, 79–92. doi:10.1007/ 978-3-642-25926-5_7

[27] Zurita-Milla, R., et al., 2013. Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. IEEE Transactions on Geoscience and Remote Sensing, 51 (4), 1914– 1921. doi:10.1109/TGRS.2012.2223218

[28] Han, J., Kamber, M., and Pei, J., 2012. Data mining concepts and techniques. 3rd ed. Waltham, MA: Morgan Kaufman, MIT Press. https://tinman.cs.gsu.edu/~zcai/course/47406740/Slides/Chapter%201%20Introduction%20to%20Data%20Mining.pdf

[29] P. Rai, S. Singh, A survey of clustering techniques, Int. J. Comput. Appl. 7 (12) (2010)1–5. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.5219&rep=rep1&type=pdf

[30] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. 2. http://arxiv.org/abs/1003.4083

[31] Esmaeil Alizadeh – Medium. (n.d.). Retrieved April 18, 2021, from https://medium.com/@ealizadeh