Location Based Covid-19 Cases Prediction

Comparison Between Linear Regression and Regression Kriging

Trent Zhang

In this report, we're going to model the spatial structure and the relationship between population and Covid-19 cases in Illinois, Iowa, Wisconsin and Minnesota. As a result, we found the spatial model is better than simple linear regression in performance. The spatial model reached a mean score of 0.48 with cross validation.

Department of Statistics
University of Illinois at Urbana-Champaign
United States
May 8, 2022

Location Based Covid-19 Cases Prediction

Trent Zhang
Department of Statistics
University of Illinois at Urbana Champaign

May 8, 2022

Abstract

In this report, we're going to model the spatial structure and the relationship between population and Covid-19 cases in Illinois, Iowa, Wisconsin and Minnesota. As a result, we found the spatial model is better than simple linear regression in performance. The spatial model reached a mean score of 0.48 with cross validation.

1 Introduction

The recent pandemic, which is commonly known as COVID-19, is an infectious disease caused by the virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Since the start of the year 2020, the infectious disease COVID-19 has started to spread globally and resulting in almost 517 million positive cases and 6.25 deaths till today.

In this project, we are going to explore the distribution of Covid-19 cases in United States, and will also model the relationship between Covid-19 cases, geographical location and population with Python.

2 Related Works

The COVID-19 pandemic is full of unknowns, and many of them have a spatial dimension that lead to understanding the phenomenon as geographical and potentially mappable. Thus from health science, the research needs include the ability to cross variables of different kinds to interpret the COVID-19 phenomenon, its spatial analysis and spatiotemporal dimensions, its geographical impact on decision-making and everyday life, and predictive modelling of the evolution of the disease.

In the paper of Kang et al., the authors explored the spatial epidemic dynamics of COVID-19 in mainland China[4]. Moran's I spatial statistic with various definitions of neighbours was used to conduct a test to determine whether a spatial association of the COVID-19 infections existed. The spatial spread of

the COVID-19 pandemic in China was observed. The results showed that most of the models, except medical-care-based connection models, indicated a significant spatial association of COVID-19 infections from around 22 January 2020.

In the paper of Slater et al., they investigated the efficacy of using telecommunication derived mobility data to induce spatial dependence in spatial models applied to two Spanish communities' COVID-19 case counts[7]. They do this by extending Besag York Mollié (BYM) models to include both a physical adjacency effect, alongside a mobility effect. The mobility effect is given a Gaussian Markov random field prior, with the number of trips between regions as edge weights. They leverage modern parametrizations of BYM models to conclude that the number of people moving between regions better explains variation in COVID-19 case counts than physical proximity data. They suggest that this data should be used in conjunction with physical proximity data when developing spatial models for COVID-19 case counts.

From the related works, we can see that the use of geo-spatial and statistical tools has become particularly relevant with the declaration of COVID-19 as a global pandemic.

3 Data

3.1 Data Source

Our data is collected from CDC[1][1]. The table 1 shows 2 data example in the original dataset.

Variable Name	Data 1	Data 2
county	American Samoa	Guam
county fips	60000	66000
state	American Samoa	Guam
county population	47,392	168,489
health service area number	901	902
health service area	American Samoa	Guam
health service area population	47,392	168,489
covid inpatient bed utilization	0	15.2
covid hospital admissions per 100k	2.1	38.6
covid cases per 100k	156.14	833.88
covid-19 community level	Low	High
date updated	3/3/22	3/3/22

Table 1: Data Example

From the original data table, we can see that county, county fips, state, county population, health service area number, health service area, covid inpatient bed utilization, covid hospital admissions per 100k, covid cases per 100k,

[1]United States COVID-19 Community Levels by County, link

covid-19 community level and date updated are given. In this project, we only use the data in the time period of 2022-04-22 to 2022-05-05, and the attributes of county, state, county population, and covid cases per 100k for the following analysis. The total data here includes 3224 counties.

3.2 Data Transformation

Since the original location data is text based, in order to make numerical analysis, we need to encode the text location into geographical coordinates. Here, we scraped data from Open Street Map to get longitude and latitude of each county in United States[2].

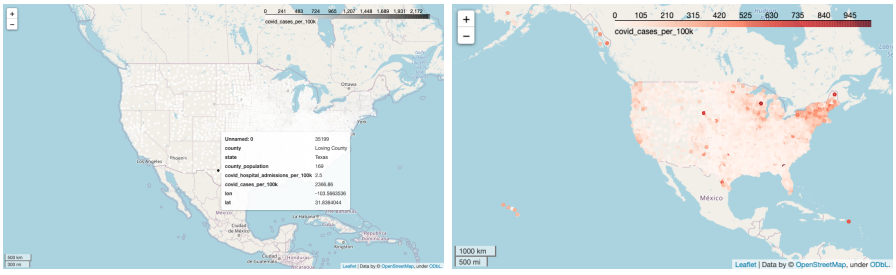
Open Street Map is a collaborative project to create a free editable geographic database of the world. The geographical data underlying the maps is considered the primary output of the project. During data scaping, some addresses are failed in search of geographical coordinates, most of them occurred in Puerto Rico and Alaska. In the end, we successfully queried geographical coordinates for 3138 counties, and we only use those data-points to continue analysis.

3.3 Spatial Data Framework

We used GeoPandas Data Structures to analysis our data. GeoPandas is an open source project to make working with geospatial data in python easier[3]. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types, geometric operations are performed by shapely and Geopandas further depends on fiona for file access and matplotlib for plotting.

3.4 Preliminary Visualization

After getting the coordinates data from Open Street Map, we visualized and covid cases per 100k as in fig. 1 . During our visualizations, an abnormal data point in texas was found, we delete that row 35199, and visualize the rest 3137 data points.



(a) Ploy With Abnormal Data (b) Plot Without Abnormal Data

Figure 1: New Cases Per 100k from 2022-04-22 to 2022-05-05

3.5 Vaccination Effect

As we all agree, vaccination coverage plays a great role in decreasing the Covid-19 cases. Since vaccination can greatly affect Covid-19 cases and vaccination coverages are significantly different between different states. Therefore, the model will be unreliable if we didn't consider vaccination rate as a influential factor.

However, The vaccination data at CDC are missing in many counties, thus requiring more time to clean the data. Due to the time limitation, we don't consider vaccination as an influential factor here.

In order to make vaccination effect ignorable, we only analyze the data in Illinois, Iowa, Wisconsin and Minnesota, which have similar vaccination coverage according to fig. 2 from CDC [2]. The final data points visualization is as fig. 3.

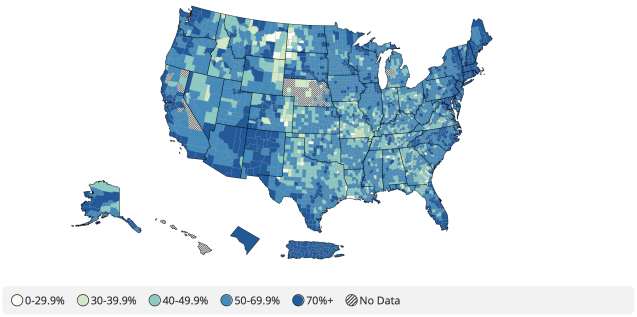


Figure 2: Vaccination Coverage

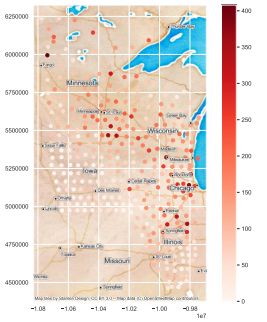


Figure 3: Data Points

4 Methods

4.1 Simple Linear Regression

In fig. 1, we first visualized the relationship between county population and covid cases per 100k, however, there seems no linear relationship between them. Then we visualized the relationship between log county population and covid cases per 100k, then we can notice increasing trend in our plot. Therefore, we use the log county population as our variable.

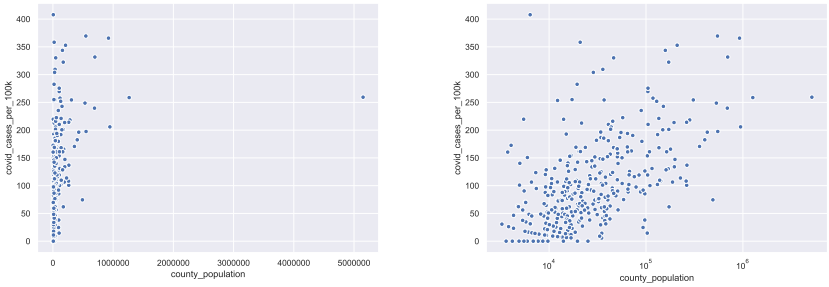
Besides, we also normalized all the data before linear regress our model, therefore, there will be no intercept term in our model. The simple linear regression model is as:

$$y = x\beta + \epsilon \tag{1}$$

where:

- y: covid cases per 100k

[2] From Centers for Disease Control and Prevention, link



(a) Without Logarithm (b) With Logarithm

Figure 4: Relationship Visualizations

- x: log county population
- β: coefficient of log county population
- ϵ: random gaussian noise

Here we used Scikit-learn Framework in python to implement simple linear regression[6]. Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

4.2 Regression Kriging

The simple linear regression didn't capture the spatial dependence between different counties. In this section, we are going to talk about Regression Kriging which can model the spatial dependence between locations.

Here we used GeoStatTools in GeoStat Framework in python to implement regression kriging[5]. GeoStatTools is a library providing geo-statistical tools like kriging, random field generation, variogram estimation and user-defined covariance models.

4.2.1 Model Form

Regression kriging is an implementation of the best linear unbiased predictor (BLUP) for spatial data, i.e. the best linear interpolator assuming the universal model of spatial variation. This model proposed that a value of a target variable at some location can be modeled as a sum of the deterministic and stochastic components:

$$y(s) = m(s) + \varepsilon'(s) + \varepsilon'' \tag{2}$$

- $y(s)$ is the target value at location s;
- $m(s)$ is the fitted deterministic part;
- $\varepsilon'(s)$ is the interpolated residual at location s;
- ε'' is the remaining residual.

In our case, the deterministic part is $x(s)\beta$, by combining the two approaches, we obtain:

$$y(s) = x(s)\beta + \varepsilon'(s) + \varepsilon'' \tag{3}$$

In order to estimate the interpolated residual at location s, which is $\varepsilon'(s)$, we introduce Kriging here. Kriging predicts the value of a function at a given point by computing a weighted average of the known values of the function in the neighborhood of the point. Depending on the stochastic properties of the random field and the various degrees of stationarity assumed, different methods for calculating the weights can be deduced, i.e. different types of kriging apply. There are 2 classical methods:

1. Ordinary kriging assumes constant unknown mean only over the search neighborhood of s_0.
2. Universal kriging assumes a general polynomial trend model, such as linear trend model $E\{y(s)\} = \sum_{k=0}^{p} \beta_k f_k(s)$.

4.2.2 Variogram models

The theoretical variogram $2\gamma(s_1, s_2)$ is a function describing the degree of spatial dependence of a spatial random field or stochastic process $y(s)$. The semivariogram $\gamma(s_1, s_2)$ is half the variogram.

In the our case, a variogram will give a measure of how much two samples taken from the area will vary in Covid cases depending on the distance between those samples. Samples taken far apart will vary more than samples taken close to each other.

In this project, we consider the following 4 variogram models:

1. Linear
$$\rho(s) = \begin{cases} 1 - \sigma \cdot \frac{s}{\ell} & s < \frac{\ell}{\sigma} \\ 0 & s \geq \frac{\ell}{\sigma} \end{cases}$$
2. Power
$$\gamma(s) = \sigma^2 \left(1 - \exp\left(-\sigma \cdot \frac{s}{\ell}\right)\right) + n$$
3. Gaussian
$$\gamma(s) = \sigma^2 \left(1 - \exp\left(-\left(\sigma \cdot \frac{s}{\ell}\right)^2\right)\right) + n$$
4. Spherical
$$\rho(s) = \begin{cases} 1 - \frac{3}{2} \cdot \sigma \cdot \frac{s}{\ell} + \frac{1}{2} \cdot \left(\sigma \cdot \frac{s}{\ell}\right)^3 & s < \frac{\ell}{\sigma} \\ 0 & s \geq \frac{\ell}{\sigma} \end{cases}$$

5 Results and Discussion

5.1 Best Model

In order to evaluate the prediction accuracy, we use coefficient of determination to evaluate our models, and we use 10 fold cross validation for model selection, the result is as fig. 5, fig. 6.

From fig. 5, we can see that Simple regression has low score around 0.1, however, the regression kriging model has a much higher score around 0.5. This means that there are indeed spatial dependence structure in our data, and it can greatly improve the accuracy of prediction.

From fig. 5, we can see that among all variograms and kriging methods we tried, the Ordinary Kriging with Spherical variogram works the best, with mean score of 0.48.

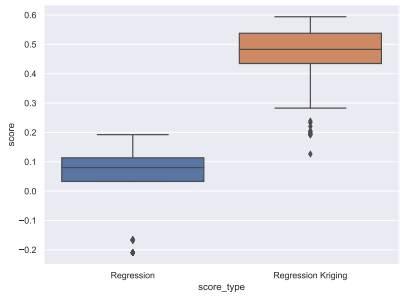


Figure 5: Simple vs Kriging

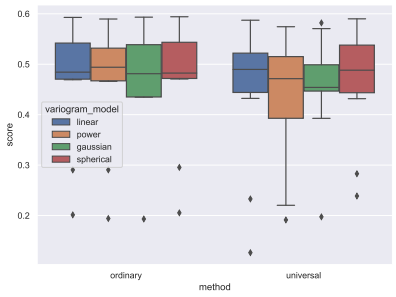


Figure 6: Regression Kriging

5.2 Prediction

Here, we will show the prediction result in Illinois, Iowa, Wisconsin and Minnesota areas. First, we need to create a prediction grid in this area, a uniformly distributed grid of 100 by 100 is created as in fig. 7.

After the grid is created, for each point on grid, we use the population its closest county as its population. Then we use our selected model to predict on the grid data points. The predicted result is as fig. 8, the background red color is the predicted value on the grid, and the red dot is the actual value from our training data.

From the plot, we can see that the dots in the places which have peak Covid-19 cases are more observable than others, which means that our model is bad at predicting extreme values.

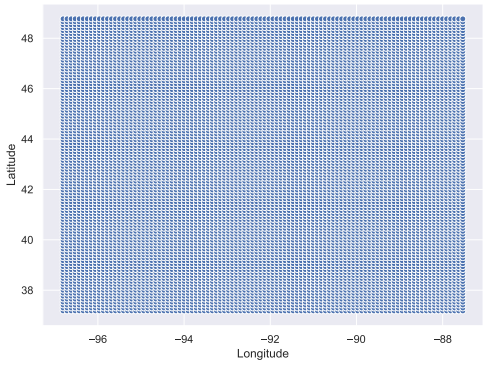


Figure 7: Prediction Grid

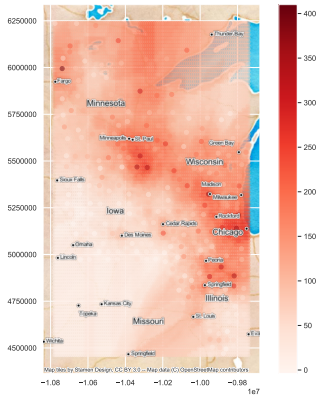


Figure 8: Prediction

6 Conclusion and Futurework

In this report, we have modeled the spatial structure and the relationship between population and Covid-19 cases in Illinois, Iowa, Wisconsin and Minnesota. As a result, we found the spatial model is better than simple linear regression in terms of performance. The simple linear regression only reach a score of 0.1, and the spatial model reached a mean score of 0.48 with cross validation. From the result, we can see that there are indeed spatial dependence between the Covid-19 cases in different counties, and there is also a weak relationship between Covid-19 cases and population.

However, there are still some limitations within our model.

- We did not using all data, we only considered Illinois, Iowa, Wisconsin and Minnesota.
- Due to time and data limitation we didn't consider other influential variables, such as age group, vaccination rate,
- We should try more deterministic models with the regression kriging model, such as tree methods, support vector regression etc.
- When we acquire the geographic coordinates, some adrees failed, we should find another way to get our data, such as Google Cloud Map API.
- There are some out-liners in our data, when modeling the deterministic part, we should try some robust regression methods

In future-works, we should focus on the limitations of our model and trying to get a better result.

References

[1] Centers for Disease Control and Prevention. United states covid-19 community levels by county. *Centers for Disease Control and Prevention*, May 2022.

[2] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

[3] K Jordahl. Geopandas: Python tools for geographic data. *URL: https://github. com/geopandas/geopandas*, 3, 2014.

[4] Dayun Kang, Hyunho Choi, Jong-Hun Kim, and Jungsoon Choi. Spatial epidemic dynamics of the covid-19 outbreak in china. *International Journal of Infectious Diseases*, 94:96–102, 2020.

[5] Sebastian Müller, Lennart Schüler, Alraune Zech, and Falk Heße. Gstools v1. 3: A toolbox for geostatistical modelling in python. *Geoscientific Model Development*, 15(7):3161–3182, 2022.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[7] Justin J Slater, Patrick E Brown, Jeffrey S Rosenthal, and Jorge Mateu. Capturing spatial dependence of covid-19 case counts with cellphone mobility data. *Spatial Statistics*, page 100540, 2021.