Analysis of Geographic, Demographic, and Socioeconomic Factors Impacting COVID-19 Case Rate in China's Hubei Province

Cindy Chen
cchenus12@gmail.com

Abstract

In this study, we investigate the relationships between COVID-19 case rate and one geographic factor (distance from Wuhan), three demographic factors (population density, registered-resident population ratio, and urban-rural population ratio), and three socioeconomic factors (average income, urban-rural income ratio, and GDP per capita) in 92 counties of Hubei Province, China. Through linear regression modeling of log-log relationships, we investigated each factor's association with COVID-19 case rate. Our results indicated that case rate was most strongly associated with a county's distance from Wuhan (R-squared = 0.601). The best model resulted from the combined factors of distance, population density, registered-resident population ratio, and urban-rural population ratio (R-squared = 0.770). Counties located close to Wuhan tend to be more urban, are more densely populated, and have more people traveling in and out. The increased frequency of contact between people results in higher case rates of COVID-19.

Introduction

The outbreak of the novel coronavirus (SARS-CoV-2) and the disease it causes (COVID-19) was first reported in Wuhan, the capital of China's Hubei Province, in December 2019. A lockdown of Wuhan and other cities in Hubei Province was implemented on January 23rd, 2020. By the time the lockdown in Wuhan was lifted on April 8th, 2020, 1.4 million cases of coronavirus had been reported worldwide, with 82,809 cases in China and 67,803 of these cases within Hubei Province.

Our study focuses on the spread of COVID-19 in 92 counties of Hubei Province from January 2020 until April 2020 (Fig. 1). We investigate the relationship between COVID-19 case rate and geographic (distance from Wuhan), demographic (population density, registered-resident population ratio, urban-rural population ratio), and socioeconomic (average disposable income, urban-rural income ratio, GDP per capita) factors.

Fig. 1: *Map of Hubei Province*

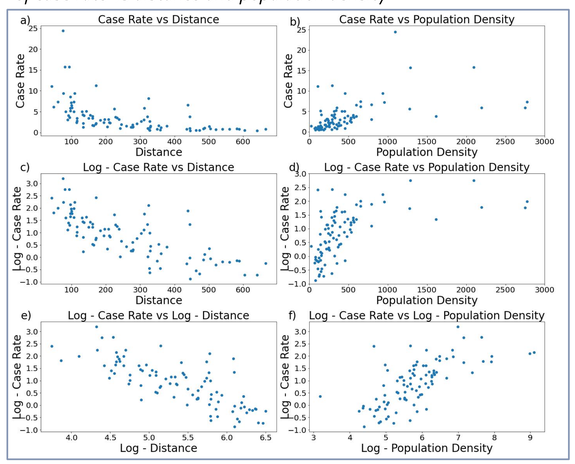
Methodology

COVID-19 case count in the 92 counties of Hubei Province as of March 31st, 2020 was reported by each prefecture's Health Commission. Data for registered population, resident population, urban-rural population ratio, GDP, and average disposable incomes were compiled from the Hubei Provincial Bureau of Statistics, City Population, Zaker, and Baidu Maps.

We chose to use the resident population in calculations requiring a population such as case rate, population density, and GDP per capita. We also divided registered population by residential population to get a ratio measuring movement within the population: a value greater than 1 suggests that more people were leaving the county, and a value less than 1 suggests that more people were entering and staying in the county. We found the ratio between the average urban and rural disposable incomes to measure income disparity.

We graphed each factor on the x-axis with case rate on the y-axis using Matplotlib in Python (Fig. 2a and 2b). The associations were non-linear, so we attempted log-linear visualization by taking the logarithm of case rate (Fig. 2c and 2d). The trends were still non-linear, so we completed log-log visualization by taking the logarithm of each factor. In the log-log graph, some factors appeared to show a linear relationship with case rate, like distance and population density (Fig. 2e and 2f).

Fig. 2: *a, b) Linear graphs; c, d) Log-linear graphs; e, f) Log-log graphs of case rate vs distance and population density*



We used ordinary least squares regression modelling in Python for each factor and combinations of factors. We compared and decided which models had better quality of fit with R-squared scores and adjusted R-squared scores. We also considered the intercept, coefficients, and coefficients' p-values. The coefficients indicated whether the relationship was positive or negative, and the coefficients' p-values were used to measure how significant the relationships were for each factor.

Results

Table 1: *Results of log - case rate vs log - individual factor models*

Factor	R-squared	Adj. R-squared
Log - Distance	0.601	0.596
Log - Population Density	0.454	0.448
Log - Registered-Resident Ratio	0.130	0.120
Log - Urban-Rural Ratio	0.161	0.152
Log - Income Ratio	0.079	0.069
Log - Average Income	0.299	0.292
Log - GDP Per Capita	0.115	0.105

Fig. 3: *a) Distance b) Population density c) Registered-Resident ratio d) Urban-rural population density models vs actual data*

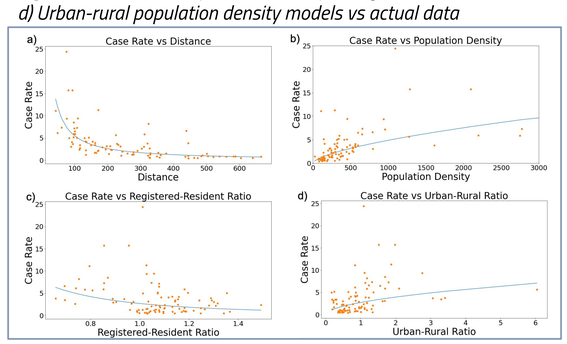
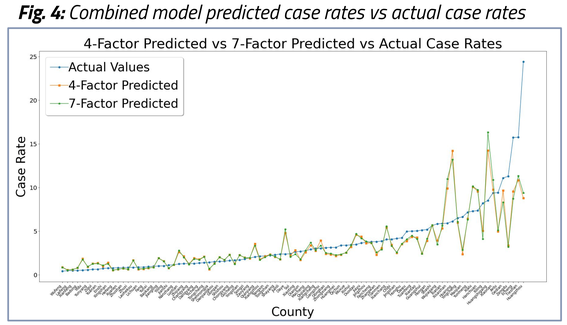


Table 2: *Results of log - case rate vs log - combined factor models*

Factors	R-squared	Adj. R-squared
Log - Distance Log - Population Density Log - Registered-Resident Ratio Log - Urban-Rural Ratio	0.770	0.759
Log - Distance Log - Population Density Log - Registered-Resident Ratio Log - Urban-Rural Ratio Log - Income Ratio Log - Average Income Log - GDP Per Capita	0.775	0.757

Fig. 4: *Combined model predicted case rates vs actual case rates*



Discussion

Association between factors and case rate:

- Distance: strong negative association
- Population density: strong positive association
- Registered-resident ratio: negative association
  - Countries with more people entering and staying (ratio less than 1) had higher case rates
- Urban-rural population ratio: positive association
  - Countries with more urban population (ratio greater than 1) had higher case rates
- Average income: negative association
- Urban-rural income ratio: weak association
- GDP per capita: weak association

Overall, the strongest model included distance and the three demographic factors. We suspect that these factors are all related—people are more likely to leave rural areas in favor of entering more urban areas in search of employment and better opportunities, and these urban areas are likely to be denser and closer to Wuhan. This influx of people entering and staying would lead to higher case rates in the counties that are closer to Wuhan, more dense, and more urban.

For socioeconomic factors, we believe the positive association between case rate and income may be because China's urban areas tend to have higher median incomes in addition to higher population density and closer proximity to Wuhan, as discussed previously. Additionally, those with higher income tend to have the means to travel more and commute to larger urban cities more often. One limitation of our study is the lack of data on the specific case rates for different income brackets, making it difficult to definitively make a conclusion on the relationship between income and case rate.

Although we lacked data on travel patterns, our study suggested a connection between case rate and frequency of people's movement. A point of discussion to consider is the possible different trends if Wuhan had not gone under lockdown on January 23rd. Wuhan is the most populous capital city of Hubei Province and a major transportation hub. if people were allowed to leave and enter Wuhan for the Lunar New Year holidays in late January, we suspect that people who had left to work and live in Wuhan would return home, causing the case rate to be higher overall. A second point of relevant discussion is the impacts of social distancing. We did not collect data on how well social distancing was enforced in each county; however, it is evident from our results that counties with less dense populations and less movement of people in and out had lower case rates. Therefore, we believe our results support the conclusion that social distancing and shelter-in-place or lockdown orders are necessary steps to reduce the case rate of COVID-19.