ADS Fundamental Project

Estimating Passenger Volume of High-speed Rail Network in China

Yunong Cao, Boya Yu, Tengfei Zheng, Jiaxu Zhou

Dec. 12. 2015

1. Introduction

1.1 Problem Addressed

As of December 2014, China had become the country that owned the longest high-speed rail (HSR) network in the world with over 9,900 mi of track in service which was more than the rest of the world's HSR rail tracks combined. What's more, there was still a length of 10,423.8 mi under construction or in planning to meet the huge traffic demand brought by the result of the rapid urbanization in China.[1] In order to maximize the transport capacity, it is absolutely significant to predict the passenger volume of each line that planned to construct.

In this project, we used the population data and the information about the setting of HSR stations and tracks to perform a network analysis on the HSR system. Our goal was to build a gravity model of the network so that we could estimate the passenger volume based solely on the information we had.

1.2 Logistics of the report

In this report, we introduced the source and the pre-processing process on our data in section 2. The methodology of our analysis was then explained in section 3. Specifically, we went through our assumptions as well as the theoretical background with respect to our model. Next, in section 4, we visualized our model and discussed its potential application. Last but not least, we wrapped up the result and talked about its application in section 5. A few recommendations were also given for the future work in this section.

2. Data

2.1 Sources of the Data

There were generally three types of data used in this project: the setting of track network, the population data and the geographical information of the high-speed train stations that involved. Information about the track network could be found nearly everywhere on the Internet. Since all these information were almost the same, we just took the one on Wikipedia and put it into a Python dictionary by manual input. The population of cities involved was taken from China 2010 Census and the population was counted based on the administration district of those cities. The geographical information (latitude and longitude) of the HSR train stations was provided by ACME Mapper and again, we manually imported them into Python.

2.2 Data Cleaning

Since both the information about the track network and geographical location of train stations were imported manually into Python, there was no concern of the cleanliness of such data. However, due to the fact that there are many homophones in Chinese characters, numerous duplicate city names resulted when we were converting Chinese characters into alphabetic characters. For example, “宜春” and “伊春” are two different cities in China, but they have the same expression “yichun” with solely alphabetic characters, which made it hard for us to differentiate them. Such phenomena occurred so frequently that adding the name of province after the name of city did was not able to solve the problem. Our solution was to add number after these names. (For example, yichun1 for “宜春” and yichun2 for “伊春”)

The China 2010 Census data was a bit dirty for two reasons. First, there were a lot of missing and deviant values in the dataset. Second, the values of population were

tabulated as string value instead of integer and symbols like comma sometimes existed in these values, too. Therefore, we filtered out all the missing or abnormal values first and then we removed all symbols and converted the value type to integer.

3. Method

3.1 Assumptions

In this analysis, we made two assumptions so as to simply the problem. The first one is that only stations in big cities were considered in the model. Since we wanted to focus our model on predicting the passenger volume among "major" stations, we only picked those cities on the HSR lines with a population over 3 million people. Another reason for doing that was due to the limitation of census data we got because the census data only provided us with the precise population information of "major" cities.

The second assumption is that he travel distance of the two stations can be represented by their geographical distance. Because we could not find any data that had the exact rail length between each station, the distance between each station is actually calculated from their geographical location (longitude and latitude). In other word, we used displacement to represent the real distance during this project.

3.2 Model Specification

Following steps were conducted to build our model. First, train stations were added as nodes in the network. Node attributes included the position of the station and the 2010 population of the administrative district which stations were located in. Next, edges were added to the graph according to the track network. The distances of edges were defined as the geodesic distance between the two cities. Then, a gravity model for

estimating passenger volume could be made based on the population, distance and the track network.

We applied the following model to estimate the passenger volume between two cities a and b. In this model, we assumed that the passenger volume was proportional to the population of each city and inversely proportional to the distance between the two cities. The distance was the length of the corresponding shortest path and K was a constant for normalizing the result.

$$e(a,b) = K\frac{w(a)w(b)}{d(a,b)}$$

Using this model, we calculated the passenger volume of high-speed train between any two cities. One should notice that a and b here were not necessarily the endpoints of an edge but could be any chosen pair of cities in the graph.

The weight of all edges was set to be 0 in the beginning. Then, for any two different cities a and b, the shortest path p (by Dijkstra algorithm) and the passenger volume $e(a,b)$ were calculated. After that we added $e(a,b)$ to all intervals on the shortest path. After such operations on all city pairs, we acquired a weight for all edges. This weight essentially showed the passenger flow on this interval. In other word, it represented the number of passengers who took the high-speed trains passing through this railway interval.

4. Results and Discussion

Table 1 shows the top 10 cities with largest pagerank centrality. Among those 10 cities, 9 of which are provincial capital cities or municipalities, indicating the high-speed railway network is being constructed surrounding and connecting major cities.

1	Zhengzhou:0.0149353465567
2	Wuhan:0.0138473966033
3	Shenyang:0.0135335579247
4	Tianjin:0.0104272392023
5	Anshan:0.0101848458507
6	Chongqing:0.0100772320646
7	Nanchang:0.00956853296629
8	Changchun:0.00956105732841
9	Shijiazhuang:0.00919587505778
10	Guangzhou:0.00904183322921

Table 1: Top 10 cities with largest pagerank centrality

Figure 1 shows the high-speed network within Yangtze River Delta and Yangtze River Downstream Plains while Figure 2 visualize the overall HSR network with passenger volume predicted. Sizes of nodes indicate the population of cities while widths of edges indicate the passenger flow in the interval calculated by the gravity model.

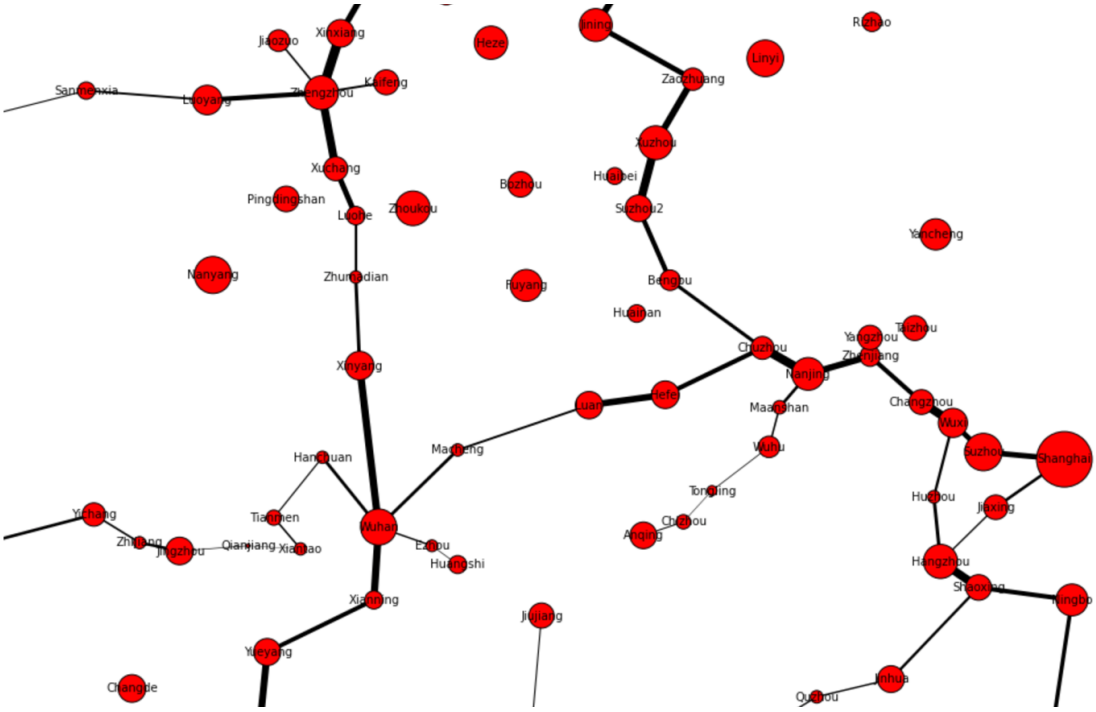

Figure 1: High-speed network within Yangtze River Downstream Plains

Figure 2: Overall HSR Network

Extracting the passenger flow value from the model, Table 2 shows the top 10 pairs of cities with highest passenger flow. It is obvious that most of these cities are located in the populous area.

City 1	City 2	Flow
Tangshan	Tianjin	249105
Tianjin	Cangzhou	239057
Tianjin	Beijing	225324
Baoding	Beijing	200228
Suzhou2	Xuzhou	142982
Taian	Jining	131345
Cangzhou	Jinan	128379
Yueyang	Changsha	122513
Taian	Jinan	109201
Wuhan	Xinyang	109150

Table 2: Top 10 pairs of cities with highest passenger flow

Because the passenger volumes on all edges of the network were estimated by our model, the performance of the model could be evaluated by comparing the estimated

passenger volume to the actual one. However, China Railway Company (CRC) did not release such data to the public and therefore we cannot evaluate the prediction power of during the current stage. If we can somehow get access to that data and the performance of the model turns out to be satisfying, the model can then be used for predicting the volume of passenger transport on the lines which are under planning. In this way, the model would be able to help maximize the capacity by just knowing some simple statistics such as distance and population.

5. Conclusions and Recommendations

In this analysis, we visualized the HSR network and built a gravity model which was able to predict the passenger volume on various railway lines. Nonetheless, we could not evaluate its performance since the real passenger volume was not available during the current stage.

There were several improvements that can be made to the model. First, one can complement the HSR network by adding those stations which located in cities with a population of 3 million or lower in that we just considered the “major” stations in this analysis. Second, one could try to find the exact distances between each two stations in a railway line instead of using the geodesic distance.

6. References

1. "China Has World's Largest High-speed Railway Network." January 30, 2015. Accessed December 13, 2015