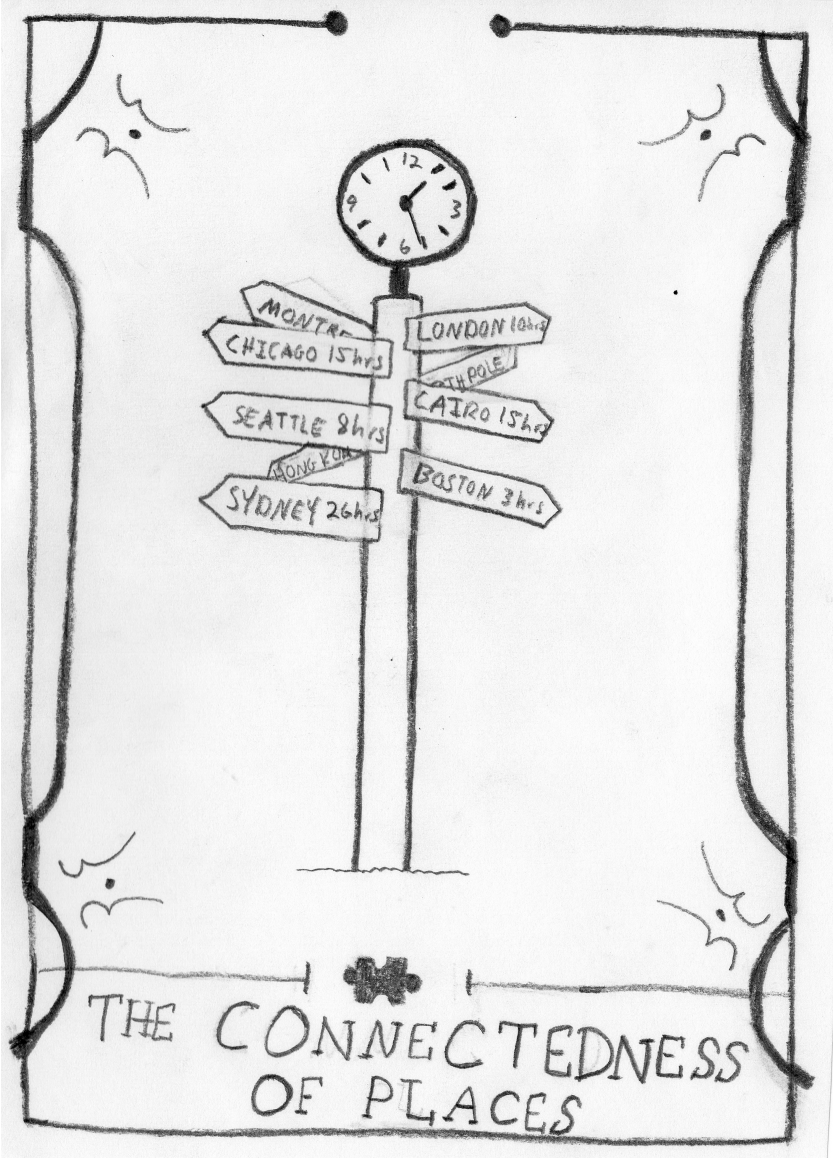
CSYS 300 Term Project:

The Connectedness of Places



Austin W. Thomas

17 December 2015

Abstract

Conventional maps depict places in relation to one another in terms of physical space. People often express their location relative to another location not in terms of distance, however, but in terms of travel time. This report explores the relative connectedness of places by using the travel time between points as the unit of distance. Closer places will be ones which can be reached faster, rather than ones which are physically closer. Postal codes stand in for places to simplify location pairing and time calculation. Two case studies, one for the State of Vermont and one for New York City, are constructed.

1 Introduction

Cartography and mapping are age old disciplines, borne out of a desire to organise and quantify one's surroundings. Maps serve to guide users and provide context for one's surroundings. Today, many people in the developed and developing world enjoy access to vast, interactive mapping databases at a moment's notice. Perhaps even more astonishing, many mapping services are made available free of charge at the point of access (that is, the maps themselves are freely available while the data transmission may or may not be). Given the maturity of mapping technologies and information access, it is no surprise that maps from large vendors such as Google, Microsoft, and MapQuest (alongside collaborative sources like OpenStreetMap) offer details beyond the typical roads, highways, rail lines, water ways, mountains, and other physical features. Users are now offered additional integrated services such as business names and trading hours, points of interest, and real time directions. It is on the latter of these examples that this report will focus. Perhaps the most powerful underlying feature of modern mapping technologies is the on-demand point-to-point routing information. A service like Google Maps can provide travel times, distances, and, in some regions, alternative modes of transport to the automobile. This report will leverage this service from Google Maps to find the travel times between ZIP codes within the State of Vermont and within New York City.

Globalisation has drawn physically disparate locations together in ways which were unimaginable just a century or two ago. Journeys between continents and hemispheres now take hours instead of weeks or months thanks to aeroplane travel. The ability of a person or a group of people to interact with other populations, other markets, even other ecosystems, is driven these days not by where you are physically located, but who you are connected to and how you are connected to them. The goal of this report is to explore the connectedness of places expressed not by physical distance, but instead by time 'distances'. Section 2 outlines the datasets used and how they were developed, section 3 visualises the data and presents some rudimentary analysis, and section 4 summarises the findings and where the next steps might be taken.

2 Data Acquisition

For the purposes of this project, ZIP codes (truncated to five digits) are used to define places between which travel times were recorded. A five-digit ZIP code offers a good compromise between location specificity and speed of analysis. More nuanced discussion of the benefits and drawbacks of this system are discussed in section 4. The State of Vermont has 309 unique ZIP codes, while New York City has 211 unique ZIP codes. A complete point-to-point matrix of travel times for a given region and mode of transport will require n^2 data requests. Travel within Vermont is restricted to driving, while New York City travel is mapped for driving, walking, bicycling, and public transportation (which encompasses buses, subway trains, ferries, and regional train services). Including requests for travel between a given ZIP code and itself (yielding a trip of zero seconds), the total dataset will hold 273,565 data points.

The Google Maps Distance Matrix application programming interface (hereafter API or Matrix API) holds adjacency time and distance information for journeys made between pairs of locations. Google restricts queries to the Matrix API to 2,500 per unique IP address per day. Each combination of origin, destination, and mode of transport counts as one query. Queries can also be made with an individual, Google-issued API key appended to them. Users may generate and hold multiple free API keys, which allow the user 2,500 queries per day, regardless of IP address. Users can also generate API keys and link their billing information; this allows for up to 100,000 paid queries per API key day. Queries involving travel by public transportation require API keys to be appended to the request. The Matrix API answers requests with XML data holding the specified journey parameters, journey time, and journey distance.

Limitations on time and budget meant that data were acquired using a combination of MATLAB scripting and a paid virtual private network (VPN) service. MATLAB's 'webread' function was leveraged to automate query construction and XML data retrieval. Over a number of days, the script was rerun to cover all possible journey combinations and transportation methods. This process was significantly accelerated because of the use of a VPN service (in this case, NordVPN). A VPN service encrypts a user's internet activity and changes the public IP address from which his or her internet activity appears to originate. In this context, the VPN was used to circumvent the daily quota imposed on an individual IP address by using the IP addresses of the VPN's various servers around the world. After the new origin server was chosen, the script was rerun such that the 2,500 requests allowed on that IP address were exhausted. This was repeated as needed until the data set was complete. For example, a single cycle of this process could complete the statewide journey times for eight Vermont original ZIP codes or eleven New York City ZIP codes. If the desired data set were acquired using just one IP address, it would take 110 days to complete. The use of API keys and the VPN service can reduce that to one tenth of the time or less.

3 Initial Presentation and Analysis

The two following plots show the arrivals data for journeys made between ZIP codes in Vermont (figure 1) and New York City (figure 2). After x seconds, a person travelling from a given origin ZIP code would have arrived at y other ZIP codes in the region of interest had all the trips been started simultaneously. Each coloured line, therefore, represents the sum of completed journeys from a single common starting ZIP code. From these 'spaghetti' diagrams, a rudimentary index of connectedness was developed which is discussed shortly. It is clear that some destinations accumulate completed journeys much more quickly than others. We first look to the Vermont case study. After $\sim$ 9000 seconds, (2.5 hours) some origin ZIP codes have seen almost all of their outward journeys completed, while others are only approaching half completion. There are of course some confounding factors inherent in this scenario, but they will be expanded upon in section 4. The rates of destination accumulation across the ZIP codes is rather consistent in the case of Vermont, which could be attributed to the relative homogeneity of Vermont's population distribution when compared to other states. An identical plot for a state like Nevada, for example, might look quite different, since it has only a few large and widely separated urban centres and very sparse populations elsewhere.

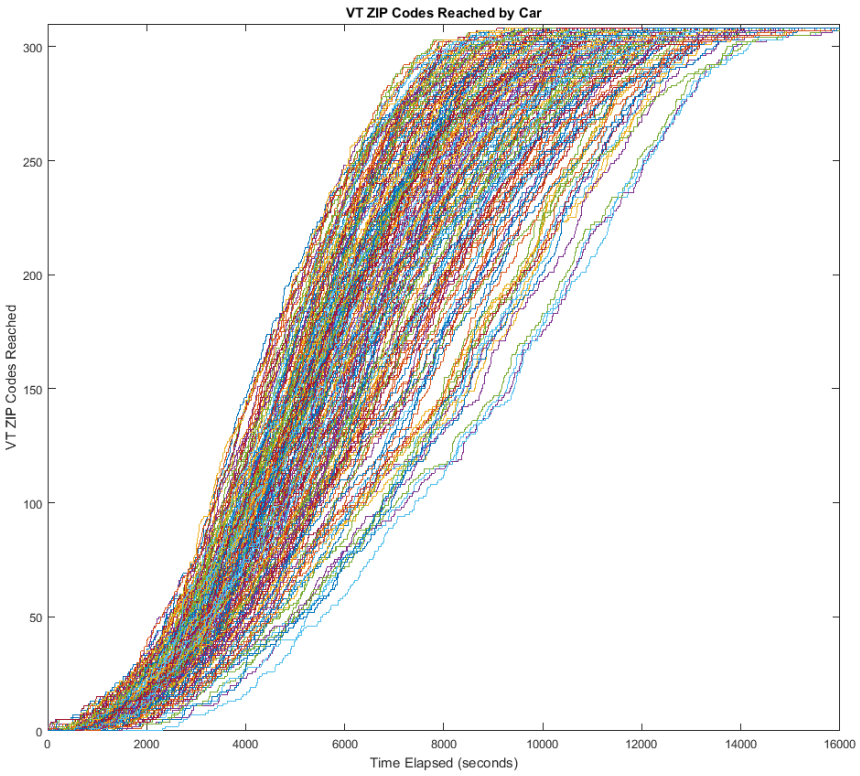


Figure 1: VT drive times

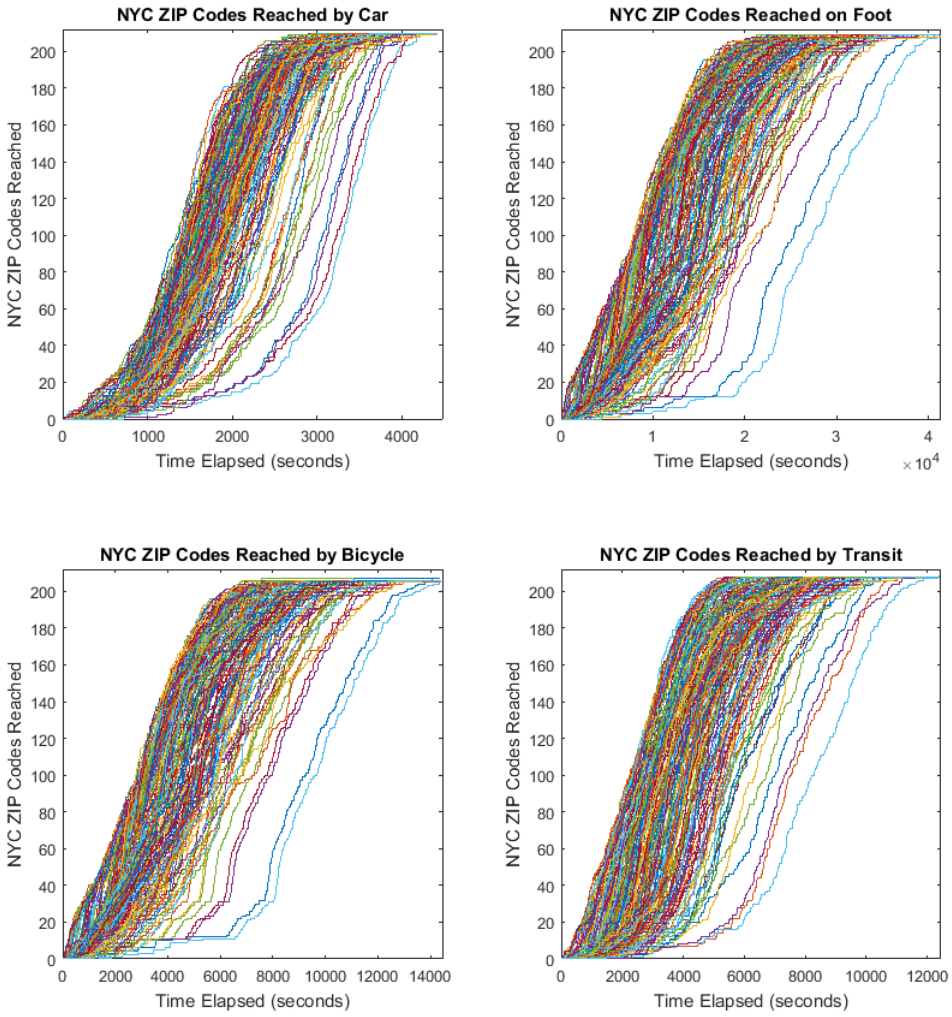


Figure 2: NYC travel times

In the New York City case, similar patterns appear in the completion of journeys by all modes of transport. There are more significant outliers on the 'slow' end of the chart, which hints that high population density is not a guarantor of high interconnectedness. New York City is more spatially fractured as compared to Vermont due to the waterways that bisect the five boroughs and separate the city from the mainland. It is likely that some of the low outliers seen in each more of transport represent a cluster of 'disconnected' locations such as Staten Island. Two ZIP codes covering Rockaway, Queens were omitted from the New York Zip code set for the non-driving modes of transport as Google does not seem to properly resolve these locations for point to point routing that does not involve cars. This area of New York was severely impacted by Hurricane Sandy in October 2012 and other Google products for the area such as Google Street View are not currently accessible.

The index of connectedness C_n is a simple metric constructed to express how well connected a given ZIP code is to its peers. From the above plots, a statistical z-score is calculated for each origin ZIP code per timestep given the distribution of arrival counts. (Peck 2005) Given a time t_{max} at which the final overall trip is completed, C_n is the mean z-score for a particular ZIP code n from $t = 1$ to t_{max}:

$$C_n = (\frac{1}{t}) \sum_{t=1}^{t_{max}} \frac{x_n - \bar{x}_t}{\sigma_t}$$

4 Discussion and Final Remarks

At time of writing, much analysis of this case in particular has yet to be carried out, but the initial forays into the data set hint at additional relationships worth pursuing further. Several general observations and caveats are already apparent. A significant limitation in this analysis is the arbitrary limiting of ZIP codes to specific geographical boundaries (in this case, state and county boundaries). Some locations near boundaries of the case study regions will undoubtedly be labelled as poorly connected relative to the state to which it belongs, but that neglects the good connections it might enjoy with it's adjacent state. The logical conclusion is to expand the analysis region to include all direct neighbours in the trip tabulations. As soon as the neighbouring region is included, then a new neighbour is excluded, and so on until a whole nation or continent is included. This would indeed create a more thorough accounting of the connectedness of a nation, but given the difficulties encountered in data acquisition to this point, further expansion of test region borders is not likely.

Another limitation point is the use of the ZIP code as a unit of location. The amount of people and land area which fall under a given ZIP code is not fixed, and thus treating all ZIP codes as being equal might obfuscate some of the actionable conclusions of the analysis. If, for example, one were to plan infrastructure investments to better balance the connectedness of places, it would be imprudent to invest heavily in a road project between two lightly populated and trafficked locations when a similar investment elsewhere would yield a better return on investment.

Despite these complications, there is much yet to be uncovered within these data sets. The immediate goals for this work as it progresses are to map the C_n data to their respective ZIP codes and identify patterns in connectedness. If additional data are gathered, a more diverse set of cases can be developed in a similar manner to the two already underway. This project is in its early stages and much work is yet to be done.

Reference

Peck, R., & Olsen, C. (2005). Introduction to statistics and data analysis (2nd ed.). Belmont, CA: Thomson Brooks/Cole.