Empirical Study on Route-Length Efficiency of Road Networks - Spring 2012 VIGRE Project Report

Karthik Ganesan

May 11, 2012

I Introduction

What does it mean to say that a physical network (such as a road network, electricity grid, or telephone network) is optimal? Different optimality criteria for different networks have been developed and studied. In the case of road networks connecting cities, an intuitive notion of an "optimal" network is one that provides the shortest routes between any two cities in the network while requiring a minimal total road network length. Some recent theoretical work of Aldous has focused on the properties of optimal random spatial networks (see [6] for an introduction). The mathematical models studied in this work place cities at random positions and the authors examine the tradeoffs between the route-lengths and the total network length as the number of cities in the network tends to ∞.

While it is not possible to study the optimality of real-world road networks on an infinite number of cities, [6] introduces two statistics (see Section II for an explanation) that can summarize the inefficiency and total length (representative of a "cost") of a road network. Using these two statistics to judge optimality of networks, do real road networks come close to being optimal? Previous undergraduate projects (see [4,5]) have attempted to answer this question by collecting road network data via the web and comparing the inefficiency and network length of these real networks to theoretically optimal random networks. These projects looked at networks on the 20 most populated cities within different states in the USA. One issue prevented the projects from analyzing larger networks: the data collection and calculation of the total network length was not fully automated.

In this project, we attempt to resolve this issue by writing R programs that perform these two tasks. One program accesses the web and collects detailed information about the routes between any two pairs of cities in the network, while a second program uses the collected data to compute the total network length. Using these two programs, we then examine road networks on the 40 most populated cities within different states in the USA. We also look at the road network on the 200 most populated cities throughout the entire USA.

II Notation & Definitions

The notation and definitions used in this work are also detailed in [6]. Let N be some number of cities. Consider these N cities and the shortest routes (via only road or rail) that connect each pair of cities. This forms a road network. Let x_i and x_j be two cities in the road network. Then we refer to $d(x_i, x_j)$ as the straight-line distance between the two cities and $l(x_i, x_j)$ as the distance of the shortest-route between the two cities. There are a total of $\binom{N}{2}$ such routes in the overall road network.

The inefficiency of the route between any two cities is represented by:

$$r(x_i, x_j) = \frac{l(x_i, x_j)}{d(x_i, x_j)} - 1$$

Notice that this route inefficiency is always a nonnegative number, since the route-length between any two cities can never be less than the straight-line distance. We also normalize the distances by dividing the

$l(x_i, x_j)$ and $d(x_i, x_j)$ terms by $\sqrt{\frac{A}{N}}$ where A is the total area containing the entire road network. This normalization makes the density of cities 1 per unit area.

Then two statistics summarize the network length and overall inefficiency:

$$\theta = N^{-1} \times \text{ (total network length)}$$
$$R = \max_{0 \le d < \infty} mean\{r(x_i, x_j) \mid d(x_i, x_j) = d\}$$

While averaging over city-pairs that are exactly the same distance apart in the definition of R is reasonable for random networks, the road networks examined in this work include only a finite number of cities and it is unlikely that any two different pairs of cities will be exactly the same distance apart. Therefore, the averaging intervals for computing the inefficiency are instead quantized and the following alternate summary statistic is used to estimate the inefficiency:

$$\rho(d) = mean\{r(x_i, x_j) \mid d - 0.5 \le d(x_i, x_j) \le d + 0.5\}$$
$$\tilde{R} = \max_{d \in \mathbb{Z}^+} \rho(d)$$

III Data Collection Methodology

Since we don't have access to a GIS database system, we have to make use of several open webservices from which necessary road network data can be extracted. The main selection criteria for the webservices used in this project was that they had to accept and allow a sufficiently large number of automated queries per day[1]. The three webservices used in this project are Hometown Locator [2], which provides population data for cities, the Yahoo! Geocoder [3], which converts city names to latitude and longitude coordinates, and the Mapquest Directions API [1], which provides detailed routes between cities. Since the R programming language offers many built-in packages and functions for webscraping, it is chosen as the language for implementing the automated data collection script.

III.A Collecting Route-Length Inefficiency Data

The webscraping script first takes a number of cities, N, and a state name as inputs. Another input parameter to the program indicates whether data on the road network on the N most-populated cities in the state or data on a road network on N random cities in the state should be collected. Then, the program accesses city population data on the web and parses a table of populations of all cities in the desired state. The city list is then arranged in descending order and the names of the top N cities (or N randomly chosen cities, depending on the input parameter) are saved.

Then, the geocoder is accessed and the latitude and longitude coordinates of each city are obtained. This is important because it allows for direct calculation of the straight-line distances ($d(x_i, x_j)$) between any two cities in the network, and more importantly, the Mapquest API requires the starting and ending points of any route to be given in latitude and longitude format. Finally, the Mapquest API is accessed and the route-lengths ($l(x_i, x_j)$) between all pairs of cities in the networks is obtained. In addition, each route between any two cities in the network is broken down and relevant information is extracted. The names of all roads traversed in each route, the starting and ending point of each road segment in latitude and longitude format, the distance of each segment, and the direction travelled on each segment are all saved and arranged in one master list for the entire road network. This detailed route data is used later to calculate the total length of the road network (see Section III.B).

A flow diagram detailing these steps is shown in Figure 1.

III.B Calculating the Total Network Length

A separate program then accesses the collected list of road segments for the whole road network and attempts to remove overlaps. All of the road segments are arranged by road name. Then for each road, all of the

[1] Since our goal was to automate this data collection.

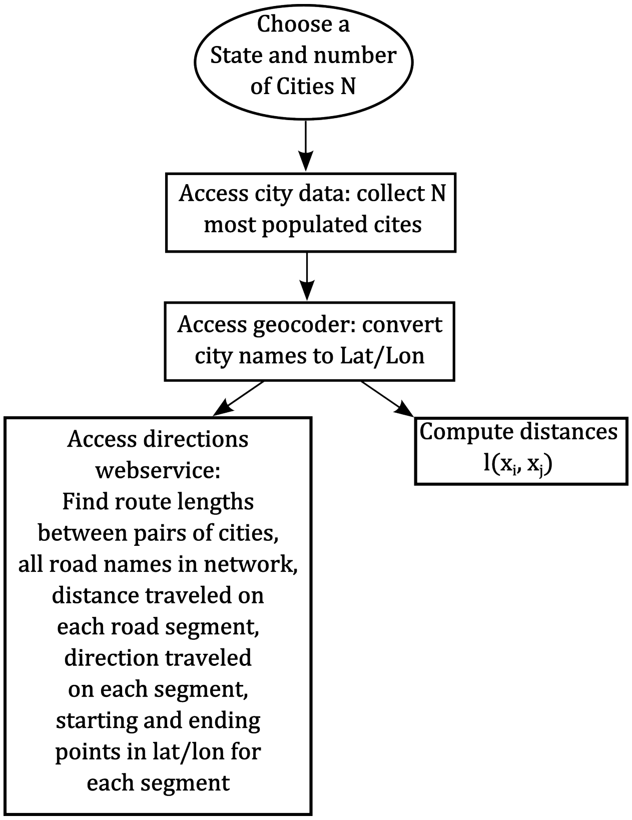
Choose a State and number of Cities N
Access city data: collect N most populated cites
Access geocoder: convert city names to Lat/Lon
Access directions webservice: Find route lengths between pairs of cities, all road names in network, distance traveled on each road segment, direction traveled on each segment, starting and ending points in lat/lon for each segment
Compute distances $l(x_i, x_j)$

Figure 1: Flow diagram showing the steps carried out by the automated data collection script.

segments traversed on that road are arranged in descending order of segment length. Then, the segments on each road are compared and overlaps are eliminated by comparing the direction, distance, and latitude and longitude points of the starting and ending point of each segment. If a segment is found to partially overlap with another segment, the distances of the non-overlapping portions are calculated via a Mapquest query and saved.

After all the overlaps are removed, the network length is calculated by simply summing up all of the distances of non-overlapping road segments in the network. A flow diagram detailing these steps is shown in Figure 2.

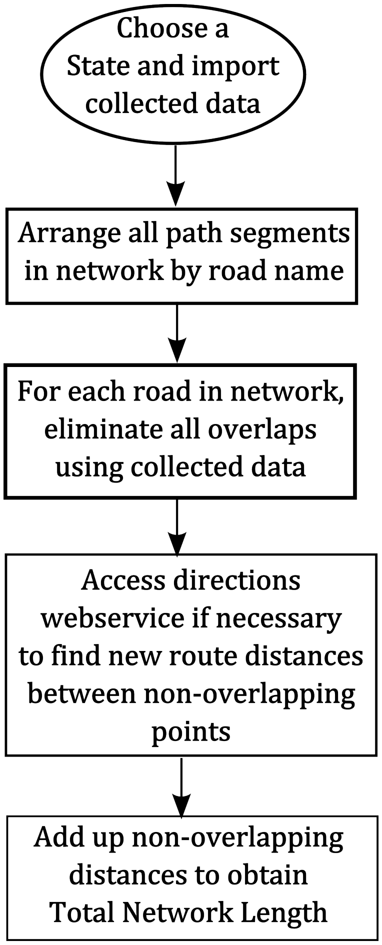
Choose a State and import collected data
Arrange all path segments in network by road name
For each road in network, eliminate all overlaps using collected data
Access directions webservice if necessary to find new route distances between non-overlapping points
Add up non-overlapping distances to obtain Total Network Length

Figure 2: Flow diagram showing the steps carried out by the network length calculation script.

IV Results & Project Status

We first collected inefficiency data for road networks on the 40 most-populated cities in 12 different states (CA, AZ, IL, PA, MI, MN, NY, MD, WA, OR, KY, and NC) and plotted scatter plots the route inefficiencies vs. the normalized straight-line distance between cities in the network. We superimposed two versions of the estimated $\rho(d)$ function for normalized integer distances d on these plots. One is the version described in Section II and the second uses a weighted average in the definition, where the weights are the products of the populations of the cities connected by the route. We also collected inefficiency data for the road network on the 200 most populated cities in the entire United States. All of these plots are given in Appendix A.

With the exception of Oregon, the estimated $\rho(d)$ function remains roughly constant with respect to the normalized straight-line distances between cities, which agrees with the theory presented in [6]. The $\rho(d)$

function for the random networks examined in [6] tends to increase to a maximum value at a normalized distance of around 2 to 3 and then remains roughly constant at larger distances (see [6, Figure 6]). From these scatter plots, we notice that the value of the estimated weighted $\rho(d)$ is often lower than that of the estimated regular $\rho(d)$. This may be expected as roads typically evolve over time starting from straight paths linking together large cities.

The total network length was also calculated for 6 of the 12 states and a plot of $\tilde{R}$ vs. θ was obtained. However, it was difficult to confirm whether the calculation was correct or not as the plot did not agree with the theory developed in [6]. Such a calculation is particularly hard to verify in an automated webscraping script, as there are hundreds of thousands of overlaps which need to be removed properly for each road network. A better approach, therefore, would be to use a GIS database in order to do this computation, as the databases possess many built-in functions that compute overlaps between paths. Hopefully, those with access to GIS databases can be convinced to investigate some of these issues.

Besides verifying current calculations for the total network length, there are further questions that can be investigated using the programs we have written. Instead of choosing the 40 most populated cities, what if we chose 40 cities at random? As mentioned before, we have this capability written in our code, but did not use it in this project. Another interesting project might be to compare road networks within the USA to road networks in other countries.

An archive file containing all of the R code, plots, collected data, and calculated numbers in this project is available online at [7].

V Acknowledgements

This project was supported by the N.S.F. VIGRE grant to the UC Berkeley Statistics department.

References

[1] Mapquest directions api web service. `http://developer.mapquest.com/web/products/dev-services/directions-ws`.

[2] Us cities and state gazetteer. `http://www.hometownlocator.com/`.

[3] Yahoo! maps web services - geocoding api. `http://developer.yahoo.com/maps/rest/V1/geocode.html`.

[4] D. Aldous, Y. Cheng, J. Friedman, Y. Huoh, W. Lee, and H. Liu. Route-length efficiency in spatial transportation networks. Oct. 2007. `http://www.stat.berkeley.edu/~aldous/Unpub/sharp_curve.pdf`.

[5] D. Aldous and A. Choi. A route-length efficiency statistic for road networks. June 2009. `http://www.stat.berkeley.edu/~aldous/Spatial/paper.pdf`.

[6] D. Aldous and J. Shun. Connected spatial networks over random points and a route-length statistic. *Statistical Science*, 25(3):275 – 288, 2010.

[7] Karthik Ganesan. Road network data, May 2012. `http://inst.eecs.berkeley.edu/~karthikg/FinalRoadNetworkDataSpring2012.zip`.

A Appendix: Plots

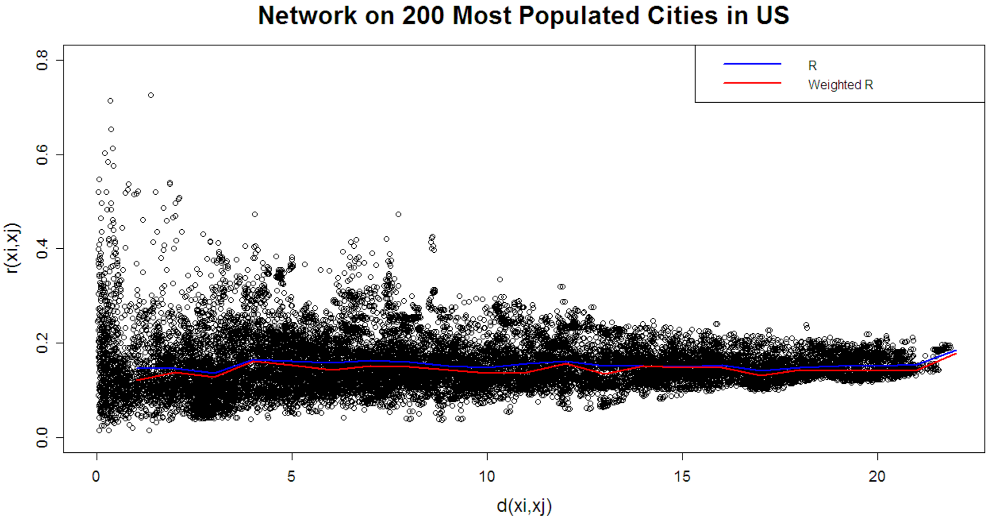

Figure 3: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 200 most populated cities in the United States. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

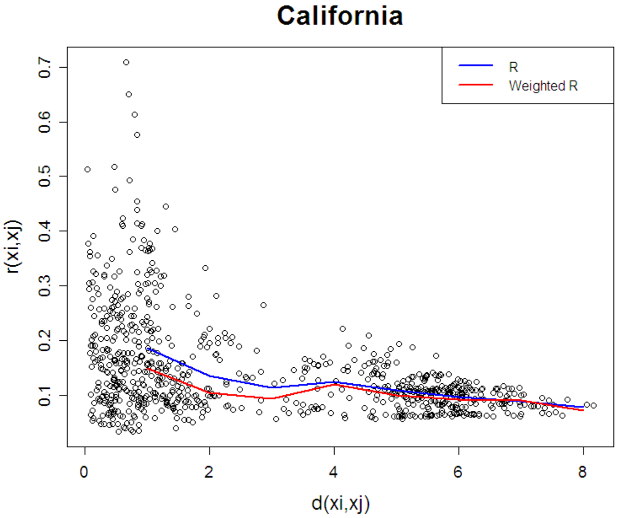

Figure 4: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of California. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

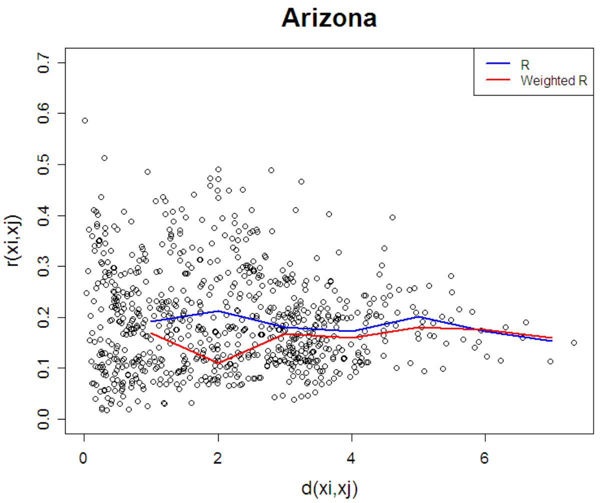

Figure 5: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Arizona. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

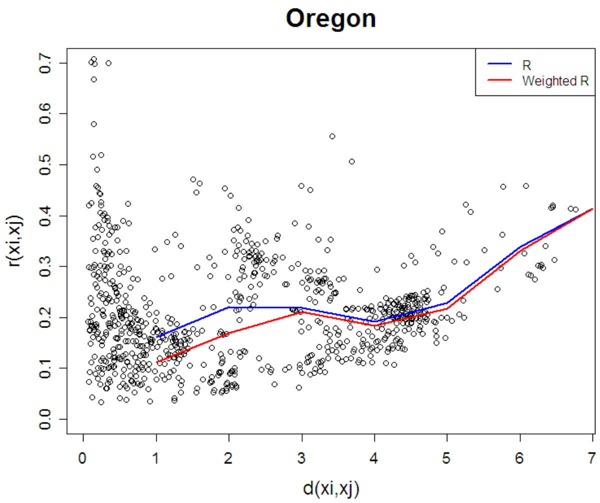

Figure 6: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Oregon. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

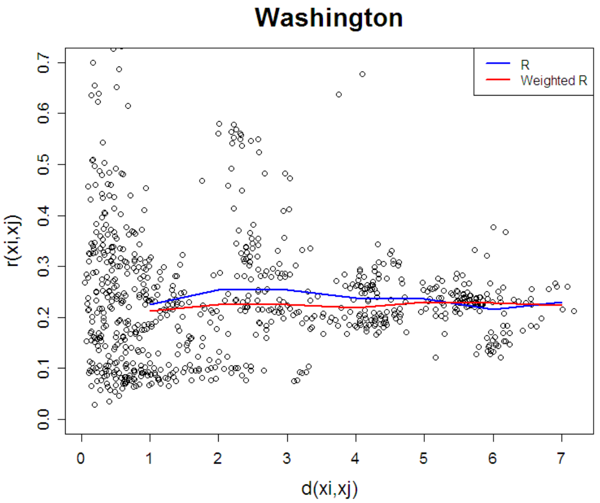

Figure 7: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Washington. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

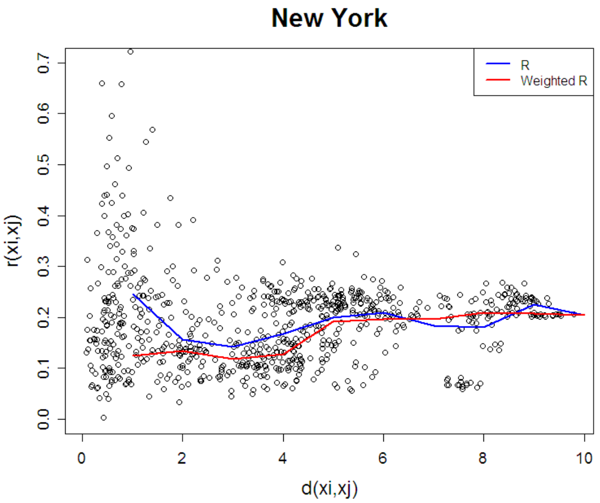

Figure 8: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of New York. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

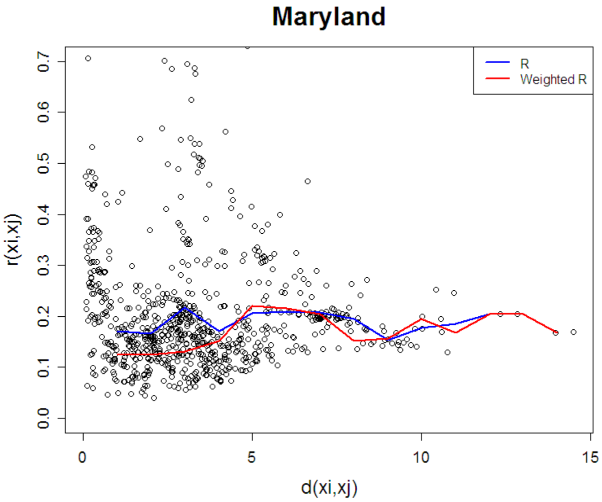

Figure 9: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Maryland. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

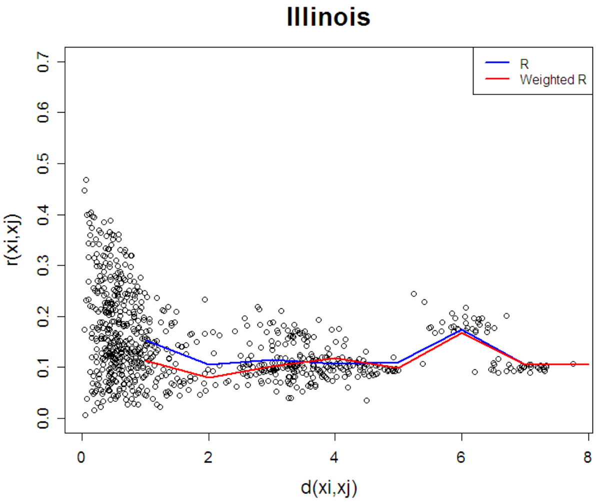

Figure 10: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Illinois. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

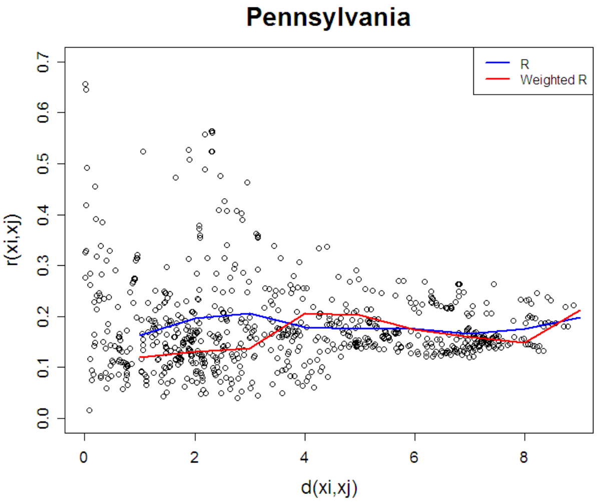

Figure 11: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Pennsylvania. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

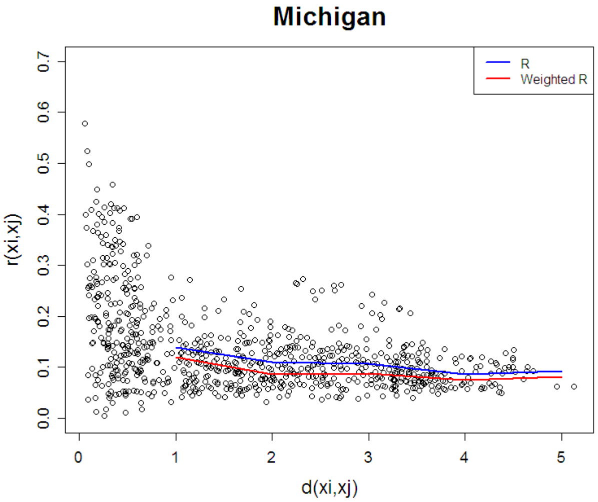

Figure 12: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Michigan. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

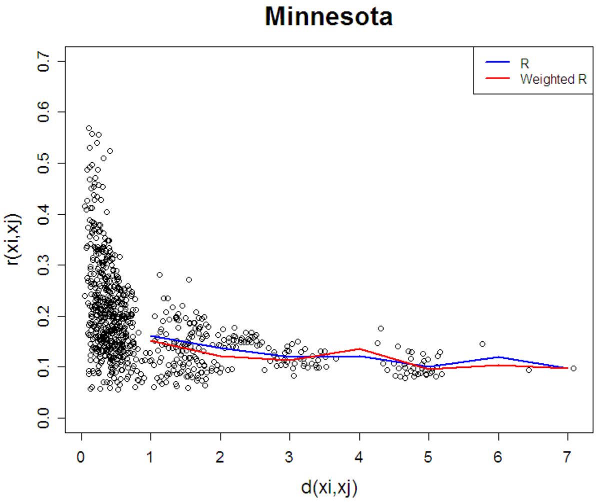

Figure 13: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Minnesota. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

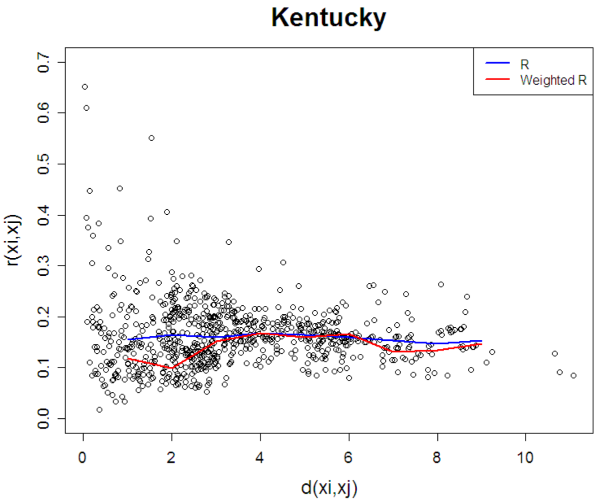

Figure 14: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of Kentucky. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.

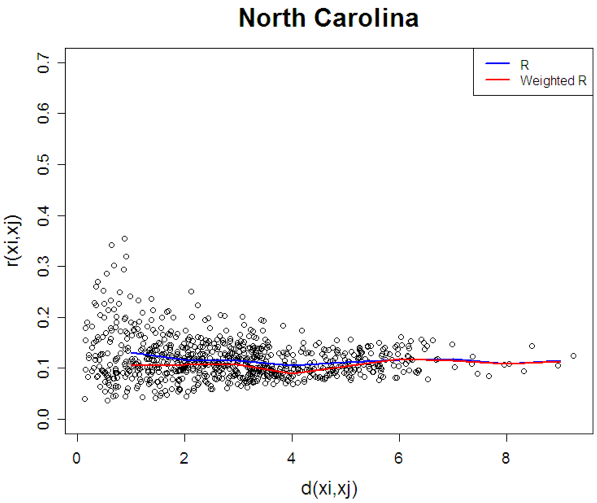

Figure 15: Scatter plot of the route-inefficiencies vs. normalized straight-line distance between 40 most populated cities in the state of North Carolina. The blue curve is the estimated $\rho(d)$ function while the red curve is the population-weighted estimate of the $\rho(d)$ function.