Understanding the Migration of Doctors in Washington, DC:

Using Historical Data to Visualize the Change in Location of Doctors Over Time

Sean Choi, Maryam Davoodi, Susanna Mostaghim

C. C. Thomas and S.A. Mitchell, *City of Washington*, Thomas, Cowperthwait & Co, Philadelphia, 1854

This project originally started as a feasibility study with the goals of creating a program that could model and visualize the movement of doctors in Washington, DC. However, due to the data having poor or no delimiters, the data cleaning took a lot longer than anticipated and the client changed the goal trajectories mid-semester. As a result, this project maps the historic locations of doctors in Washington, DC using heatmaps built in `R` to map the addresses to the modern roads of Washington, DC. After which a historical map sourced from the Library of Congress[1] was superimposed on the heatmaps using `MATLAB`.

Due to the length of the timeline and the data limitations from historical sources, additional work that would be desired could not be completed within the semester. However, future work can easily be done using the census data for demographics associated with the years that have been visualized. The `R`-code itself can help clean census and medical license data for additional years in order to understand the locations of doctors in relation so social and economic context.

[1] C. C. Thomas and S.A. Mitchell

1. Problem Statement The initial objective of the project was to see if it would be feasible to create a model prototype of Washington, DC that gives users an understanding of the change in healthcare accessibility during different time periods with varying economic influences and crime levels. Some of the important data we initially planned on collecting were yearly population demographics such as race, gender, and age as well as the relevant factors: yearly income rates, education level, environment transformations, and the economic status of the doctors in addition to the overall population.

However, after meeting with the client, there was a trajectory change in the project goals due to the limitation of the given timeline and lack of available information. Due to this, Dr. Ewing requested that we focus on visualizations of different years over time instead of creating a mathematical model to explain the migrations. The visualization is done through a heat map which can be superimposed on top of a historical map of Washington, DC. This was done using the datasets of information about physicians such as name, specialty, and address. This shows the availability of the physicians and overall change in location over the years.

This is important to our client as a historian that the historical data is implemented in order to truly see where the doctors have been located during periods of economic and social significance. As such, this method can be applied to other historical data concerning Washington, DC or other cities if they followed the same procedure when it came to data cleaning and coding to create heatmaps.

The resulting heatmaps from contemporary data could be used with a model in various fields as part of an explanation for underlying reasons to create a public policy or for individuals starting a business. Visually inspecting the location can escalate the questions on why one location is more populated than other. Not only this can be used in Washington, DC, it is feasible that it could also be used for another city by creating a program that an inexperienced user could implement with specific instructions.

1.1 Ethical Considerations of the Problem As this data is historical it has no ethical considerations, however the analysis and visualization done on the historical data could used if applied to contemporary city data. The information from the analysis and visualization could then be used either in a positive or negative manner depending on who possesses and uses the data.

Potential decisions based on contemporary demographic and doctor location data would allow for people to identify areas that are lacking in services. While our data is pertinent to Washington, DC and the medical profession, the ethics of the data can be generalized to any profession that has publicly available location data as well as other cities. Analyzing the data to find out where certain demographics and professions are centered could affect decisions of adding service and health locations to areas. This could allow for the identification of areas that lack human, medical, and public services which can lead to the potential for these services to be provided to areas in order to increase access. In turn, this could lead to the creation of better infrastructure by identifying specific areas that are in need. In addition, the data could be used to analyze the amount of the crime in comparison to the demographics and expand public services such as police and fire departments in that area in order to protect the population.

There is definitely the potential for misuse of contemporary data. Consider if a private party or corrupt government were to get their hands on the data, it could potentially be disastrous for low-income areas of Washington, DC (and other cities this could be applied to). The misuse of demographic data in relation to services could be used to mass-gentrify an area. This may not sound too bad at first as gentrification is viewed as cleaning up a city. However, gentrification displaces low-income households in the interest of development and could potentially create a homelessness problem. A private company offering services that an area needs could potentially monopolize that neighborhood's need for a service and raise their prices for those services and essentially take economic control of a portion of a city's population.

In addition to forced gentrification, a politician could use this data to gerrymander an area in order to gain votes to keep him in office and pander to what he knows voters in specific areas need based on the services available in their neighborhoods. Along the same lines, a lobbyist could use the data as leverage over a politician by providing services and favors to his community that help keep the politician aligned with their interests and influence his policy as long as he holds office.

1.2 Literature Review A large problem with this project originally was that there was no published similar work to approach this problem. However, there were many different theoretical models that could be used in as references to build a data structure. Models that explain human migration are very hard to come by, however the oldest standing model is the push-pull model[1] that gives a very generalized result of movement. Other models that could potentially have been used were more complex and required more historical data than we would be able to find, such as the human capital model [2] which was developed by Michael Greenwood and discussed in different papers [3]. Greenwood's model does not offer much outside theory and require more information than purely mathematical models based on non-homogeneous Markov chains[5] or nodal discontinuous Galerkin methods[4].

However, there are faculty in other universities doing research in similar areas that we could have potentially contacted for some guidance on this project when it came to the development of a model. By reviewing some sites we were able to find a historical map of Washington, DC via the Library of Congress[6], pictured below in Figure 1.

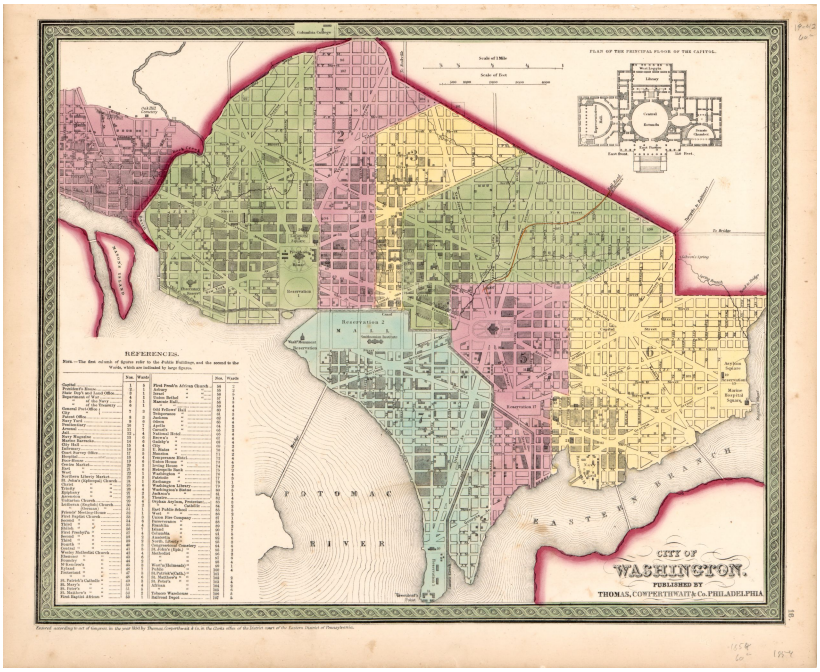

Figure 1: Washington, DC in 1954

2. Design and Data Our project was mostly focused on finding historical data about doctor locations in Washington, DC the proceeding to clean and visualize it via heatmaps in R. Originally we were asked to create a migration model and designed our project around it before we had a trajectory change from our client Dr. Ewing, which changed our design criteria and solutions.

2.1 Design Criteria When we initially began this project, we identified four design criterion: model accuracy, run time, application to other cities' datasets, and ease of use. As we were ambitious, we defined model accuracy as the modeled migration of doctors to and from Washington, DC as being within 10% of the brute-forced calculations from the data. The criteria of run time was defined as the program finishes running the model within a 30 minute time period. Third, the criteria of program versatility to be applied to other cities' datasets was set such that the program could be used for other cities so long as the data adheres to specific guidelines set by the program. The final criteria of the project was that it could be easily used by someone with little experience in using a data analysis program.

However, due to our trajectory change, we are no longer using a model for this project and most of the criterion we initially determined as necessary are no longer needed. Our main criteria was using data to create an accurate visualization.

The visualization aspect itself does not depend heavily on feasibility as it is possible to visualize the data on a dynamic

map of Washington DC. A dynamic map is done through a heatmap that is superimposed on top of the Washington DC map. This generates graphical representation of physicians data where the individual values contained in the map are represented by the density of a color. Our program is designed to demonstrate time-lapse using the heatmap which shows the changes in the area by each year of the physicians dataset.

The versatility of the visualization of the program can easily be applied to other cities datasets as long as the data adheres to specific guidelines set by the program. Having said that, this could be easily used by someone with little experience in using a data analysis program.

2.2 Data Validity Due to the nature of the older datasets, it can be assumed that there is a relative level of inaccessibility in comparison to newer data. This is a result of unreported practices as well as potentially illegal practices that are highly relevant to the socioeconomic focus of this project. The unreported information also reflects poorly on the concentration of physicians, giving a skewed sense of accessibility to healthcare. There is also quite a bit of missing data as some of the physicians did not have addresses tied to their practices and had to be removed from the data sets. Just as the lack of accessibility influences the accuracy of the data, the city boundaries and addresses influence the accuracy of our visualizations.

As the addresses had to be modified to add the quadrants of Washington, DC associated with the cardinal directions, there is some potential for a skewed visualization in the 1864 heat map shown in Figure 2. Although there are a number of factors negatively influencing the old visualization's validity, the data was collected from Boyd's Directory which is a respected data source for Washington, DC from 1860-1909.

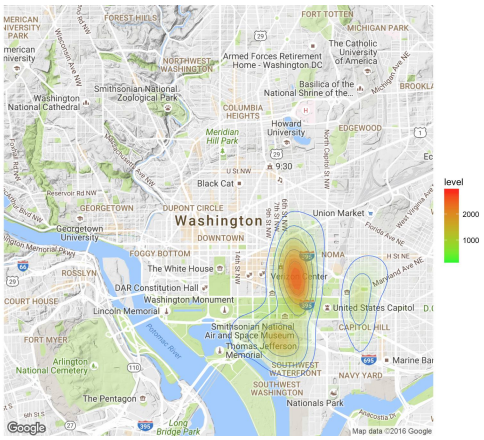

Figure 2: Washington DC, 1864

The newer data is being collected by the Open Government Advisory group, which is held to a very high standard of data validity. This makes for highly accurate visualizations when the data comes from a scrape of the open data site that the Open Government Advisory group runs. However, one issue using this data for the visualization is from the potential data loss during scraping.

2.3 Design Solutions Originally our design solutions to model the movement of doctors in Washington, DC was between three different model bases: push-pull, human capital, and non-homogeneous Markov chains. When the different models were analyzed it was determined that the push-pull model was too simple for the data and even if accurate would be too generalized to have validity. This resulted in narrowing the model selection to two choices: the human capital model and non-homogeneous Markov chains.

Markov chains have been used by sociologists and economists to model human migration before. In these models, the states of the Markov chains are various geographical locations. In ours, these states would be locations in Washington, DC over time. The transition probabilities are either empirically estimated or assumed to possess certain properties. However, there is a problem using homogeneous Markov Chains to model the migration of the studied population over time as they do not take into account the variation of transition matrices over time. This would require the use of a Markov chain that uses an irreducible stochastic matrix and allows for multiple time-steps as well as chain

migration.[5] However, Markov chains are incredibly computationally expensive and require a large amount of time in order to implement. This model would have been implemented in `MATLAB` or `C` while drawing on the data.

The second choice of human capital model was much easier to implement. The model is based on a collection of resources of individuals knowledge, talents, skills, habits, social, and personality attributes as a form of capital. It is mostly useful for a broad spectrum of data and could be applied to the individuals in the medical field. This model would have been implemented in `C` or Python. This model in particular was chosen over the non-homogeneous Markov chains due to the computational complexity and run time associated with each model.

During early model development, our objective of the project changed mid-semester by Dr. Ewing from creating a model using the data to visualizing with reference to the demographic. This was due to the data being in forms that weren't easy to clean due to the poor delimiters if delimiters within the data were present at all. Much of the data cleaning had to be done manually by opening the `.xlsx` files in Microsoft Excel and going through them in order to find doctors and their associated addresses. However, once the data was cleaned it was easy to begin mapping where doctors were using `R` and packages such as `ggmap` and `ggplot2`. The code itself used to clean and map the data is Code 1 in Appendix 1: `R` Codes. Unfortunately, due to `R`'s limitations.

3. Progress Overview With our project, there were some significant setbacks and obstacles we did not anticipate. However, we did accomplish the cleaning of datasets that allowed us to see snapshots of historical locations of doctors in Washington, DC.

3.1 Obstacles Due to the trajectory change that occurred for our project, there is less concern about mathematical accuracy and computational efficiency of a model. Due to this, our main obstacle for the project became data collection and cleaning. There is a limited amount of data publicly available between 1850 and 1970 due to lack of collection or copyright issues. The only datasets that could feasibly be collected from the Boyds directory between 1850-1900, however these datasets were not easy to clean as each directory is in a scanned PDF document from a hard copy which had optical character recognition automatically applied to allow for searching. This, in turn, allows for the document to be scraped through using our algorithm in `R`. However, there are a large number of characters that are garbled and creates a need for manual scraping.

The National Archives and Record Administration holds the Washington, DC directory from 1867 to 1970 with some gaps. After 1920, a large amount of data is not available to the public digitally. However, the majority of the missing datasets are available in hard copy at the Robert M. Warner Research Center in Washington, DC. The missing gaps can be collected from going through the hard copies and manually entering the data into a data table or an excel file. Doing such a task would be incredibly time-consuming and also violate the objective this project being a feasibility study.

3.2 Results Focusing on the visualizations from years 1861-1894 there is a dispersion of medical practices from a central point. Most of the medical practices at the beginning of the study are located between Capitol Hill and the White House, but as we see in Figure 3, a very large concentration of physicians began to practice close to the downtown area and more physicians dispersed their practice to areas that had less access to healthcare in 1864.

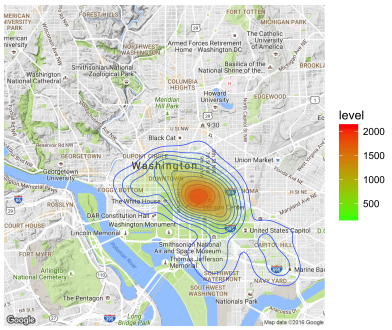

Figure 5: Washington, DC 1871

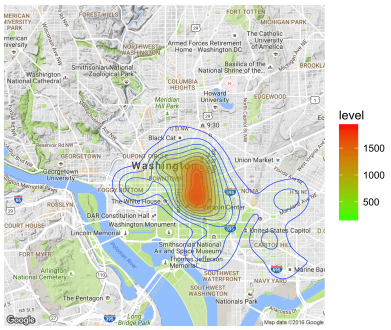

Figure 6: Washington, DC 1878

Comparing 1871 to 1878 and 1881, there are also practices appearing on Maryland Avenue and in the Georgetown area. Comparing 1881, as shown in Figure 5, to 1893 in Figure 6, we can see further dispersion of practices and a

stronger concentration in the downtown area of Washington, DC.

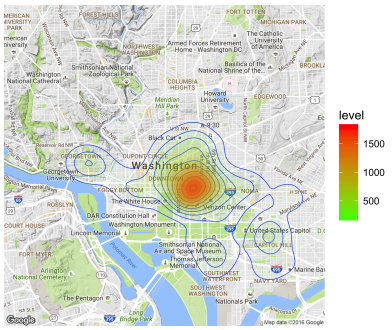

Figure 3: Washington, DC 1881

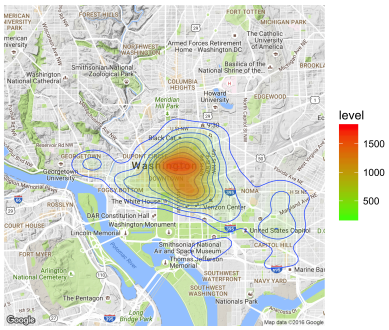

Figure 4: Washington, DC 1893

Due to our struggles with data collection, even though we saw the trends we were expecting in the older visualizations, there is not much that can be answered without census data or modern data. There are a variety of socioeconomic questions such as healthcare accessibility in inner cities that can't be answered without this data. With the addition of historical census data, the visualizations would be able to show accessibility of healthcare to people of color in times of segregation as we would be able to see which areas people of color were concentrated in as well as which physicians were willing to treat them.

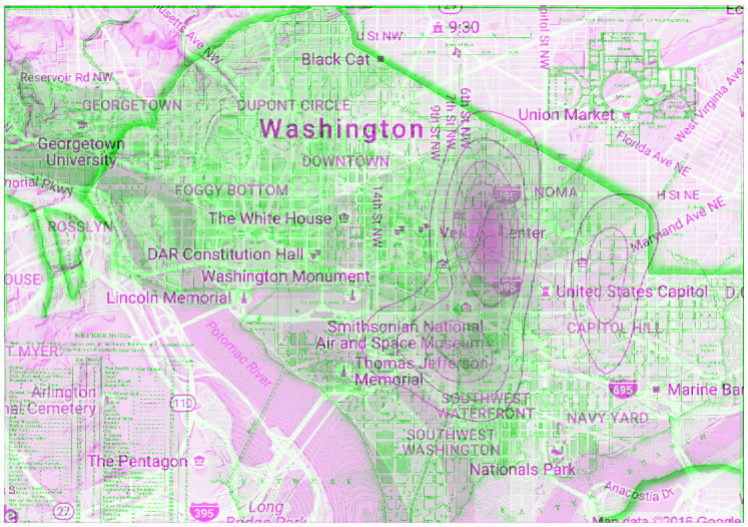

Figure 7: Composite map of Washington, DC 1864

However, we were able to superimpose a historic map on several of the heatmaps using variations of the `MATLAB` code in Appendix 2. A early example of which is pictured above in Figure 7.

3.3 Future Work When looking into factors influencing healthcare accessibility in the 1800s, segregation cannot be ignored as it would be detrimental in studying social change over time. Further studies into hospital segregation are important to the understanding of healthcare accessibility to people of color in the late 1800s, as well as understanding the location of physicians, including local population's income and race.

If possible, it would also be helpful to have research with a more historic undertone into private practices to further study segregation in 1800s healthcare. As the project's focus is socioeconomic impact on the movement of doctors in Washington, DC, collecting census data on racial, gender, and income demographics would aid in determining healthcare accessibility in more urban areas over time. In addition, there would be benefits into researching which hospitals physicians were physically close to as it could be assumed that addresses close to specific hospitals indicate

which races they tend to. This would not only aid in identifying accessibility to healthcare , it would give insight into understanding the specializations of physicians in the 1800s. This is due to the fact that most of the data does not indicate the specialization of a doctor, only that they are a physician.

The other focus would be to collect more data on modern day physician addresses and personal attributes in the Washington, DC area. This potentially could allow us to create a working Human Capital Model or another mathematically sound model that would take into account the socioeconomic events that influence the migration of doctors in Washington, DC.

4. Conclusion As we have done mostly visualizations, this project shows that as a feasibility study, the modeling of the migration of doctors in Washington, DC is currently not feasible. If data could be cleaned faster and become more accessible, it is possible that a model could be constructed to explain the movements in relation to socioeconomic events over time.

Appendix 1: R Codes

Code 1

```
library(psych)
library(xlsx)
library(ggplot2)
library(ggmap)

#1864 GGMAP
data1864 = read.xlsx("washington1864.xlsx", 1, header = FALSE)
data1864 = data1864[,1:3]
colnames(data1864)<-c("Name", "Address", "Profession")

#downloaded map
dc_map <- get_map(location = "washington dc", zoom = 13)

names1864<-data1864[,1]
profession1864<-data1864[,3]
addresses<-as.character(data1864[,2])
addresses
longLat1864<-geocode(addresses)
longLat1864
#DATA PREPPED FOR GGMAP
newData1864<-cbind.data.frame(names1864 ,longLat1864$lon ,longLat1864$lat,
    profession1864)

png(filename="1864headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1864, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1864,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)

dev.off()



#1871 GGMAP
data1871 = read.xlsx("washington1871clean.xlsx", 1, header = FALSE)
colnames(data1871)<-c("Name", "Address", "Profession")

names1871<-data1871[,1]
profession1871<-data1871[,3]
addresses<-as.character(data1871[,2])
longLat1871<-geocode(addresses)
```

```
#DATA PREPPED FOR GGMAP
newData1871<-cbind.data.frame(names1871 ,longLat1871$lon ,longLat1871$lat,
    profession1871)

png(filename="1871headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1871, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1871,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)
dev.off()

#1878 Heat Map
data1878 = read.xlsx("washington1878clean.xlsx", 1, header = FALSE)
colnames(data1878)<-c("Name", "Address", "Profession")

names1878<-data1878[,1]
profession1878<-data1878[,3]
addresses<-as.character(data1878[,2])
longLat1878<-geocode(addresses)

#DATA PREPPED FOR GGMAP
newData1878<-cbind.data.frame(names1878 ,longLat1878$lon ,longLat1878$lat,
    profession1878)

png(filename="1878headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1878, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1878,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)
dev.off()

#1881 GGMAP
data1881 = read.xlsx("washington1881clean.xlsx", 1, header = FALSE)
colnames(data1881)<-c("Name", "Address", "Profession")

names1881<-data1881[,1]
```

```
profession1881<-data1881[,3]
addresses<-as.character(data1881[,2])
longLat1881<-geocode(addresses)
longLat1881

#DATA PREPPED FOR GGMAP
newData1881<-cbind.data.frame(names1881 ,longLat1881$lon ,longLat1881$lat,
    profession1881)

png(filename="1881headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1881, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1881,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)
dev.off()

#1893 GGMAP
data1893 = read.xlsx("washington1893clean.xlsx", 1, header = FALSE)
colnames(data1893)<-c("Name", "Address", "Profession")

names1893<-data1893[,1]
profession1893<-data1893[,3]
addresses<-as.character(data1893[,2])
longLat1893<-geocode(addresses)

#DATA PREPPED FOR GGMAP
newData1893<-cbind.data.frame(names1893 ,longLat1893$lon ,longLat1893$lat,
    profession1893)

png(filename="1893headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1893, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1893,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)
dev.off()
```

```
#1894 ggmap
data1894 = read.xlsx("washington1894clean.xlsx", 1, header = FALSE)
colnames(data1894)<-c("Name", "Address", "Profession")

names1894<-data1894[,1]
profession1894<-data1894[,3]
addresses<-as.character(data1894[,2])
longLat1894<-geocode(addresses)

#DATA PREPPED FOR GGMAP
newData1894<-cbind.data.frame(names1894 ,longLat1894$lon ,longLat1894$lat,
    profession1894)

png(filename="1894headmap.png",
    units="in",
    width=5,
    height=4,
    pointsize=12,
    res=1200)
ggmap(dc_map, extent = "device") + geom_density2d(data = longLat1894, aes(
   x = lon, y = lat), size = 0.3) +
  stat_density2d(data = longLat1894,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level
                     ..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low =
                     "green", high = "red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)
dev.off()
```

Appendix 2: MATLAB Codes

Code 1

```
%% Scaled heat map from R read in and display
dcheatmap = imread('washingtondc.png'); %read in your heatmap
figure(1) %create separate figures in order to see differences
original = imshow(dcheatmap); %display heatmap being used for fusing
                              % purposes
pause

%% Use scaled historical map
historicmap = imread('historic_map.jpg'); %read in your historical map
figure(2)
imshow(imread('historicoriginal.jpg'));
pause
figure(2)
map = imshow(historicmap); %display historical map being used
pause
%% Creating a composite image using two spatially referenced items
figure(3)
dcheatmapref = imref2d(size(dcheatmap)); % creates image reference data
                                         % type for 2d image
historicmapref = imref2d(size(historicmap));
historicmapref.XWorldLimits = dcheatmapref.XWorldLimits
historicmapref.YWorldLimits = dcheatmapref.YWorldLimits
[composite, compositeref] = imfuse(dcheatmap,dcheatmapref, ...
    historicmap, historicmapref);
compositeimage = imshow(composite);
```

Bibliography

[1] G. DORIGO AND W. TOBLER, *Push-pull migration laws*, Annals of the Association of American Geographers, 73 (1983), pp. 1–17.

[2] M. GREENWOOD, *Modeling migration*, Encyclopedia of Social Measurement, 2 (2005), pp. 725–734.

[3] ——, *Modeling Migration*, Elsevier, Philadelphia, 2008.

[4] J. S. HESTHAVEN AND T. WARBURTON, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, Springer, New York, 2008.

[5] J. PAN AND A. NAGURNEY, *Using markov chains to model human migration in a network equilibrium framework*, Mathl. Comput. Modeling, 19 (1994), pp. 31–39.

[6] C. . C. THOMAS AND S. A. MITCHELL, *City of Washington*, Thomas, Cowperthwait & Co, Philadelphia, 1854.