Using population weighted county centroids to calculate migration distance for historical linked census data: A case study of male migrants in the New York and New Jersey area, 1880-1910

Sula Sarkar, Minnesota Population Center, University of Minnesota
Rebecca Vick, Minnesota Population Center, University of Minnesota

Background

The availability of complete count historical microdata has permitted the linking of individuals between census years for longitudinal analysis. These linked datasets provide very important research opportunities regarding historical population mobility. In previous work, we estimated migration distance with geographical country centroids. This work underlies the MILEMIG variable in the IPUMS Linked Samples[1]. The geographical centroid assumes that the population is evenly distributed throughout the county; this assumption is problematic when much of the population resides in urban areas. This paper extends and refines that measure by calculating migration distance using *population weighted* county centroids, which we argue relies on more realistic assumptions about how county populations are distributed. We describe our method of calculating population weighted county centroids. We then compare the distances calculated using geographical versus population weighted county centroids. Finally, we describe the effects of population weighted centroids on the migration distances in one of the IPUMS Linked Samples.

Data

Our analysis utilizes the complete count historical microdata from the North Atlantic Population Project (NAPP)[2] and Integrated Public Use Microdata Series (IPUMS) [1] sample and linked data. We use the United States 1880 complete count data and the United States 1910, 1-in-100 national random sample for population estimates. We use the 1880-1910 IPUMS Linked Sample of males for linked records. Each linked record already includes a variable called MILEMIG that contains a migration distance estimate that was calculated using unweighted geographical county centroids.

For map files, we use National Historical Geographic Information Systems (NHGIS) historical county shapefiles for 1880 and 1910[3]. The county shapefiles reflect changes in county boundaries between the 2 census years. We also use the urban area polygons from NHGIS, and extract only the cities that are found in the CITY variable in the 1880 and 1910 NAPP census data files.

Methods

To compute a distance migrated we need a starting point and end point for each migrant. We use U.S. state and county of residence, as listed in the IPUMS STATE and COUNTY variables, to locate each

---

[1] **Steven Ruggles, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek.** ***Integrated Public Use Microdata Series: Version 5.0*** **[Machine-readable database]. Minneapolis: University of Minnesota, 2010.**

[2.] Minnesota Population Center. North Atlantic Population Project: Complete Count Microdata. Version 2.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center, 2008.

[3] Minnesota Population Center. National Historical Geographic Information System: Pre-release Version 0.1. Minneapolis, MN: University of Minnesota 2004.

person's place of residence in each year. The county is the most precise level of geography available for place of residence for all U.S. records in the 1880 and 1910 U.S. census data. This exemplifies the challenge of working with historical data, where the readily available information is often very limited. To make our study more manageable, we narrow the scope to the New York and New Jersey area. We chose this area because New York and New Jersey with several counties and urban areas in the late 19th century was an important hub for migrants.

We start by computing the total population for each county in New York and New Jersey using the NAPP 1880 complete count and 1910 sample census data. We then add the population counts to the NHGIS county boundary shapefiles. Next, we identify all cities listed in the IPUMS variable, CITY, in the two states in the two years. If a county has no cities based on the CITY variable, we proceed as if the population is evenly distributed throughout the county. In other words, the geographical centroid will be unchanged from our weighting process. If the county contains one or more than one city area, we assign the city's population, as calculated using the census files, to the corresponding polygons in the NHGIS city shapefile. For counties with city areas we subtract the total city population from the total county population to get the population for the 'city-free' part of the county. 61% of the counties in 1910 contained 1 or more than 1 city. 5% of the counties in 1910 were entirely covered by one city (e.g. New York City in Kings County) or an amalgamation of cities. The remaining 34% of counties in 1910 contained no city areas.

With the availability of Geographical Information Systems (GIS) technology, the method for computing a population weighted centroid is straightforward. We use ArcGIS9.3 to compute the weighted centroids based on the mean centers of the city areas and mean centre of the county. The mean center tool in ArcGis uses population to weight geographical centers of counties. If most of the population in a particular county lives in cities, the population weighted centroid shifts considerably towards the city and away from the geographical centre. The output of the analysis is a point shapefile of population weighted county centroids. We do this for both 1880 and 1910. Then we use the point distance tool in ArcGIS9.3 to create a large table of distances between every possible pair of weighted county centroids, 1880-1910. These are the distances we use to compare with the distances calculated without population weights for the current 1880-1910 IPUMS Linked Sample.

Preliminary results

Our weighted county centroids are illustrated in Figure 1 in red. In the figure, we highlight three different scenarios. In Map A, we show a part of New York State that has many counties with no city areas. In Map B we show the New York City area, where two entire counties are entirely city. In both Map A and Map B the population weighted centroid is located in the exact same place as the geographical centroid. However, in Map C, where we illustrate a part of New Jersey, the weighted centroid has shifted considerably because most of the counties have one or more cities. For example, the shift in the Essex County, New Jersey (marked by a purple star) centroid is 4 miles. For reference, Essex County has an area of approximately 124 sq. miles.

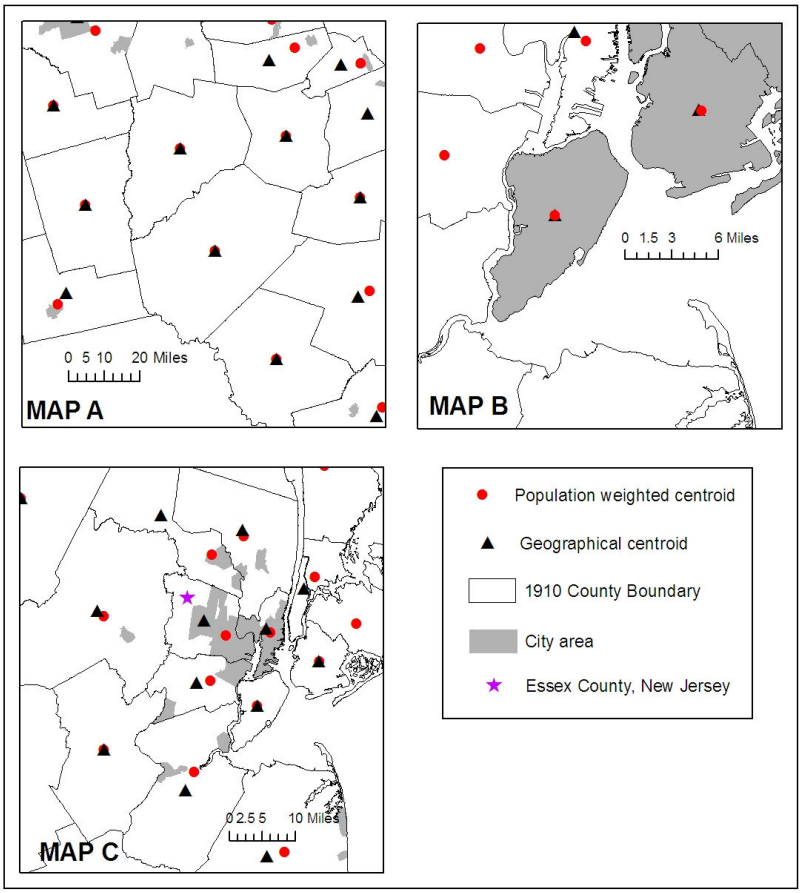

Figure 1: Map showing the shift in centroid location using geographical and population weighted methods. Map A shows a part of the state of New York where most of the counties contain no city areas. Map B shows the New York City area where the entire county is a city. Map C shows the part of the state of New Jersey where most of the counties have one or more city areas.

Next, we compare the distance calculated with and without population weights for all possible county combinations in our test area. There are 82 counties in 1910 in New York and New Jersey, and 81 counties in 1880 in New York and New Jersey, creating 6642 possible county combinations. We find that in 60% of the cases there was a change in distances between the two counties after applying population weights to the centroids (Table 1). Table 2 shows the degree of change in the 60% of cases identified in Table 1.

	No change	Any change	Total
N	2643	3949	6642
%	40	60	100

Table 1: Occurrence of change in county-to-county distance after applying population weights. The total N is the number of county combinations within the New York and New Jersey area, 1880-1910.

Change in distance (miles)	N	%
1 to 5	3270	83
6 to 10	650	17
> 11 (max = 16 miles)	29	Less than 1
Total	3949	100
Table 2: Amount of distance change for county combinations in Table 1 that had 'any change'.		

To answer our final question as to the importance of using population weighted centroids as opposed to geographical centroids for calculating migration distances for historical linked data sets, we use the 1880-1910 IPUMS Linked Sample of males. Table 3 shows the change in the miles migrated computed for the linked males using geographical centroids versus population weighted centroids.

Linked males in New York/New Jersey 1880-1910	N	%
Migration distance did not change	110	32
Migration distance changed	234	68
Total	344	100
Table 3: Change in migration distance for the IPUMS linked males, 1880-1910, after applying population weights to the geographical county centroids.		

Preliminary conclusions

By using population weighted county centroids, we make the migration distances in the IPUMS Linked Samples more accurate and spatial representations of travel distance more precise. We use the New York/New Jersey area as a test case. Although there is a change in migration distance for a substantial percentage of the linked males from the 1880-1910 IPUMS Sample, the degree of change (number of miles) is small. However, on average, the county areas for the eastern United States are substantially smaller than the county areas in the western United States. Our next step will expand on the finding in this preliminary study by testing the western states with larger counties, including California, Montana, or Arizona.