Computing error measures for migration distance estimates in historical linked data sets

Rebecca Vick, Minnesota Population Center, University of Minnesota

Sula Sarkar, Minnesota Population Center, University of Minnesota

Introduction

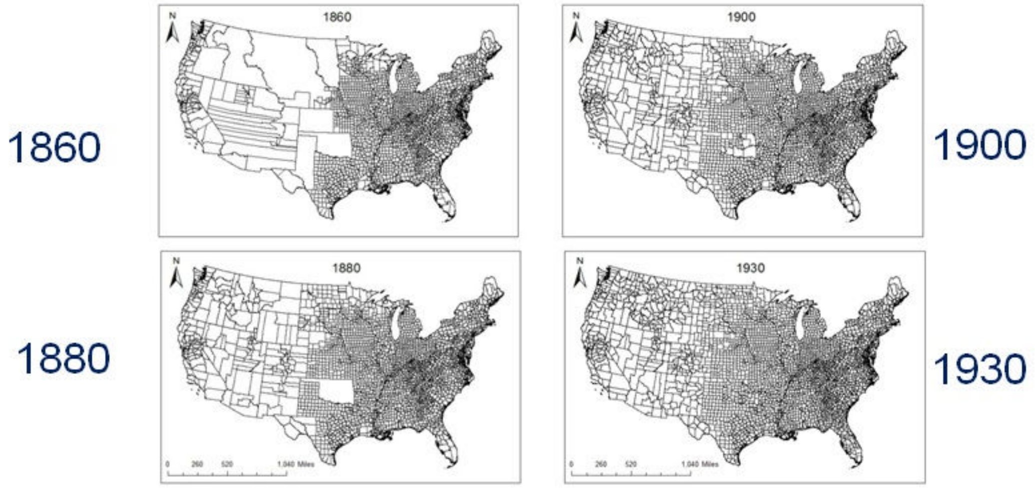
The Integrated Public Use Microdata Series (IPUMS) group at the Minnesota Population Center (MPC) has constructed twenty-one linked samples of United States census data from the 19th and early 20th centuries (MPC 2010). Each file contains a set of linked individuals at two historical time points. This longitudinal data will allow for new ways of researching population migration. Included in all twenty-one linked sample files is an IPUMS-constructed variable called MILEMIG, which contains the number of miles a linked individual migrated between census years. This distance measure provides a unique perspective on migrant patterns during this great migration period in United States history. However, the precision of the MILEMIG data is on the county level and should be used with caution. Users need to be aware of its limitations.

Providing a measure of error for MILEMIG would counteract researchers having a false sense of precision when using it in their research. Computing error measures for each U.S. division (as defined by the United States Census Bureau) separately would be particularly useful because in the U.S. there has always been great variation in average county size divisionally, particularly the further you go back as is shown in Table 1 and Figure 1. Furthermore, divisional error measures would be more relevant if a researcher is using a subset of the U.S. linked data that geographically represents only part of the country. The purpose of this paper is to describe one method of constructing divisional error measures for historical distance estimates such as those found in the MILEMIG variable and to discuss their potential importance to researchers.

Table 1. Descriptive county measures for each division, 1880

Division	**Number of counties**	**Median county perimeter in miles**
New England	67	162
Middle Atlantic	148	130
East North Central	423	99
West North Central	562	109
South Atlantic	485	116
East South Central	351	111
West South Central	357	140
Mountain	113	357
Pacific	107	281
Total	*2613*	*120*

Figure 1. United States county boundaries over time, 1860-1930



Background

A transportation revolution occurred in the 18th and 19th century with trains, which were then often referred to as an "annihilator of space and time". The revolution continued in the 20th century with automobiles. The human relationship with distance was changed during the nineteenth and early twentieth century by these extraordinary machines, making the study of distance and humans during this time period relevant to our history.

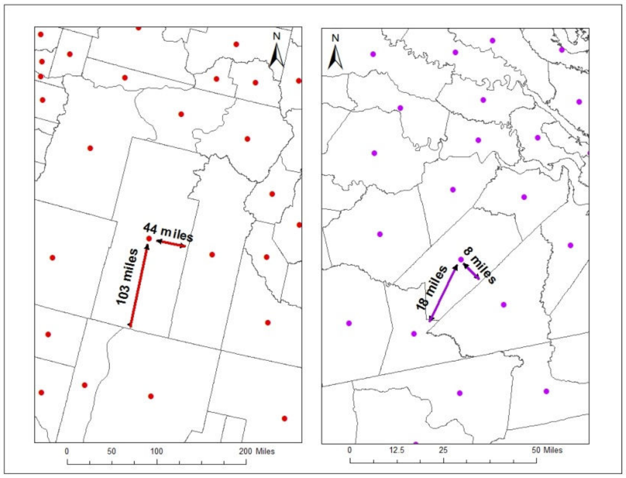
Literature on the 19th and early 20th century United States (U.S.) tells us that migration distance has been related to the demographic make-up of different migration flows in the past, and consequently, has had an enduring impact on the social, economic and cultural grain of the sending and receiving areas. Tolnay et al. (2005) found that whites migrating out of the South moved further than blacks migrating out of the South between 1910 and 1970 during the Great Migration, particularly in the first waves. Differentials in distance migrated between southern whites and blacks narrowed with time, likely due to improvement to the national transportation infrastructure and access to information (Tolnay et al. 2005; Tarver and McLeod 1970, 1976). Tolnay et al. also found that those who were literate and those who were single moved further than their counterparts. Daniel Price (1948) studied internal migration in the United States between 1935 and 1940 by loosely estimating migration distance.

The studies mentioned above measured distances between centers of regions and centers of states. Price (1948) measured distances between census divisions and Tolnay et al. measured between the centers of the state of birth and state of residence. This is problematic in that measurements on those scales miss moves within regions and states.

Including shorter distance moves in migration studies would add to the literature helping to expand our understanding of migration. The IPUMS Representative Linked Samples include data that will allow for study of intra-region and intra-state migrations. Each linked sample contains a migration distance variable — MILEMIG — that was constructed using county centroids.[1] We represented the "start" and "end" locations of each linked person with the centroids of their counties of residence in the two linked years. We systematically computed distances between every possible centroid pair for each pair of linked years in a Geographical Information Systems (GIS) creating look-up tables of distances that were used used to populate the MILEMIG variable. This is a highly generalized approach, which is obvious when we think about a person who lived very close to a county boundary being represented by a county centroid. Figure 2 provides a visual representation of this point.

[1] In this paper a centroid is the geographical center of an area (Vick & Sarkar, 2010).

Figure 2. Comparison of centroid-to-border distances in two different states. The counties in RED (on the left) are from the western state of Oregon and the counties in PURPLE (on the right) are from the eastern states of New York and New Jersey.



The rationale behind using county centroids for MILEMIG was that the county is the most precise location information that is standardized *and* available for *every* individual in the linked samples and is easily mapped in a GIS.[2] Although a migration distance computed with county centroids have a varying amount of imprecision among linked individuals in the IPUMS linked samples, it can provide meaningful results for migration studies if accompanied by a measure or accuracy or an error boundary that informs its use.

Reporting error is nothing new to the sciences. Common ways of reporting error in the social sciences include confidence intervals, p values, and standard error values. There are a myriad number of statistical tests that are used to evaluate data for the soundness of making a conclusion. But many of these methods do not easily translate to spatial data (Chrisman, 1991). The method we present in this paper is one way of doing so. We are sure there are other methodologies. We hope this paper will spark discussion about spatial error reporting particularly for historical spatial data.

[2] This information is easily mapped by using NHGIS historical county boundary files downloaded from http://usa.ipums.org/NHGIS.

Data

We acquired the key data files for our paper from the National Historical Geographic Information System (NHGIS) (MPC 2011). The NHGIS provides free GIS boundary files of the United States from 1790 to 2010. NHGIS Historical county boundary files for 1860 and 1880 contained the spatially referenced Inter-university Consortium for Political Social Research (ICPSR) state and county codes necessary for our project. We also used an NHGIS boundary file of the contiguous United States to create a uniformly-spaced grid of points spaced five miles apart covering the U.S. after first re-projecting the file into an equidistant projection that would limit distance distortion.

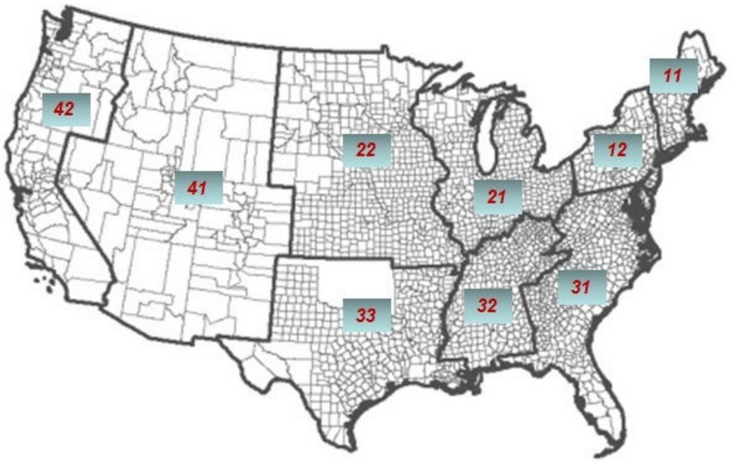
We also used the look-up table of all possible county centroid-to-centroid distances for 1860-1880 previously created at the Minnesota Population Center (MPC) to populate the MILEMIG variable in the IPUMS Representative Linked Files for that year pair. This file also contains state and county ICPSR codes.

Methods

We decided to calculate error measures for nine U.S. divisions[3]: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain and Pacific. We limited the error measures to divisions for several reasons. Making one error measure for the entire country given the method we chose would have been prohibitively demanding computationally, which will be explained further in this section. In addition, it would not serve users whose focus is a smaller area within the United States well. Computing error measures for each state would be significantly more demanding in the processing of the data in ArcGIS. Computing divisional error measures provides a balanced approach. They will provide more precision than one measure for the whole country, but will be specific enough to be selectively applied depending on the geographic area of interest. If a research project includes more than one division the user can simply use the largest divisional error measure applicable.

[3] We used the IPUMS detailed version of the variable, REGION to define the nine U.S. divisions (http://usa.ipums.org/usa-action/variables/REGION#description_tab).

Figure 3. 1990 Census Bureau divisions of the United States overlaying county boundaries from 1880



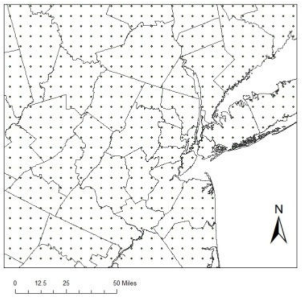
11 New England
12 Middle Atlantic
21 East North Central
22 West North Central
31 South Atlantic
32 East South Central
33 West South Central
41 Mountain
42 Pacific

To summarize our method, we measured the potential error in MILEMIG by taking a large sample of all possible migration distances and comparing them to the migration distances we calculated for the IPUMS linked samples. The large sample of distances was made using a uniformly-spaced grid of points spaced five miles apart. We calculated every possible inter-sample point distance within each division; then attached the corresponding IPUMS migration, or MILEMIG distance based on matching ICPSR state and county codes. We then took the difference between the two distances for each sample pair. The result was a frequency distribution of distance difference which can be used to describe the error bounds for each division.

We can describe the method in another way. The 'MILEMIG' value is the distance we use to approximate the real distance moved for a linked person because we don't know the exact location of that person's start and end points. We can think of each sample point pair as representing one possible migration that an individual could make. Taking the difference between the MILEMIG distance and the grid 'sample' distance we determine the amount of error in MILEMIG for a linked migrant who followed that particular sample line in real life. We use this distribution of distance differences to describe the accuracy of MILEMIG for an entire U.S. division. The way we chose to do this is to report the error bounds in miles at a certain percentile of the distribution for that division. For example, we can say that the MILEMIG distance for migrants who moved within 95% of the division is accurate to within +/-30 miles.

When creating the large sample of grid points we specified an inter-point distance of five miles. Our rationale for this spacing was to ensure that every county contained at least one point. The smallest county in the two linked years of 1860 and 1880 is Baltimore City in Maryland in 1880, which is 13.7 square miles and spans a maximum of approximately 5.25 miles. The resulting grid contained 120,151 points. Obviously, using a uniformly-spaced five-mile grid of points does not represent every possible move, however, there are many more locations represented by the grid than by all the county centroids, of which there are 2126 in 1860 and 2613 in 1880.

Figure 4. Five mile grid points.



The migration distance variable in the IPUMS linked samples is computed only for linked individuals whose county of residence changed from one census year to the other. For this reason, when computing error measures it did not make sense to include inter-grid point distances between points located within the same county, so we removed these computations from each divisional set.

As you can see in Table 2, using a five-mile point grid resulted in billions of calculations. This work could not be completed without the computer power of the MPC. It required so much data crunching that it would not have been practical if we did not have access to a server version of statistical software like, the one we used, SPSS. It also put our macro skills to very good use for the repetitive tasks in SPSS. We did not escape repetitive work however, because in ArcGIS, where our scripting skills were lacking, we had to break up the five mile grid into shapefiles that contained only 6,000 points in order to prevent ArcGIS software from crashing on our individual machines.

Table 2. Number of grid point distance calculations by division

Division	1860-1880 distance calculations
New England	6,630,626
Middle Atlantic	16,221,586
East North Central	97,488,301
West North Central	426,895,417
South Atlantic	118,494,108
East South Central	52,954,729
West South Central	302,116,542
Mountain	1,188,939,361
Pacific	166,345,506
Total	

Once each point-to-point distance had been calculated we applied a script to crunch the results. Essentially, the script combined the results each division into one large SPSS file, and based on the county and state ICPSR codes, attached the corresponding 'MILEMIG' distance and computed the differences.

Constraints

The method assumes that the population is evenly distributed, which is obviously not the case. But digitally mapping population distributions from the 19th and early 20th centuries on the county level would be very difficult. It may be possible to use modern data to make such digital maps, but the un-weighted method we used in this paper was doable immediately, whereas using population distributions would require much more time and resources. Another constraint is that the grid points are five miles apart and it is unclear how to incorporate that information into the error measures.

Results

Using the method described we computed error measures frequency distribution for the nine regional divisions. We show the +/- mile error bounds at different percentile of each divisional distribution in Table 3, below. One of the most widely-accepted benchmarks in many scientific areas is the 95% confidence interval. Since it is so commonly used and accepted it is what we highlighted in for each division.

Table 3. Maximum potential error for MILEMIG at different percentiles of land area by U.S. division, 1860-1880

	Maximum Error in Miles								
Percentile	**New England**	**Middle Atlantic**	**East North Central**	**West North Central**	**South Atlantic**	**East South Central**	**West South Central**	**Mountain**	**Pacific**
25th	7	5	4	10	5	4	10	30	15
50th	14	10	9	25	9	8	23	66	33
75th	25	16	15	84	16	14	62	124	64
95th	**46**	**27**	**28**	**223**	**33**	**23**	**148**	**232**	**129**
99th	62	36	43	352	56	30	229	299	190
100th	116	72	126	585	188	70	528	532	358

What these mile values portray is, for example, for those who moved within 95% of the New England division between 1860 and 1880, the migration distance measure in the IPUMS linked samples will be accurate to within +/- 46 miles; for those moving within 95% of the Middle Atlantic division, the distance will be accurate to within 27 miles, and so on.

We can see a vast difference in error values between the eastern and western divisions, as expected. But seeing these numeric values provides us with a more definite idea about the variation. You can also see that in the East in particular, users could reason that it would be appropriate to use the migration distances to, for example, study migration within states, but would not be appropriate to study migration to adjacent counties. Another way researchers can use MILEMIG to improve the fitness of use of the data for their particular purpose is to select cases whose MILEMIG is greater than a given value to ensure that they moved far enough to be considered a true migrant by the individual researcher's standards . For example, someone who wanted to make sure to exclude migrants who moved only a few miles over a county border, can select migrants whose MILEMIG values are greater than the error value at the 95th percentile for the applicable division with the largest error value. For example, If a researcher was interested in migrants who stayed in New England they could excluded those whose MILEMIG <= 46 miles.

This is a good spot to address an idea that has been brought up to us several times. The idea is to create a variable in the linked samples that indicates if a migrant moved to an adjacent county. This would be excellent information to include and a sure-fire way of excluding border-hopping migrants. Thus far we have not done any work to produce this information, partly because it would be more complicated that it seems due to changing county boundaries between linked years.

In addition to using the error measures for filtering records, they could serve as benchmarks for ‘real’ differences between different groups being analyzed, improving the confidence of results.

Focusing on a few divisions with high values- West North Central and West South Central — clearly the areas in what is now North Dakota and South Dakota, and Oklahoma are driving the error values up. Given that the linkable population moving in or out of those areas is very small it seems reasonable to recalculate the error measures for those two divisions after removing the data from these areas, and adding a footnote.

Conclusions

The MILEMIG variable in the twenty-one IPUMS Representative Linked Samples provides valuable migrant information over a long period of time and on a large geographic scale and is likely the only such measure readily available for historical migration studies. Although it is certainly not a perfect measure of migration distance, MILEMIG will be a relevant quantitative measure if used appropriately. Appropriate use will be informed by the divisional error boundaries calculated using the method described here. Although our method required a lot of computer time and resources, the intra-divisional point-to-point distances that we computed can be reused to calculate error measures for any set of boundaries that fit neatly into the divisions. There are certainly other methods of computing error, although the path for other methods is not clear in the spatial data literature. Ultimately, our error measures will help to inform users on the quality of the MILEMIG data and improve the communication of their research when the 95th percentile plus/minus error boundary is included in their results.

Future Work

In the future we would like to explore error measures for migration distances constructed using population weights. We also plan to mimic the study done by Tolnay and others on race differentials in distance migrated by Southern movers during the Great Migration with a few key changes. First, we will use the county-based distances instead of state-based; and second, our study will include *intra*-regional migration. This study will take advantage of the higher precision of the county-level distance measures by including short moves, something the current literature lacks, and, we will be able to apply the error measures we constructed with the methods explained here to inform our use of the data, help us make sound conclusions, and present our results clearly.

References

Chrisman, N. R. 1991. "The Error Component in Spatial Data" in Geographical information systems - Volume 1: Principles (D.J. Maguire, M.F. Goodchild, and D.W. Rhind, Eds.), Longman, New York, pp. 165-174.

Minnesota Population Center (MPC). 2010. IPUMS Linked Representative Samples, 1850-1930. Final Data Release. Minneapolis, MN: University of Minnesota, June 2010. Http://usa.ipums.org/usa/linked_data_samples.shtml.

Minnesota Population Center (MPC). 2011. National Historical Geographic Information System: Version 2.0. Minneapolis, MN: University of Minnesota. [http://www.nhgis.org]

Ruggles, S., et al. 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota.

Tarver, J. 1970. "Trends in Distances Moved by Interstate Migrants." Rural Sociology. 35:4 (Dec.) p.523.

Tarver, J. and McLeod, R. 1976. "Trends in Distance of Movement of Interstate Migrants." Rural Sociology, 41:1 (Spring) p. 119.

Tolnay, S., Curtis White, K., Crowder, K., and Adelman, R. 2005. "Distances Traveled during the Great Migration: An Analysis of Racial Differences among Male Migrants." Social Science History. 29:4 (winter), 523-48.