Historical Population Informatics: Comparing Big Data of Family Trees and the U.S. 1880 Census for Migration Analysis

Diansheng Guo
Department of Geography
University of South Carolina
709 Bull St., Columbia, SC 29229
+1(803) 777-5234
guod@mailbox.sc.edu

Alice Bee Kasakoff
Department of Geography
University of South Carolina
709 Bull St., Columbia, SC 29229
+1(803) 777-5234
Kasakoff@sc.edu

Caglar Koylu
Department of Geography
University of South Carolina
709 Bull St., Columbia, SC 29229
+1(803) 777-5234
koylu@email.sc.edu

Yuan Huang
Department of Geography
University of South Carolina
hyuan2011@gmail.com

Jack Grieve
School of Languages and Social Sciences, Aston University
j.grieve1@aston.ac.uk

ABSTRACT

Throughout history, migration has been a key force in shaping and transforming humanity. Many cultural, political, economic, and demographic patterns in the modern world can be traced back to the process of human migration. However, historical migration data over a long time span (e.g., the past 300 years) at either local or global scales have been scarce and difficult to obtain. In the past human migration has been studied primarily through qualitative analyses of survey data and archival sources, which are often too generalized or too small to reveal the spatiotemporal complexity of human migration patterns. To address this challenge, we collect, process and analyze massive amounts of user-contributed family trees on the web, combined with full-count historical U.S. census data, to obtain new datasets on human migration at a global scale over several centuries, with a focus on North America and Europe. This paper reports preliminary results of this on-going study, particularly focusing on the assessment of the representativeness of the family tree data, compared with the more complete and official census data of 1880, which is a full-count data (having each individual) with names, locations, and parents' locations.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining;* I.5.3 [**Pattern Recognition**]: Clustering—*algorithms, similarity measures;*

General Terms

Algorithms, Data Mining, Spatial Data, Migration

Keywords

Population, Migration, Family Tree

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015.

1. INTRODUCTION

Many cultural, political, economic, and demographic characteristics and patterns in the modern world can be traced back to the process of human migration. For example, it is well known that the origins of early settlers (migrants) in a region have had a strong influence on the formation of dialects. However, migration data over a long time span at either local or global scales are scarce, scattered in different sources, and have been difficult to obtain and analyze.

Nowadays, more and more historical sources are being digitized and shared, and interest in genealogy has grown rapidly among both academic scholars and the general public. Moreover, the general public has been using the available digital sources to trace their ancestors, build family trees, and share the trees on genealogy websites, which can be a source of large-scale and comprehensive data on historical population and mobility. However, it is extremely challenging for scholars to sift through such a vast amount of records to track movements of individuals.

We process a big data set of user-contributed family trees posted on ancestry.com in order to understand large-scale historical migration data of the US population. The aim is to extract information on the origins of immigrant groups and their subsequent migration within the United States. Such a dataset, if successful, will provide a unique and rich set of information on migration in a family context over a very long period of time extending back many generations. This was a period when many immigrant groups colonized in the United States and the economy changed from one based on farming to manufacturing and to the post-manufacturing regime of today. We want eventually to see if regional patterns of migration set down in earlier generations are reflected in the circulation of cultural information today.

2. RELATED RESEARCH

It has long been thought that family trees are only for the elite or the earliest groups to come to the US. Kasakoff et al. concluded that the data from family trees was not biased towards the wealthy

[1]. They compared a sample of New Englanders taken from family trees alive in 1850 who were descended from 9 progenitors who were among the earliest to settle New England since they had come before 1650, with men who had been born in that region taken from in the public use sample from the 1850 US census. The genealogical sample was slightly wealthier than men in the public use sample but this was to be expected since many in that sample had been descended from more recent immigrants, and fewer were farmers. So in the case of descendents from New England progenitors, a genealogical sample may be quite representative. But other groups might not be well represented. Registration of births and baptisms was more widespread in New England than in other regions of the US and those sources were made available in printed form in the 1930's to people who wanted to compile their genealogies. Many of the subsequent migrants to the US are hard to trace due to the existence of only a small number of surnames (e.g. Irish and Mexican). The African American population did not have surnames until emancipation and many could not read or write so their vital events were probably not registered. The slave censuses, in which they were listed before 1870, only contained first names and are difficult to match with the later censuses. Other groups might have experienced death rates that were so high that fewer descendents were left to compile trees.

Although there has been considerable research that has examined the regional structure in migration in different countries with different data, there is no systematic examination of the change and shift of migration patterns over time. The main challenge is related to the lack of migration data at a large scale and across a long time period. Systematic census survey data on migration are only available for the past decade or two. Therefore, past research on migration often rely on qualitative analysis with small datasets. In recent years, historical anthropologists and demographers have turned to *quantitative* methods to study migration with a range of newly digitized data sources such as manuscript census data, population registries, passport registers, ship passenger lists, and local parish records, which are now available at genealogy websites such as www.Ancestry.com.

More importantly, nowadays more and more users have produced data by contributing and sharing their research outcomes (i.e., family trees). But the ability to use such unprecedented and unique data sources is a challenge for research in the social sciences and humanities.

Migration analysis can be classified into two broad categories: (1) those that are model-based and (2) those that are data-driven and exploratory. The first category includes various spatial interaction models [e.g., 2, 3] and regression analysis methods [e.g., 4, 5], which are used to model, analyze, and/or predict migration using spatial and other derived factors. The advantage of such a model-based approach is the ability to work with small datasets, incorporate known theory, and be able to predict. However, model-based approaches have to assume a model (such as a gravity model), which can be mis-specified or based on inappropriate assumptions [6].

The second category is to discover spatial structures and patterns in migration with exploratory methods such as spatial clustering, graph partitioning, and flow mapping. Particularly, discovering migration regions is one of the most important research topics, as noted in [7] decades ago: "… *there is a fundamental spatial organization in the pattern of movements and suggest that the discovery of this inherent system of migration regions is the most profitable avenue of approach to the present problem*". A migration region is a spatially contiguous area with maximum internal migration flows and minimum external interactions, which is similar to the concept of "community structure" or modularity in complex network analysis or graph theory [e.g., 8, 9]. More broadly, a different category of "migration region" is defined as a collection of places sharing similar origins or similar destinations [10]. Such spatial structures in migration can be detected with either computational methods, such as spatial clustering [11-13] and graph partitioning [14, 15], or visual methods such as flow mapping.

To detect migration regions and understand its effects are of great interest in various application domains, including linguistics. For example, in history the U.S. settlement or migration patterns primarily extended East-West [16, 17] due to the constraints of crops which could be grown only in particular latitudes [18], and also to transportation–-most rivers drained mountains and ran across rather than down the continent. However, beginning in the 20^{th} century, there was a migration trend that shifted to a North-South pattern due to industrial opportunities in the North and, more recently, to retirement to warmer climates and the spread of industry to the American South. In addition there was migration to cities starting in the late 19^{th} century.

In this paper we collect and evaluate a big dataset of family trees to examine how each segment of the population is represented in the data, by comparing family tree data to the 1880 full-count census data of the U.S.

3. MIGRATION STUDY WITH FAMILY TREES

3.1 Family Trees

There are a number of genealogy websites that collect user-contributed trees. We focus on the data at Ancestry.com, which has over 700 *million individuals* organized in family trees, contributed and shared by users. This database of family trees are constantly growing and updated as each user does research on his/her family history. The spatial and temporal coverage of the data is worldwide with a concentration on North America and Europe. An individual may have a birthplace and date, death place and date, parents, wife/husband, and such information for their children.

There are a number of important properties for user-contributed data. First, the data may contain duplicate information on individuals, events (e.g., birth, death) and metadata on family relations (e.g., spouse, parent-child), due to various reasons such as overlapping trees, multiple versions of the same family, conflicting records, and referencing census records that represent multi-person events. Resolving such duplicate information can lead to more information for an individual by combining event records from different trees. Second, in order to link (match) individuals in different trees one can compare general information about an individual such as name, surname and birth year, as well as family connections such as father, mother, spouse and children. We have experimented with using the combination of first name, surname and birth year as a unique identifier for individuals.

To evaluate the potential bias in the family tree data we compare it to the US Census 1880. For this task, we processed the largest user-contributed trees (which contain more than 100 million individuals). To extract individuals that were likely to be alive by 1880 in the US we employ the following procedure:

1) Identify unique individuals based on unique combination of first name, surname and birth year, who
    a) have location information for at least one of their parents and have both location and birth year of each of their children; or
    b) have both birth and death years (with the year1880 in between); or
    c) have a birth year later than 1800 if the death year is missing (this is based on the assumption that those lacking death dates lived to an age of 80).
2) Extract the possible location of each individual and geocode locations.

Table 1 illustrates the statistics of individuals (up to 4 children) that we extracted in the first step of the process. These records represent individuals with (1) birth year and location; (2) at least one of the parents' birth locations; and (3) children's birth locations and years if available.

Table 1 Individuals that were likely to be alive in 1880

Information available	*# Individuals*
Birth year & location	12,653,120
Death year	5,503,534
Death location	4,506,833
Father's birth location	12,002,328
Mother's birth location	11,102,963
1st child's birth location	3,340,307
2nd child's birth location	2,734,769
3rd child's birth location	2,316,253
4th child's birth location	1,934,804

The potential location of an individual in 1880 was estimated based on the location of the nearest event to 1880 for the individual. The events may include the individual's birth and death, and his/her child's birth event. For example, the individual shown in Figure 1 had five events (birth, births of three children, and death) and we use the birthplace of the third child (born in 1884) as the location of this individual in 1880.

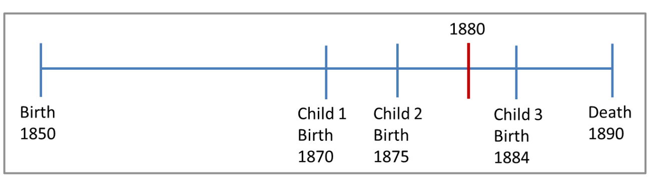

Figure 1: Estimation of an individual's location in 1880.

We built a procedure to geocode the locations with reference data from different sources: U.S. Gazetteer (https://www.census.gov), the U.S. Board of Geographic Names (http://geonames.usgs.gov), National Historical Geographic Information System (https://www.nhgis.org), Great Britain Historical Geographical Information System (http://www.gbhgis.org), and Wikipedia (https://www.wikipedia.org). For the 752,074 unique place names extracted from the family trees, 92% of all the names can be geocoded. With a manual examination of 1% of the geocoded names, we find that 98% of the geocoded names are correct.

With geocoded locations of each individual in family trees, we can map the migration of individuals and families over multiple generations for several centuries. For example, Figure 2 maps the migration of a single family tree, whose family members had migrated over two centuries (1700–1900) within and beyond the U.S., with a root in Massachusetts. Further research will synthesize the migration patterns of millions of families to discover time varying trends. In this paper, we focus on reporting preliminary assessment of the representativeness of the family tree data, compared with the more complete and official census data of 1880, which is a full-count data (having each individual) with names, locations, and parents' locations.

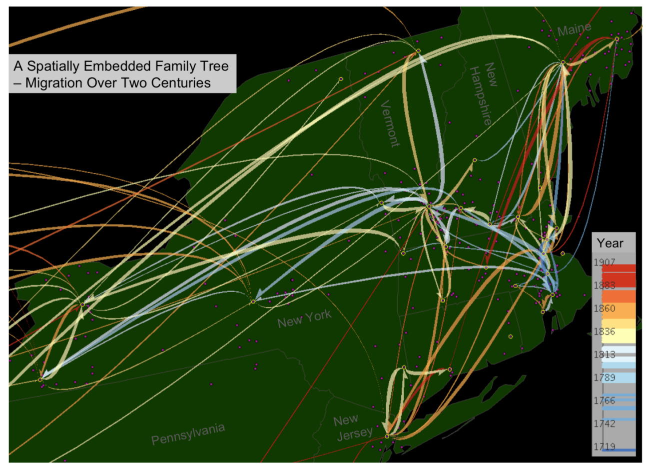

Figure 2: A migration flow map of a single family tree (1700–1900). Note: this map is a partial view that focuses on the Northeast. Line colors represent average time (with red being more recent) and width represents the number of migrants.

3.2 U.S. Census 1880

The 1880 U.S. census dataset is a complete count of population in the U.S., and thus it should be less biased than the user-contributed trees. We use this data set to evaluate the quality of the family tree data. Since the 1880 census was the first to ask for the birthplaces of the father and mother of each individual enumerated, we can construct shallow trees from the census to compare with the family tree data. Slightly more than 50 million individuals were enumerated in the 1880 census, of which 86.5% were born in the United States, 70% of their mothers born in the United States and 68 % of their fathers. The birthplace was listed as the state if it is inside the U.S. or as the country if outside the U.S.

Table 2 shows the percentage of births in each continent. Besides America, most people were born in Europe. There is little difference comparing mothers' and fathers' birth places, except that more mothers were born in the Americas. Table 3 shows the top 11 origin Europe countries of population based on mothers', fathers', and individuals' birthplaces.

Table 2: Birthplaces (by continent) for U.S. Population (and their parents) in the 1880 census.

Birth Place	*Individual birth place*	*Mother's birth place*	*Father's birth place*
Americas	88.10%	71.35%	69.50%
Europe	11.46%	24.96%	26.80%
Asia	0.21%	0.21%	0.21%
Oceana	0.01%	0.01%	0%
Africa	0%	0.02%	0.02%

Table 3: Births of U.S. Population (and their parents) in selected European countries in the 1880 census data.

Country	*Individual birthplace*	*Mother birthplace*	*Father birthplace*
Germany	3.86%	8.74%	9.25%
Ireland	3.69%	8.97%	9.52%
U.K.	1.83%	3.64%	4.14%
Sweden	0.39%	0.57%	0.58%
Norway	0.36%	0.65%	0.66%
France	0.25%	0.44%	0.54%
Switzerland	0.18%	0.35%	0.38%
Czechoslovakia	0.18%	0.30%	0.30%
Poland	0.13%	0.21%	0.22%
Demark	0.12%	0.19%	0.21%
Netherland	0.11%	0.22%	0.24%

3.3 Comparison of Family Trees and Census

The 1880 full-count census can be used to assess the representativeness of the family tree data. From the family tree data, we selected 3,048,418 individuals who we determined were alive in 1880 using the method described above. 92.41% of the individuals' birth places, and both of their parents' birthplaces were geocodable. Among them, 86% (2,423,045) of the individuals' birthplaces were in the U.S, of which the majority can be geocoded to the level of the state.

Table 4 shows the comparison of between the family tree data and the 1880 full count census based on mothers' and fathers' birthplaces for those who were born in the U.S. It seems that immigrants (whose parents were born outside the U.S.) were less represented in the family trees. For example, there were about 6% of the 1880 population having parents born in Ireland while the family trees only has less than 1%. This, however, can be due to the fact that we so far only processed the larger trees, while immigrants may have relative small trees. For example, German and Irish migrants came to the US later and thus there would be fewer descendants. When we finish processing all trees, this bias may be alleviated. Another potential reason may be that U.S. people are more likely to upload their family trees on website than people from other countries.

Table 5 shows the comparison of individuals' and parents' birthplaces within the contiguous U.S. between the family tree data and the 1880 full count census data. For individuals' location, the people born in the Northeast regions were relatively less represented, while the remaining regions were proportionally similar to the census data. For parents' locations, the family tree data is very similar to the census data. One of the possible reasons for this might lie in the procedure that we use to "guess" the location of an individual based on the nearest event (which might not be very near to 1880). Further and detailed inspection is needed to fine-tune the procedure and understand the potential bias.

For some states the percentages are quite different, for example, 5.92% of the individuals' birth places were in New York based on the family tree data, while 10.95% were in New York based on the 1880 census data. For some states, the percentages are quite similar when individuals', mothers', and fathers' birthplaces are all considered, for example, Illinois and Vermont. Difference in parents' percentages between family trees and the 1880 census are smaller than the individuals' percentages, which indicate family tree data may be a good data source for studies in historical migration.

Table 4 Comparison of mothers' and fathers' birth places for those born in the U.S. between the family tree and 1880 full count census

	Mothers' birthplaces		Fathers' birth places	
	Family tree	1880 census	Family tree	1880 census
US	93.78%	79.84%	92.65%	77.72%
Ireland	0.73%	5.85%	0.91%	6.15%
Germany	1.76%	5.68%	2.23%	6.56%
UK	1.80%	2.09%	2.09%	2.63%
Other European Countries	1.03%	1.78%	1.16%	2.09%
Canada	0.85%	1.05%	0.89%	1.08%
Other countries	0.05%	3.71%	0.06%	3.78%
Sum	100%	100%	100%	100%

Table 5: Comparison of family trees and 1880 census with individuals' and parents' birthplaces inside the contiguous US

	Individuals' birth places		Mothers' birth places		Fathers' birth places	
State	Family tree	1880 census	Family tree	1880 census	Family tree	1880 census
Northeast	**20.6%**	**31.1%**	**28.0%**	**34.1%**	**30.4%**	**35.3%**
Pennsylvania	7.38%	9.65%	10.2%	11.4%	10.78%	11.60%
New York	5.92%	11.0%	7.58%	10.6%	8.06%	10.67%
Maine	2.17%	1.73%	2.40%	2.18%	2.51%	2.25%
Massachusetts	1.68%	3.12%	2.37%	2.97%	2.74%	3.19%
Vermont	0.97%	0.99%	1.47%	1.45%	1.67%	1.60%
Connecticut	0.95%	1.25%	1.70%	1.66%	2.00%	1.81%
New Hampshire	0.73%	0.86%	1.08%	1.17%	1.29%	1.31%
New Jersey	0.57%	2.09%	0.94%	2.23%	1.02%	2.32%
Rhode island	0.20%	0.46%	0.29%	0.48%	0.32%	0.53%
South	**42.8%**	**38.7%**	**51.7%**	**48.7%**	**52.7%**	**50.0%**
Kentucky	7.39%	4.27%	8.92%	5.85%	8.87%	6.03%
Tennessee	6.63%	4.14%	8.47%	5.70%	8.57%	5.81%
Virginia	4.68%	4.92%	9.64%	8.93%	11.08%	9.70%
Georgia	4.09%	3.96%	4.92%	5.26%	4.76%	5.35%
Alabama	3.86%	3.04%	3.07%	3.15%	2.66%	2.97%
North Carolina	3.04%	3.78%	5.93%	6.00%	6.38%	6.64%
Texas	2.85%	2.08%	0.68%	0.61%	0.39%	0.41%
West Virginia	2.54%	0.99%	1.77%	0.72%	1.87%	0.72%
Arkansas	1.91%	1.19%	0.79%	0.55%	0.58%	0.42%
Mississippi	1.68%	2.43%	1.23%	2.11%	0.99%	1.88%
South Carolina	1.39%	2.73%	3.23%	4.40%	3.32%	4.71%
Louisiana	1.23%	1.87%	1.02%	1.63%	0.88%	1.48%
Maryland	0.84%	2.22%	1.64%	2.85%	1.93%	2.98%
Florida	0.51%	0.45%	0.25%	0.30%	0.19%	0.27%
Delaware	0.08%	0.36%	0.14%	0.46%	0.17%	0.49%
Oklahoma	0.06%	0.01%	0.02%	0.01%	0.01%	0.01%
DC	0.04%	0.23%	0.03%	0.12%	0.02%	0.11%

Midwest	**33.3%**	**28.6%**	**19.7%**	**16.7%**	**16.6%**	**14.3%**
Ohio	8.84%	7.60%	8.53%	7.51%	7.98%	7.07%
Indiana	5.69%	4.16%	4.02%	3.26%	3.47%	2.83%
Illinois	5.22%	5.17%	2.71%	2.21%	2.06%	1.69%
Missouri	5.40%	3.59%	2.70%	2.06%	2.02%	1.61%
Iowa	2.81%	2.19%	0.66%	0.42%	0.39%	0.25%
Michigan	1.74%	2.12%	0.58%	0.71%	0.43%	0.54%
Wisconsin	1.45%	2.04%	0.36%	0.42%	0.19%	0.24%
Minnesota	0.93%	0.78%	0.06%	0.04%	0.02%	0.02%
Kansas	0.84%	0.63%	0.05%	0.02%	0.02%	0.01%
Nebraska	0.32%	0.26%	0.03%	0.01%	0.02%	0.00%
Dakota	0.05%	0.05%	0.00%	0.00%	0.00%	0.00%
West	**3.33%**	**1.69%**	**0.59%**	**0.60%**	**0.36%**	**0.49%**
Utah	2.05%	0.20%	0.31%	0.03%	0.14%	0.02%
Oregon	0.47%	0.18%	0.08%	0.04%	0.04%	0.01%
California	0.34%	0.81%	0.05%	0.16%	0.02%	0.10%
Washington	0.19%	0.05%	0.09%	0.02%	0.10%	0.01%
Idaho	0.11%	0.02%	0.00%	0.00%	0.00%	0.00%
Colorado	0.04%	0.07%	0.00%	0.00%	0.00%	0.00%
Nevada	0.04%	0.04%	0.00%	0.01%	0.00%	0.01%
New Mexico	0.03%	0.26%	0.03%	0.32%	0.03%	0.32%
Montana	0.02%	0.02%	0.01%	0.00%	0.00%	0.00%
Arizona	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%
Wyoming	0.02%	0.01%	0.01%	0.00%	0.01%	0.00%

4. CONCLUSION

User-contributed Family trees are a unique source of data that can be utilized for research on migration. Most of the information on location can be geocoded to the level of the state at least within the United States. However, the larger trees that we processed so far may be biased towards the earlier arrivals from Europe. Once the smaller trees are processed we may find more information on African Americans, Mexicans and later arrivals from Europe.

In future research we will be using tree based methods to eliminate duplicates and also match specific individuals from family trees to the 1880 census in order to identify the family trees that are most complete and accurate to further specify the segment of the population that can be best studied with these sources. Currently we use the combination of first name, surname and birth year as a unique identifier. We applied this procedure to the 1880 census data and found that around 95% of the individuals in the census can be uniquely identified by this combination of first name, last name, and birth year. In future we will examine ways to further distinguish individuals with identical names.

Research on family trees can identify fundamental differences between migration regimes over long periods of time. We plan to extend this research to the UK, which had a very different migration regime. We expect smaller migration regions reflecting the longer period of settlement. Also the urban system was dominated by one city, London, which was not the case in the US. We expect this migration history to have led to more dialect differentiation than in the US. With fewer immigrants we also expect that family tree data will be more representative of the population there than it is in the US.

ACKNOWLEDGMENTS

This paper is partly based upon work funded by NSF Grant No. 0748813 and Digging Into Data Award LG-00-14-0030-14 by the Institute of Museum and Library Services (IMLS).

REFERENCES

1 Kasakoff, A.B., Lawson, A.B., and Meter, E.V.: ‘A Bayesian analysis of the spatial concentration of individual wealth in the US North during the nineteenth century’, Demographic Research, 2014, 30, pp. 1035-1074

2 Fotheringham, A.S., Brunsdon, C., and Charlton, M.: ‘Quantitative Geography’ (SAGE Publications, 2000. 2000)

3 Chun, Y., and Griffith, D.A.: ‘Modeling Network Autocorrelation in Space-Time Migration Flow Data: An Eigenvector Spatial Filtering Approach’, Annals of the Association of American Geographers, 2011, 101, (3), pp. 523-536

4 Chen, J.: ‘Internal migration and health: Re-examining the healthy migrant phenomenon in China’, Social Science & Medicine, 2011, 72, (8), pp. 1294-1301

5 Kley, S.A., and Mulder, C.H.: ‘Considering, planning, and realizing migration in early adulthood. The influence of life-course events and perceived opportunities on leaving the city in Germany’, Journal of Housing and the Built Environment, 2010, 25, (1), pp. 73-94

6 Sargento, A.L.M.: ‘Empirical Examination of the Gravity Model in two Different Contexts: Estimation and Explanation’, Jahrbuch für Regionalwissenschaft, 2007, 27, (2)

7 Ng, R.C.Y.: ‘Recent Internal Population Movement in Thailand’, Annals of the Association of American Geographers, 1969, 59, (4), pp. 710-730

8 Newman, M.E.: ‘Fast algorithm for detecting community structure in networks’, Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69, (6 Pt 2), pp. 066133

9 Newman, M.E.: ‘Finding community structure in networks using the eigenvectors of matrices’, Phys Rev E Stat Nonlin Soft Matter Phys, 2006, 74, (3 Pt 2), pp. 036104

10 Pandit, K.: ‘Differentiating between subsystems and typologies in the analysis of migration regions - a united-states example’, Professional Geographer, 1994, 46, (3), pp. 331-345

11 Slater, P.B.: ‘Hierarchical regionalization of state economic areas based upon migration flows’, Review of Public Data Use, 1976, 4, (6), pp. 32-56

12 Masser, I., and Brown, P.J.B.: ‘Hierarchical aggregation procedures for interaction data’, Environment and Planning A, 1975, 7, (5), pp. 509-523

13 Poon, J.P.: ‘The cosmopolitanization of trade regions: Global trends and implications, 1965-1990’, Economic Geography, 1997, 73, (4), pp. 390-404

14 Guo, D.: ‘Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data’, IEEE Transactions on Visualization and Computer Graphics (TVCG: Proc. of InfoVis'09), 2009, 15, (6), pp. 1041-1048

15 Thiemann, C., Theis, F., Grady, D., Brune, R., and Brockmann, D.Y.N.: ‘The Structure of Borders in a Small World’, PLoS ONE, 2010, 5, (11), pp. e15422

16 Crosby, A.W.: ‘Ecological imperialism : the biological expansion of Europe, 900-1900’ (Cambridge University Press, 1993. 1993)

17 Fischer, D.H.: ‘Albion's seed : four British folkways in America’ (Oxford University Press, 1991. 1991)

18 Steckel, R.H.: ‘The Economic Foundations of East-West Migration during the Nineteenth Century’, Explorations in Economic History, 1983, 20, pp. 14-36