Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Smoothing locational measures in spatial interaction networks

Caglar Koylu [1], Diansheng Guo *

Department of Geography, University of South Carolina, 709 Bull Street, Columbia, SC 29208, USA

ARTICLE INFO

Article history:
Received 12 July 2012
Received in revised form 6 February 2013
Accepted 2 March 2013
Available online 3 June 2013

Keywords:
Smoothing
Spatial interaction
Spatial network
Network measure

ABSTRACT

Spatial interactions such as migration and airline transportation naturally form a location-to-location network (graph) in which a node represents a location (or an area) and a link represents an interaction (flow) between two locations. Locational measures, such as net-flow, centrality, and entropy, are often derived to understand the structural characteristics and the roles of locations in spatial interaction networks. However, due to the small-area problem and the dramatic difference in location sizes (such as population), derived locational measures often exhibit spurious variations, which may conceal the underlying spatial and network structures. This paper introduces a new approach to smoothing locational measures in spatial interaction networks. Different from conventional spatial kernel methods, the new method first smoothes the flows to/from each neighborhood and then calculates its network measure with the smoothed flows. We use county-to-county migration data in the US to demonstrate and evaluate the new smoothing approach. With smoothed net migration rate and entropy measure for each county, we can discover natural regions of attraction (or depletion) and other structural characteristics that the original (unsmoothed) measures fail to reveal. Furthermore, with the new approach, one can also smooth spatial interactions within sub-populations (e.g., different age groups), which are often sparse and impossible to derive meaningful measures if not properly smoothed.

1. Introduction

Spatial interactions, such as migration and airline travel, naturally form a location-to-location network (graph). In the network a node represents a location (or an area) and a link represents an interaction (flow) between two locations. Locational measures, including both simple ones such as in-flow, out-flow, and net-flow and more complicated ones such as centrality, entropy and assortativity, are often derived to understand the structural characteristics and roles of locations in generating interactions. However, due to the dramatic differences in size (such as population) among locations and the small-area problem, locational measures derived with the original flow data often exhibit spurious variations and may not be able to reveal the true underlying spatial and network structures.

Scaling approaches such as iterative proportional fitting procedure (IPFP) are often employed (Clark, 1982; Pandit, 1994) to remove the confounding effects of origin and destination sizes on flows. However, such transformation procedures may distort the relative significances of nodes in a network (Fischer, Essletzbichler, Gassler, & Trichtl, 1993; Holmes, 1978). Alternatively, several studies have applied existing spatial kernel smoothing methods to remove spurious data variations (Porta et al., 2009; Sohn & Kim, 2010), which treat a locational measure (e.g., centrality) as a regular attribute and apply a traditional spatial kernel smoothing method to directly smooth the derived measure values. However, directly smoothing the measure values may generate unreliable or even misleading results for two main reasons. First, the original measure values may be unstable due to varying unit sizes and small flows between units. Second, traditional smoothing methods do not differentiate flows within and beyond a neighborhood and it is inappropriate to directly smooth original locational measures. For example, the net flow ratio (i.e., net flow/total flow) for a neighborhood (i.e., a group of contiguous spatial units) cannot be calculated as the average of unit-level net flow ratios within the neighborhood.

We introduce a new approach to smoothing locational measures in spatially embedded networks. For each location, the new method first smoothes the flows to/from that location considering flows to/from its neighborhood and then calculates its locational measure with the smoothed flows. The same procedure is repeated for each location, using the original flows (i.e., the smoothed flows for the previous location are not used). The neighborhood of a location is defined as the minimum set of nearest neighbors that meet a size constraint (such as a minimum population threshold or a distance threshold). To demonstrate the usefulness of the approach, we use the county-to-county migration data in the US and smooth the net migration rate and entropy measure for each

* Corresponding author. Tel.: +1 803 777 5234; fax: +1 803 777 4972.
E-mail addresses: koylu@email.sc.edu (C. Koylu), guod@sc.edu (D. Guo).

[1] Tel.: +1 803 777 5234; fax: +1 803 777 4972.

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.03.001

county. The smoothed results clearly help discover natural regions of attraction (and depletion) and a variety of structural characteristics that the original measures fail to reveal. Furthermore, we also smooth measures for sub-populations (e.g., different age groups), which can help discover not only distinctive regions of attraction and depletion but also show that attractiveness changes in both geographic space and multivariate space (e.g., migrants of different ages).

2. Related work

2.1. Locational measures

Locational measures (network/graph measures) have been extensively used in spatial interaction analysis to examine structural characteristics such as centrality (Hughes, 1993; Irwin & Hughes, 1992), entropy (Limtanakool, Schwanen, & Dijst, 2009), connectivity (Estrada & Bodin, 2008), assortativity and disassortativity (Fagiolo, Reyes, & Schiavo, 2009) and weighted clustering coefficient (De Montis, Barthelemy, Chessa, & Vespignani, 2007). Similar measures have also been introduced in application-specific domains such as migration. For example, many index approaches have been developed and used to quantify migration characteristics such as spatial focusing of migration streams (Plane & Heins, 2003; Plane & Mulligan, 1997; Rogers, 1992; Rogers & Raymer, 1998; Rogers & Sweeney, 1998). The index measures are usually derived for each location with the graph data (e.g., migration network). Commonly-used measures include net migration rate (Rogers, 1992), Gini index (Plane & Mulligan, 1997), coefficient variation (Long, 1988) and migration efficiency (Plane & Rogerson, 1991). However, due to the dramatic difference in unit size (e.g., population) and the small-area problem, derived locational measures often exhibit spurious data variations, and may conceal (instead of reveal) the true underlying spatial and network structures.

2.2. Iterative proportional fitting procedure (IPFP)

In order to remove the effects of location sizes on flows and capture patterns that are not necessarily associated with larger volumes, scaling approaches have been employed (Clark, 1982; Pandit, 1994; Slater, 1975). The most commonly used scaling approach is the iterative proportional fitting procedure (IPFP), which can be used to standardize a migration network by transforming the flows among locations so that all locations have the same inflow and outflow. Scaling does not change the cross-product ratio of the diagonal elements of the original matrix, and as a result the flow structure is preserved. However, IPFP transformation can distort the relative significances of nodes in a spatial interaction network in which the variability of node sizes is large (Fischer et al., 1993; Holmes, 1978).

2.3. Kernel density estimation and smoothing

Kernel density estimation or smoothing methods are commonly used for smoothing lattice spatial data, e.g., point- or area-based location attribute data, which are different from connection-based spatial interaction data. A spatial kernel smoothing method recalculates the attribute value of a location using a weighted average of the attribute values of its spatial neighbors (Borruso & Schoier, 2004; Carlos, Shi, Sargent, Tanski, & Berke, 2010), where the weight is calculated considering geographic distance. Alternative to spatial kernel smoothing, locally weighted average smoothing that uses a background value such as population to calculate weights is commonly used in smoothing disease rates (Kafadar, 1994; Shi, 2010). Bandwidth and kernel function selection are two important parameters in a spatial kernel smoothing method. The choice of the bandwidth determines the maximum radius (e.g., the extent of the neighborhood) or the number of neighbors that is considered to have an effect on the point of interest. The kernel function determines how each neighboring observation will be weighted and considered in the smoothing process. Previous research on kernel density estimation proved that the performance of the estimation is greatly affected by the choice of the bandwidth while the kernel function usually does not have a significance effect (Bors & Nasios, 2009; Silverman, 1986).

The most commonly used kernel functions include Gaussian kernel, triangular kernel, and Epanechnikov's kernel (Danese, Lazzari, & Murgante, 2008; Wand & Jones, 1995). There are two main types of bandwidth: *fixed* and *adaptive*. In a fixed-bandwidth approach, the radius that defines the extent of the neighborhood is assumed to be the same throughout the dataset. An adaptive bandwidth allows the radius to vary from one data point to another. Domain knowledge is commonly used to obtain a fixed bandwidth. However, it is widely acknowledged that a fixed bandwidth causes biased estimations for most spatial data sets, where the underlying density often exhibit significant spatial heterogeneity (Davies & Hazelton, 2010). Alternatively, various adaptive bandwidth approaches have been developed (Abramson, 1982; Carlos et al., 2010; Sain & Scott, 1996; Yang, Luan, & Li, 2010), which can be categorized into model-based and domain-based approaches.

In model-based bandwidth selection approaches, the goal is to improve a statistical model fit such as in geographically weighted regression. A statistical criterion is often used to provide guidance on selecting an appropriate bandwidth among a large number of possible bandwidth values (D'Amico and Ferrigno, 1990). Cross-validation (CV), Akaike Information Criterion (AIC_c) and Bayesian Information Criterion (BIC) are among the most commonly used criteria to select an appropriate bandwidth for local spatial statistics such as geographically weighted regression (Fotheringham, Brunsdon, & Charlton, 2002). In model-based approaches, an appropriate bandwidth is the one that gives the best model fit among a large number of possible bandwidth values. However, model-based approaches are not applicable for spatial smoothing in which there is no statistical model to fit and the goal is to smooth each unit with the neighborhood values. In domain-based bandwidth selection approaches, a relevant attribute (e.g., population) is used to determine the bandwidth. For example, to account for the underlying heterogeneous population distribution common in public health research, some studies (Carlos et al., 2010; Shi, 2009) have utilized a population threshold (i.e., the size for a neighborhood) to determine the adaptive bandwidth. Therefore, the bandwidth stops expanding when the threshold value is reached.

2.4. Smoothing network measures

Traditional smoothing methods introduced above have been adopted and used in transportation analysis research (Porta et al., 2009; Sohn & Kim, 2010) in order to accommodate the neighboring effect in calculating centrality measures. Existing smoothing practices treat the locational network measure (e.g., centrality) as a regular attribute and apply an existing spatial kernel smoothing method to directly smooth each locational measure with neighboring values. However, since a network measure summarizes the structure of the flow incidents on a node in a network, it is inappropriate to directly smooth measure values without considering the flow structure within and beyond the neighborhood.

3. Methodology

The new smoothing approach consists of four steps. First, for a location (node) s in a spatial interaction network, identify its spatial neighborhood N_s based on a geographic distance threshold (fixed-bandwidth) or a size threshold such as a minimum population (adaptive-bandwidth). The neighborhood N_s is represented with a gray circle in Fig. 1.

Second, *temporarily* remove the flows within the neighborhood, i.e., those with both origin and destination in the same neighborhood. Note that these flows are excluded only for this specific neighborhood and are still eligible for consideration for other neighborhoods. Then we weigh flows from/to the nodes (including s) in the neighborhood based on their distances to location s. The result is a smoothed sub-graph, in which flows to/from location s are modified considering flows to/from its neighbors. Fig. 1B illustrates the smoothed sub-graph of a location s where flows within N_s are removed and flows to/from N_s (shown by dashed lines) are weighted and partially considered as flows to/from location s.

Third, calculate the needed network measure for location s with the smoothed sub-graph. In other words, the weighted flows to/from the neighborhood are used in calculating the network measure for the location.

Fourth, repeat the above three-step process for each location (node). After the measure is obtained for a location, the smoothed flows are discarded and their original flows are restored. In other words, the smoothing (Step 2) is only temporary for each neighborhood.

In following subsections, we introduce each of the steps. To demonstrate the approach, we use county-to-county domestic migration data between 1995 and 2000 in the contagious US provided by census surveys, which includes 3075 counties (of the 48 continental states and Washington DC) and millions of migrants moving between these counties. Each data record has an origin county, a destination county, the count of migrants, and migrant characteristics, e.g., counts of migrants for each income level or age group that move from the origin to the destination.

3.1. Bandwidth selection

There are two potential alternatives for choosing the bandwidth. If applicable to the context of the spatial interaction network, a domain-based approach could be employed, which uses an attribute and a threshold value to configure the size of a neighborhood, e.g., the population or total flow of a neighborhood. Alternatively, a data-driven approach could be employed to determine the bandwidth according to the properties of the spatial interaction network. In this research we primarily focus on the first approach (domain-based) to configure neighborhood and discussed the alternative (data-driven) approach in the conclusion section.

In spatial interaction data, locational measures can be sensitive to the volume of flows or population of involved locations. It is more meaningful to make each neighborhood be of a similar and sufficiently large size so that the flows to/from different neighborhoods can be compared. Therefore, we employ a domain-based approach and use a population threshold to determine the adaptive bandwidth (or neighborhood size) for each unit. Other than population, the total volume of in-flow or out-flow may also be used for defining the size threshold. The choice depends on its applicability to the locational measure. For example, a net migration rate represents the net-flow of a location normalized by its population, in which case it makes sense to make each neighborhood have a similar population.

Let the population threshold be p. The neighborhood N_s of a location s is the smallest set of nearest neighbors that has a total population $P(N_s)$ greater than p. Specifically, the neighborhood N_s for unit s is constructed with two steps: (1) initially, let $N_s = \{s\}$and sort all other units based on their distance to s; (2) the nearest neighbors are added to N_s until $P(N_s) > p$. The bandwidth for s is then the distance to the farthest unit in its neighborhood N_s.

For cases where the population attribute does not exist or is inappropriate for the context of the analysis, alternative variables can be used to define neighborhood (bandwidth). For example, the in-flow entropy measure quantifies the diversity of flows that go into a location. Thus, it is appropriate to use the total in-flow to a neighborhood (excluding flows within the neighborhood) to define the bandwidth in calculating the in-flow entropy measure. Similarly, for the out-flow entropy measure we may use the total volume of out-flows from a neighborhood to define the adaptive bandwidth.

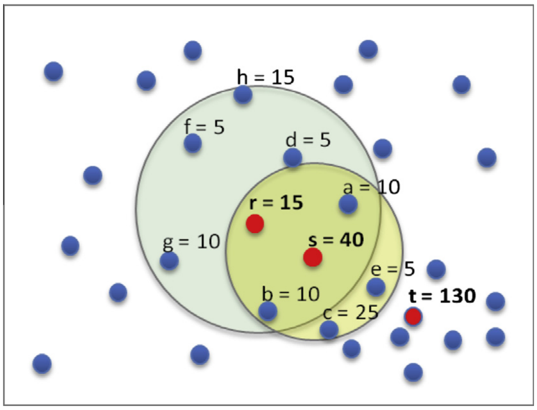

Fig. 2. Illustration of the bandwidth selection process. The neighborhood N_s of a location s is the smallest set of nearest neighbors that has a total population $P(N_s)$ greater than a given population threshold p, which is 100 in this example. The map shows the neighborhoods of three locations r, s, and t.

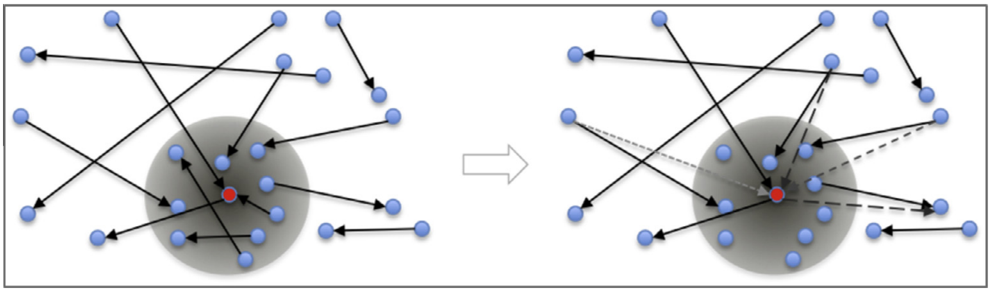

Fig. 1. An illustration of the smoothing approach for spatial interaction data. The left map shows the original data. The map on the right shows smoothed flows related to a location (in red, at the center of the circle) and its neighborhood (gray circle). Dashed lines represent weighted flows to/from the neighborhood that are now partially considered as flows to/from the location in red and used in calculating the network measure for the location. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

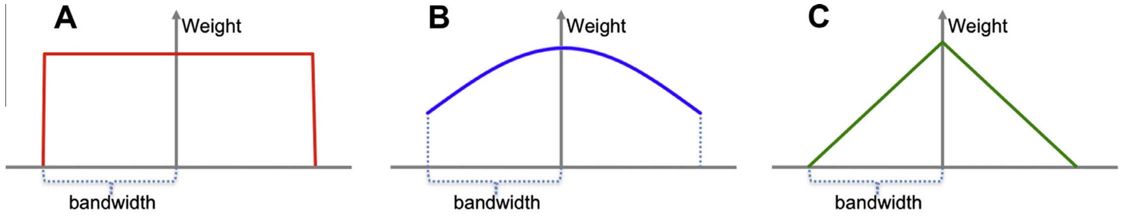

Fig. 3. Three commonly used kernel functions. (A) Uniform: $W_{si} = 1$ if $d_{si} \leqslant B_s$; else 0. B_s is the bandwidth and d_{si} is the distance between location s and its neighbor i. (B) Gaussian: $W_{si} = \exp(-(d_{si}/B_s)^2)$ if $d_{si} \leqslant B_s$; else 0. (C) Triangular: $W_{si} = 1 - |d_{si}/B_s|$, if $d_{si} \leqslant B_s$; else 0.

Fig. 2 illustrates the bandwidth selection process with a simple data set. Let the population threshold $p = 100$. Three nodes r, s, t are highlighted and their population values are $P(r) = 15$, $P(s) = 40$ and $P(t) = 130$. Since node t is sufficiently large, it forms a neighborhood by itself and thus no smoothing is needed. Nodes r or s need to add neighbors to meet the threshold p. Following the procedure outlined above, we have $N_r = \{r, s, d, a, b, f, g, h\}$ and $N_s = \{s, a, e, r, b, c\}$, with $P(N_r) = 110$ and $P(N_s) = 105$.

Choosing the population threshold for determining bandwidth involves a tradeoff between over-smoothing and under-smoothing. On one hand, the bandwidth should be sufficiently large to avoid artifacts caused by small neighborhood and under-smoothing. On the other hand, if the neighborhood is too large, interesting local patterns may disappear. Smoothing results change in a predictable way with decreasing/increasing bandwidth, with larger bandwidths generating more smoothed result (we will show the experiments with different bandwidths in Section 4.4). For the county-to-county migration data of the US, we experimented with different population thresholds to examine flow patterns at different scales and chose a population threshold of one million, which is about the population of a medium-sized metropolitan area.

3.2. Smoothing flows

For a specific location s and its neighborhood N_s, $s \in N_s$, we smooth the flows that go into or out of the neighborhood. Let B_s be the bandwidth. First, flows within N_s are temporarily removed, i.e., a flow is ignored if its origin and destination are both in N_s. Removing flows within the neighborhood is necessary because the entire neighborhood is considered as a single unit (i.e., location s) in calculating a network measure. Second, a kernel function is incorporated to weigh each flow from/to N_s based on the distance between s and the flow origin or destination (whichever is inside N_s). In other words, each flow to/from N_s is partially (according to its weight) considered as a flow to location s even if the flow does not directly involve s in the original data, which essentially reassigns weights to existing edges or adds new edges to location s.

The most commonly used kernel functions include the uniform kernel, the Gaussian kernel and triangular kernel (Fig. 3). Previous research and our experiments show that the smoothing results are not significantly affected by the choice of models (Bors & Nasios, 2009; Silverman, 1986). In this research, we have experimented with the above three models and report the results using the Gaussian kernel.

In Fig. 4 we show an example of a smooth graph that includes the connections to/from a location s (in red[2]) and its neighborhood N_s (gray circle). In addition to edges (a, s) and (s, k) that exist in the original data, the smoothed sub-graph for location s also has newly added edges (b, s), (c, s), (j, s) and (s, i), which will be included in calculating the location measure for s. The value for the new "flow" (b, s), for example, is the product of the value of flow (b, e) and its weight W_{se} according to a chosen kernel model. Note that flows within N_s are ignored.

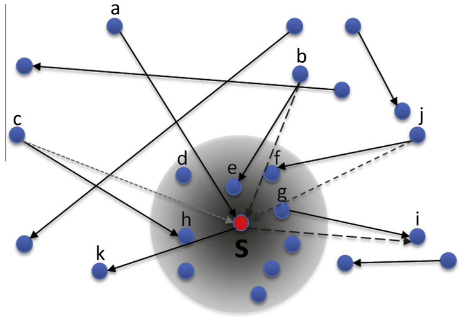

Fig. 4. An illustration of a smoothed sub-graph. Dashed lines are newly added edges.

3.3. Calculating a locational measure

Using the smoothed sub-graphs, it is straightforward to calculate a variety of network measures for the focal location, which are more stable (with less spurious variation) than those calculated without smoothing. Here we use the net migration rate and an entropy measure as two case studies to demonstrate the approach and evaluate its results.

3.4. Net migration rate

Net migration rate is the difference between in-migration (inflow) and out-migration (out-flow) of an area in a period of time, divided by the population of the area. Net migration rate is usually multiplied by 1000 to represent the number of migrants per 1000 inhabitants. To obtain a smoothed net migration rate for a neighborhood, we smooth the flows for the neighborhood (as introduced earlier), calculate the inflow and outflow of the neighborhood with the smoothed graph, and then divide (inflow – outflow) with the total weighted population of the neighborhood of s, denoted by $P(N_s)$. In other words, the same weighting is applied to both the flows and the population.

3.5. Entropy

The variation of flow volumes across the links to/from a location can provide important insights about the structure of the network and the characteristic of the location. Local entropy measures (Limtanakool et al., 2009) are often used for this purpose. Entropy of a location s (i.e., its neighborhood) is calculated using the formula in the following equation:

$$EI_s = -\sum_{j=1}^{J} \frac{x_{sj}\ln(x_{sj})}{\ln(J-1-n)} \qquad (1)$$

[2] For interpretation of color in Figs. 4–6 and 12–14, the reader is referred to the web version of this article.

where EI_s is the Entropy Index of location s, x_{sj} is proportion of flow sj in relation to the total flow connected to s, J is the total number of locations in the network, and n is the number of locations inside the neighborhood N_s. The maximum number of links that location s may have is $J - 1 - n$. EI measures the variation in the magnitude of interactions across the connections of a node. The index value ranges between 0 and 1. A small inflow entropy value indicates that the flows to the location vary greatly (with large flows from a few locations and small flows from elsewhere), whereas a large inflow entropy value indicates that a location receives similar amount of flows from all locations. Entropy can also be calculated for outflows or all flows (inflow and outflow together).

With the county-to-county migration data in the US, we calculate and map the smoothed net migration rate and the entropy measure for each county, which clearly help discover natural regions of attraction or depletion and a variety of structural characteristics that the original measures fail to reveal. Furthermore, our smoothing method make it possible to calculate measures for a subset of flows (e.g., flows of a specific age group), which are impossible to obtain without smoothing due to the small-area problem.

4. Results

4.1. Smoothed net migration rate

In this section, we show the smoothed net migration rates and compare them to the original measures. For the county-to-county migration dataset of the US, we chose a population threshold of one million, which approximates the population of a medium-sized metropolitan area. To enable comparison of the two measures, we used a custom classification for both in which 0 was chosen as the critical midpoint and the Jenks natural breaks classification was applied separately to each side of the midpoint. A diverging color scheme is used to represent different value ranges, with red representing attraction and blue for depletion (i.e., negative net migration rate).

The original net migration rates are shown in Fig. 5, in which it is difficult to distinguish regions of attraction and depletion because of unstable values caused by the dramatic population differences among counties and the small-area problem in the data. On the contrary, the smoothed net migration rates (with a neighborhood size of one million population) shown in Fig. 6 can clearly reveal the regions of attraction and depletion with differing magnitudes. For example, major attraction regions (i.e., those of darker red hues) include Florida, Arizona, Greater Las Vegas region, north-east outskirts of the Atlanta metropolitan area, counties surrounding Denver, Dallas, Houston and San Antonio, and the metropolitan counties in North Carolina. On the other hand, large metropolitan counties such as Los Angeles, New York City, Chicago and Miami and rural counties in Montana, North and South Dakota can be easily recognized as regions of depletion. The smoothing results also reveal contrasting patterns locally within metropolitan areas, such as Chicago, Denver, Washington DC, Dallas and Miami, where the core metropolitan areas have a push effect on migrants while the counties surrounding these core metropolitan areas have a pull effect on migrants as a result of suburbanization and urban sprawl.

4.2. Smoothed net migration rate for sub-populations

Migration patterns of sub-populations such as different races, ethnicities or age groups are expected to be spatially and structurally different from each other. Locational measures for sub-population flows are even less reliable because of much smaller volumes of flows and small base populations. To illustrate the effectiveness of our approach to overcome this problem, we smooth the flows within each age group, calculate net migration rate with the smoothed flows and compare them to their original net migration rate results. After examining the smoothing results for each age group, we focus on two age groups, namely 20–24 and 25–29, because they have the highest mobility and distinctive migration patterns (we will explain this below in Fig. 9). The original net migration rates for age groups 20–24 and 25–29 are shown in Figs. 7 and 8. It is difficult to interpret both maps because of unstable measure values that have spurious variation.

After we smooth net migration rates within each age group, we use box-plots of the smoothed measure results to give an overall understanding of migration behaviors for different age groups by comparing their distributions (Fig. 9). One of the most interesting and contrasting patterns that can be observed in Fig. 9 are those for age 20–24 and age 25–29. On one hand, the age group 20–24 has a large number of outliers with very high positive net migration rates and a larger upper quartile with a median around 0. On the other hand, age group 25–29 has a lower median below 0, a larger lower quartile and some outliers with negative net migration rates. The migration flows within these two age groups are likely related to education and employment specific flows. From Fig. 9, we may also observe patterns related to elderly migration (Plane & Jurjevich, 2009; Rogers & Sweeney, 1998). For example, there are outliers that disproportionately attract migrants of age groups 55–75. We can map the net migration rates for each age group to further investigate the observed patterns. Due to limited space, we only show the smoothed results for age groups 20–24 (Fig. 10) and 25–29 (Fig. 11).

The smoothing results highlight distinctive patterns that agree with existing migration studies. For example, migration of students and young adults for education and employment purposes (Slater, 1976; Whisler, Waldorf, Mulligan, & Plane, 2008) can be seen clearly in Figs. 10 and 11. While metropolitan areas attract age group 25–29 because of employment opportunities, places with a substantial number of universities attract age group 20–24. This divide can be seen in many places across the country. For example, in Texas, though the region surrounding Austin attracts age group 20–24, there is an opposite tendency among age group 25–29 to move away from this region and possibly targeting the Dallas Metropolitan area. A similar pattern is also observed in Florida. Because of the presence of many university campuses, the region that includes counties around Tallahassee, Gainesville and Jacksonville in Florida attracts age group 20–24, whereas age group 25–29 migrate from this region, targeting the Orlando and Miami Metropolitan areas for jobs. Also, metropolitan areas including Las Vegas, Atlanta, Raleigh, Charlotte, Denver and Minneapolis also attract both of the age groups 20–24 and 25–29.

4.3. Smoothed entropy

In addition to discovering regions of attraction and depletion, it is also important to gain insight into the structure of flows. A variety of measures such as entropy (Limtanakool et al., 2009), Gini index (Plane & Mulligan, 1997) and coefficient variation (Long, 1988) could be used to measure the diversity of flow volumes among the links to/from a location. In this section, we use the inflow and outflow entropy measures to capture the differentiation of the magnitude between the links to/from each location. We also compare the smoothed entropy measures to their original measure results. We use the total volume of in-flow to determine the neighborhood in calculating in-flow entropy whereas we use the total volume of out-flow to determine the neighborhood in calculating out-flow entropy. To balance between over-smoothing and under-smoothing we heuristically chose 100,000 as a threshold volume both for in-flow and out-flow values to determine the neighborhood for

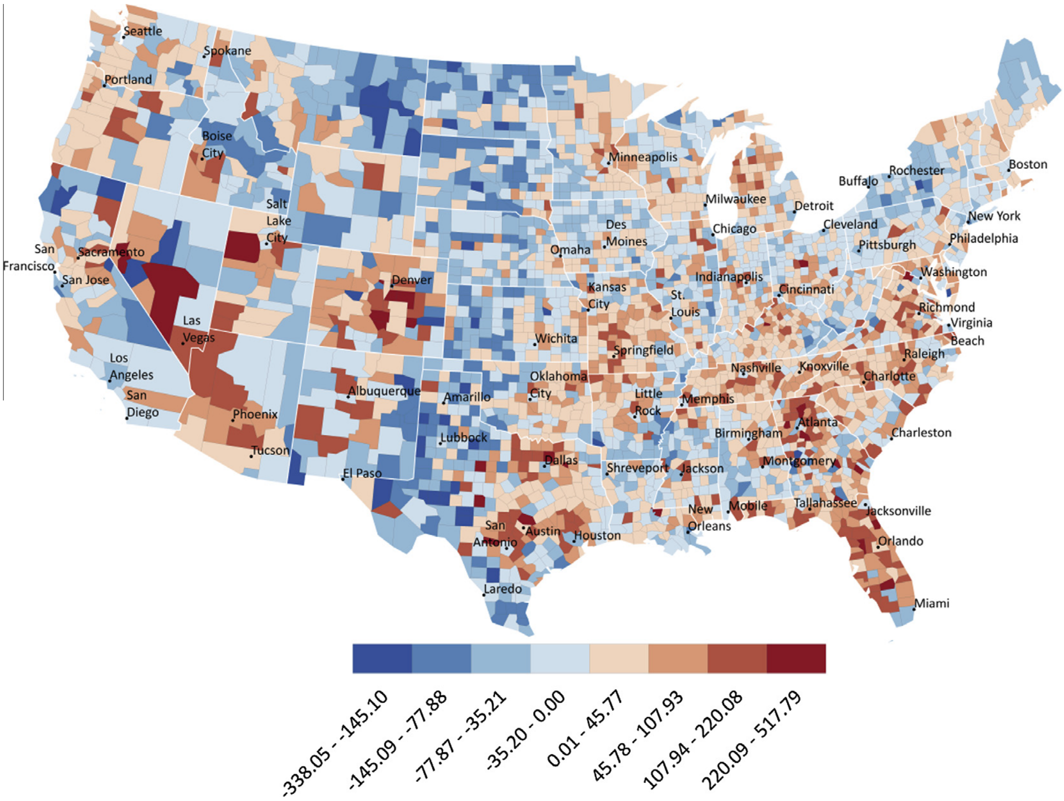

Fig. 5. Original net migration rates.

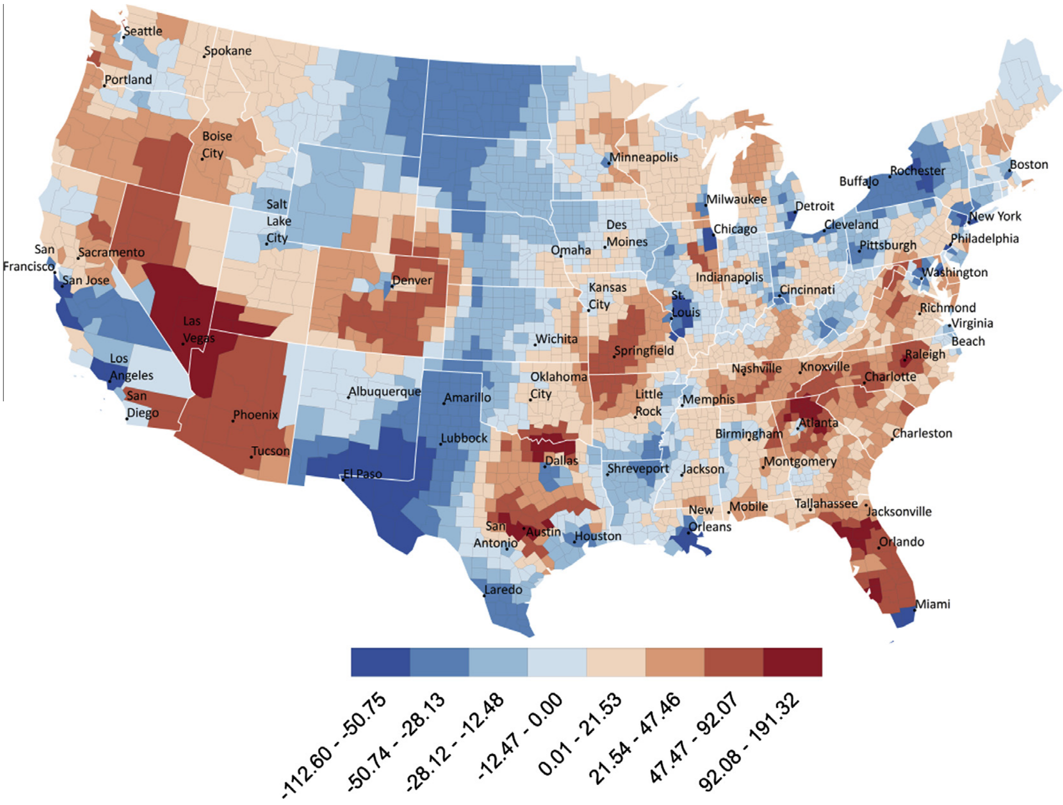

Fig. 6. Smoothed net migration rates.

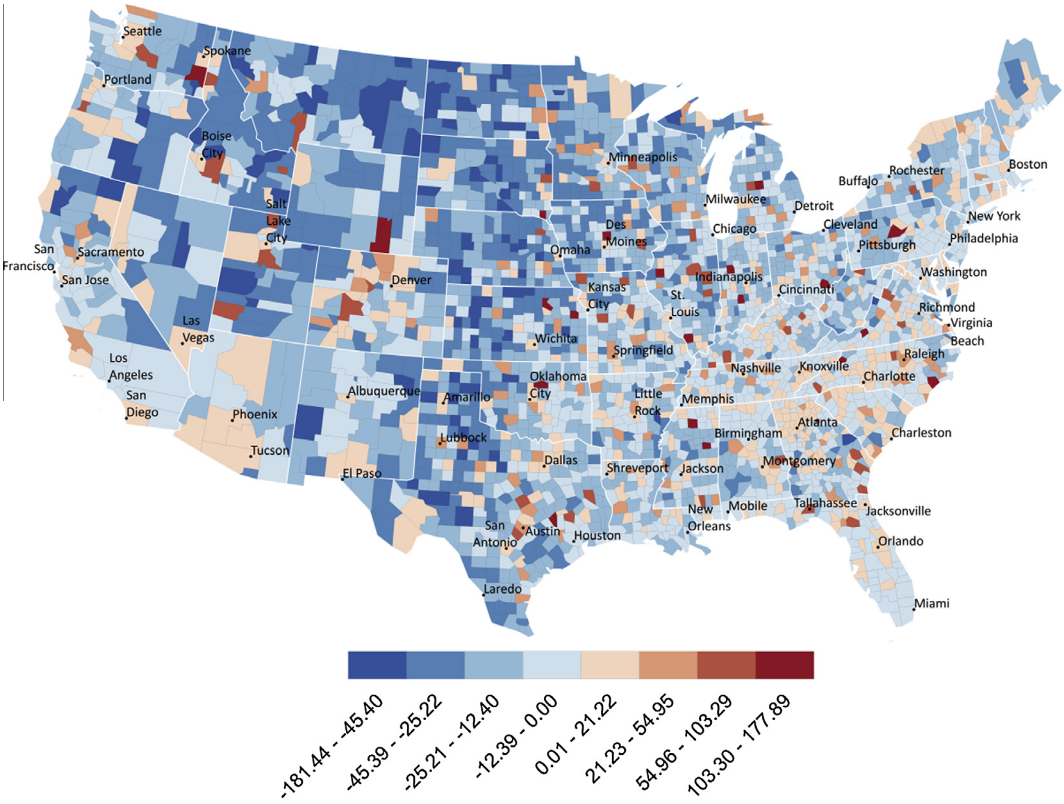

Fig. 7. Original net migration rates for age group 20–24.

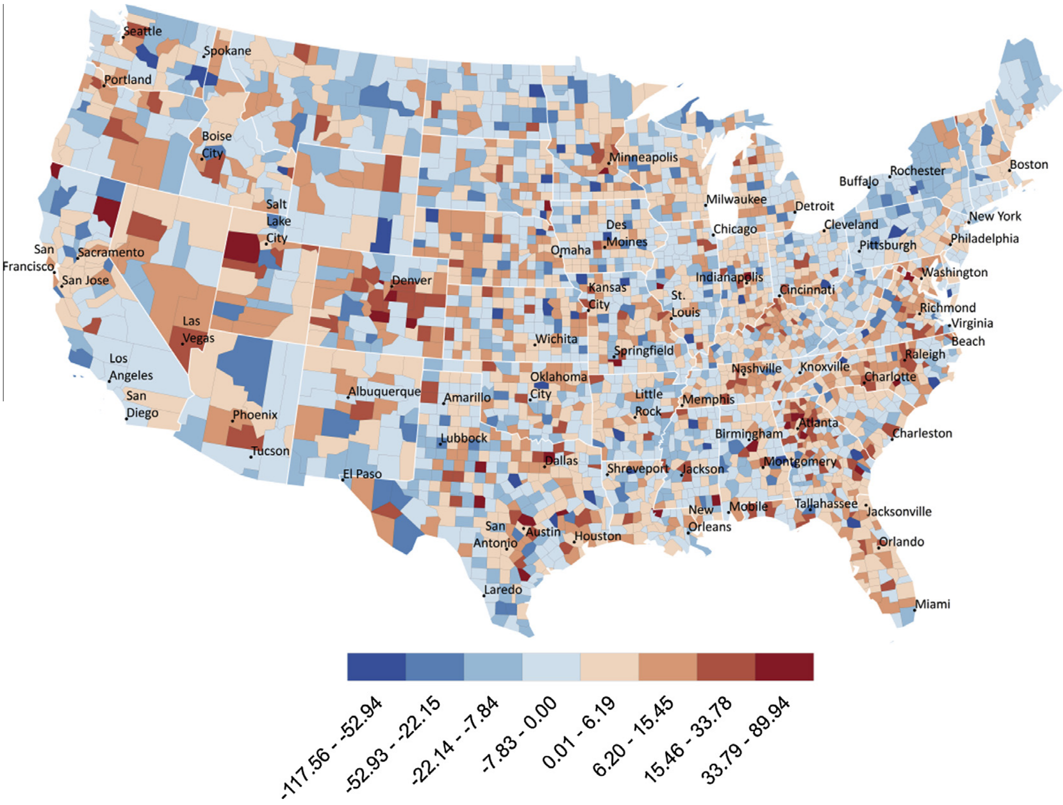

Fig. 8. Original net migration rates for age group 25–29.

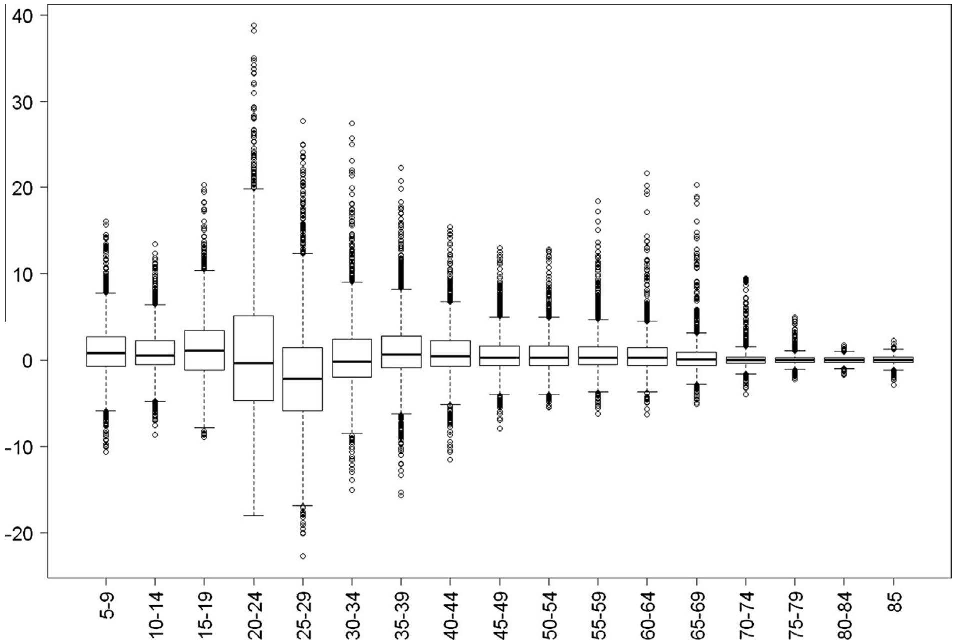

Fig. 9. Box-plots of smoothed net migration rate results for age groups.

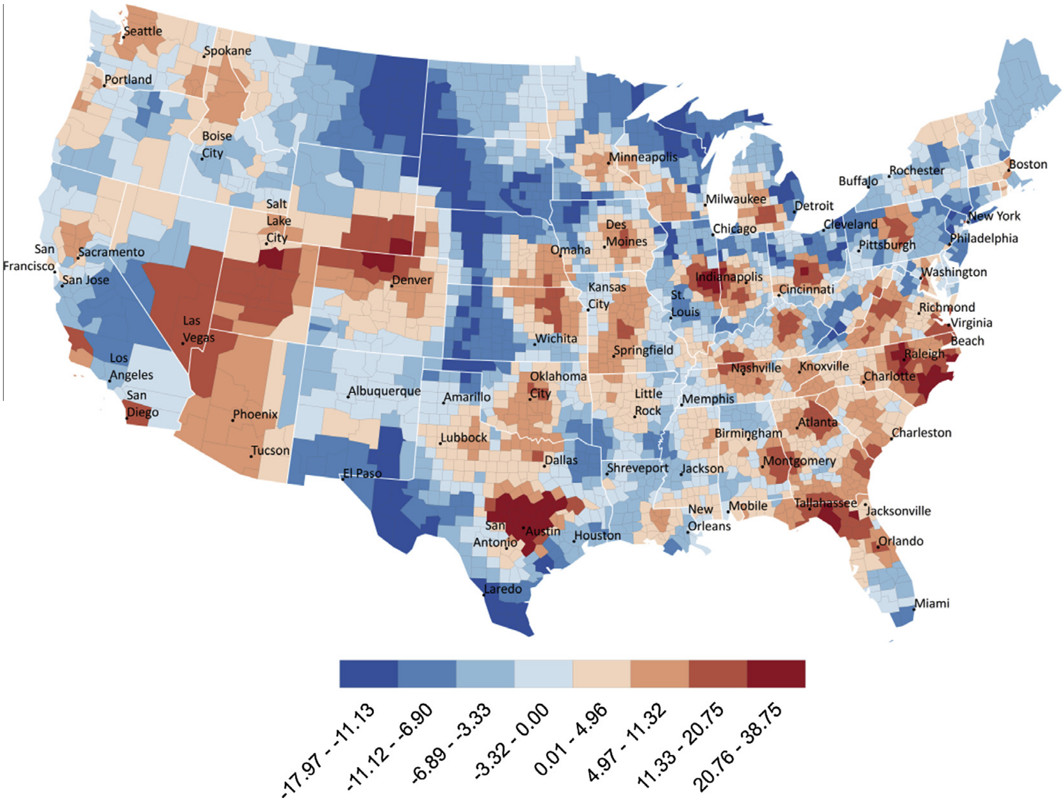

Fig. 10. Smoothed net migration rate for age group 20–24.

calculating in-flow and out-flow entropy measures. To enable comparison, we again use the Jenks natural breaks classification with a sequential color scheme, with darker colors of red representing low entropy values (which highlight spatially focused (targeted) flows) and darker colors of blue for high entropy values (which show more evenly spread flows to/from the other locations).

The original inflow entropy and outflow entropy are shown in Figs. 12 and 13, respectively. Both maps are difficult to interpret because of large and spurious variation in measure values. Moreover, the entropy values correlate with size and, as a result, smaller counties have always relatively low entropy since they normally have much less links than larger counties (see Eq. (1)). Similarly,

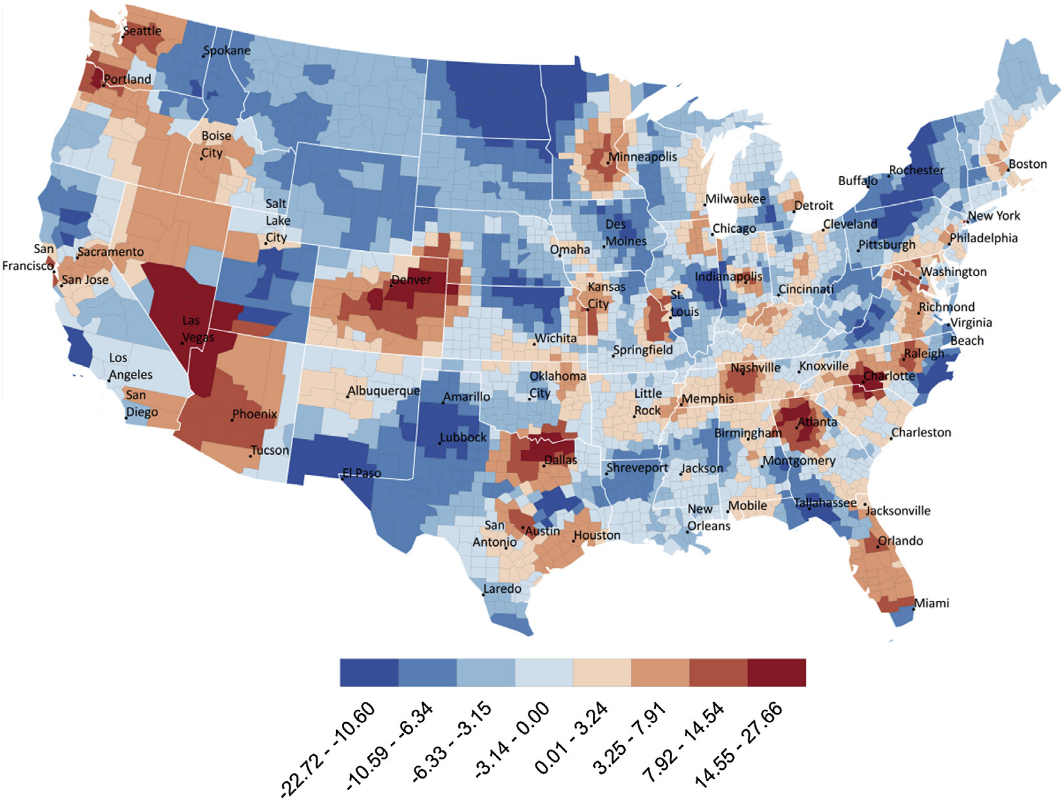

Fig. 11. Smoothed net migration rate for age group 25–29.

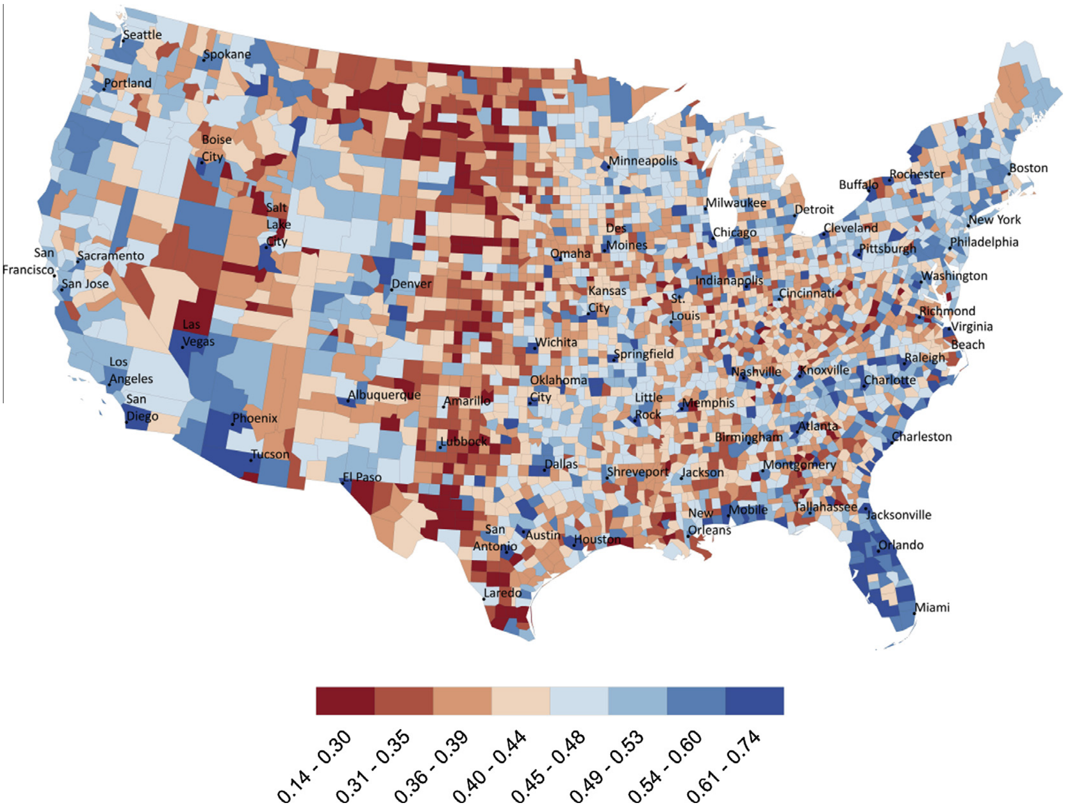

Fig. 12. Original in-flow entropy values.

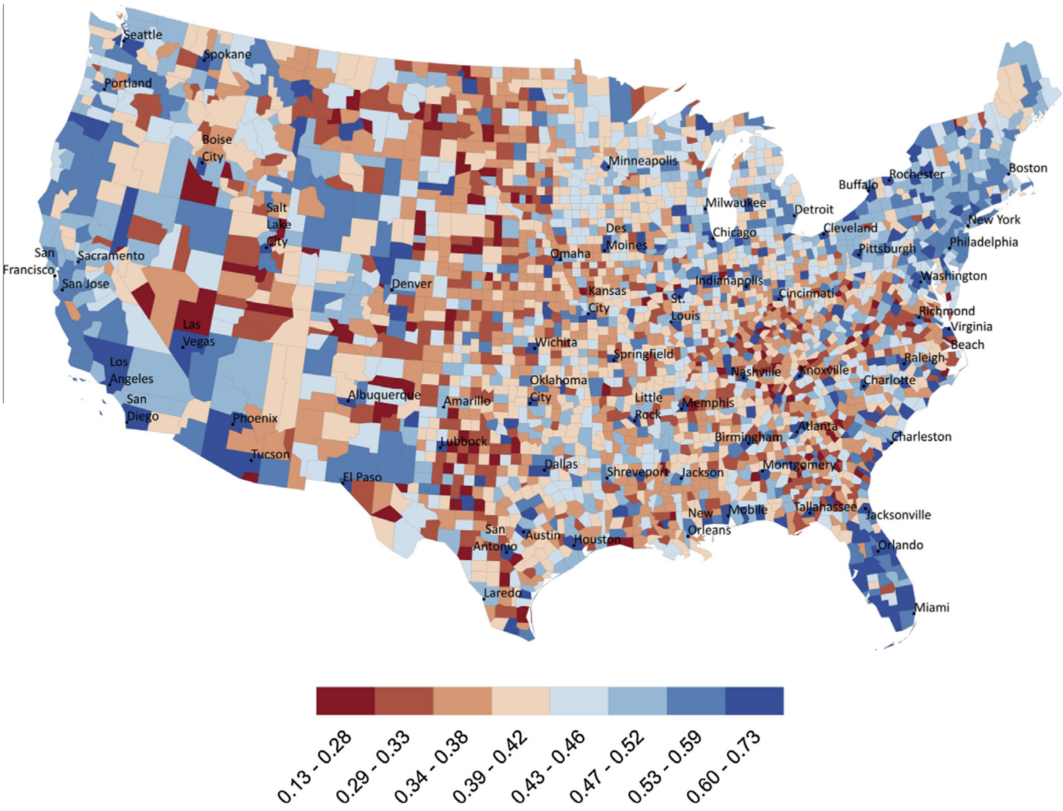

Fig. 13. Original out-flow entropy values.

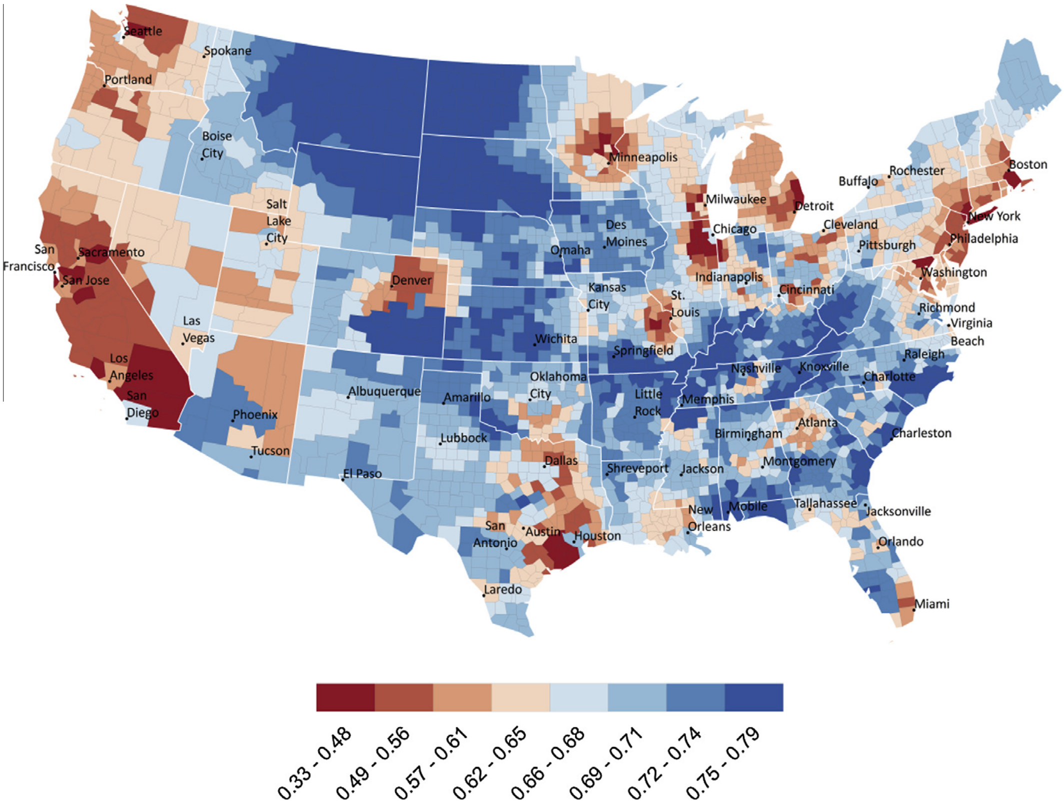

Fig. 14. Smoothed in-flow entropy.

larger places tend to have connections with many other places, and hence the entropy value is larger (i.e., more evenly spread).

The smoothed in flow entropy values shown in Fig. 14 show distinctive and different patterns from their original unsmoothed measures. It is interesting to see that the low inflow entropy clusters (red color) center around a number of major cities (such as Atlanta, Houston, San Diego and Chicago) but exclude the urban counties at each cluster center. The low entropy values indicate that these places draw focused flows, i.e., major flows are from certain places. On the other hand, clusters of high entropy values (blue color) represent places that receive migrants in similar volumes from many locations (i.e., more evenly spread). From Fig. 14, we also observe contrasting patterns within some regions. For example, while the core counties of the Chicago, Houston, San Antonio and Dallas metropolitan areas have high entropy values, the counties surrounding these metropolitan cores have low entropy values. This could potentially be explained by the tendency of metropolitan cores to attract migrants (especially young adults) in similar volumes from many places in the country as opposed to the tendency of the outskirts attracting migrants (e.g., families and retirees who prefer suburban lifestyle) from metropolitan cores disproportionately more than they attract migrants from other places.

The overall extents of the clusters in the smoothed out-flow entropy map (Fig. 15) are similar to the ones in the smoothed in-flow entropy map. However, there are local differences between the clusters of in-flow and out-flow entropy values. For example, in the Dallas, Atlanta and Chicago metropolitan areas, we observe lower in-flow entropy values for the periphery of some metropolitan areas and higher in-flow entropy values for the metropolitan cores. However, we observe the opposite of this pattern in the out-flow entropy map where the cores of Dallas, Atlanta and Chicago metropolitan areas have lower out-flow entropy as opposed to the counties surrounding them. Thus, this pattern indicates that migrants leaving these cores are more targeted (focused) towards a fewer number of places in much higher volumes. In addition to these contrasting patterns, we observe that high outflow entropy clusters match the high inflow entropy clusters, indicating that migrants leaving these places do not target certain areas and migrants moving into these places come from many locations in similar volumes.

4.4. Sensitivity analysis

In this section we evaluate the sensitivity to population thresholds and compare the results of our approach (smoothing local network and then calculating the measure) and the conventional approach (calculating measures and then smoothing measures). Due to limited space we only present the sensitivity analysis results for smoothing net migration rate. In order to select an optimal population threshold (bandwidth) that reveals general patterns in the data and reduces the instability caused by small populations (Shi et al., 2007), we experimented with a series of population thresholds using both the conventional smoothing approach and our approach. Both approaches use the same spatial kernel and the same bandwidth. We plot the variance of smoothed rates from both approaches using a series of population thresholds. Plot B in Fig. 16 shows the total variance of the resulting rates, whereas Plot A shows the difference in variances between two consecutive thresholds. As expected, Plot B shows that variance decreases as population bandwidth increases. Our approach produces rates with less variance than the conventional result since the latter still uses the original instable rates caused by small base population. Although the general trend is that variance decreases with larger thresholds, Plot A reveals several thresholds where the variance reaches a local minimum, including 300 k, 800 k, 1 million and 1.5 million. Smoothing results at these different thresholds show patterns at different scale levels, from finer to coarser resolutions.

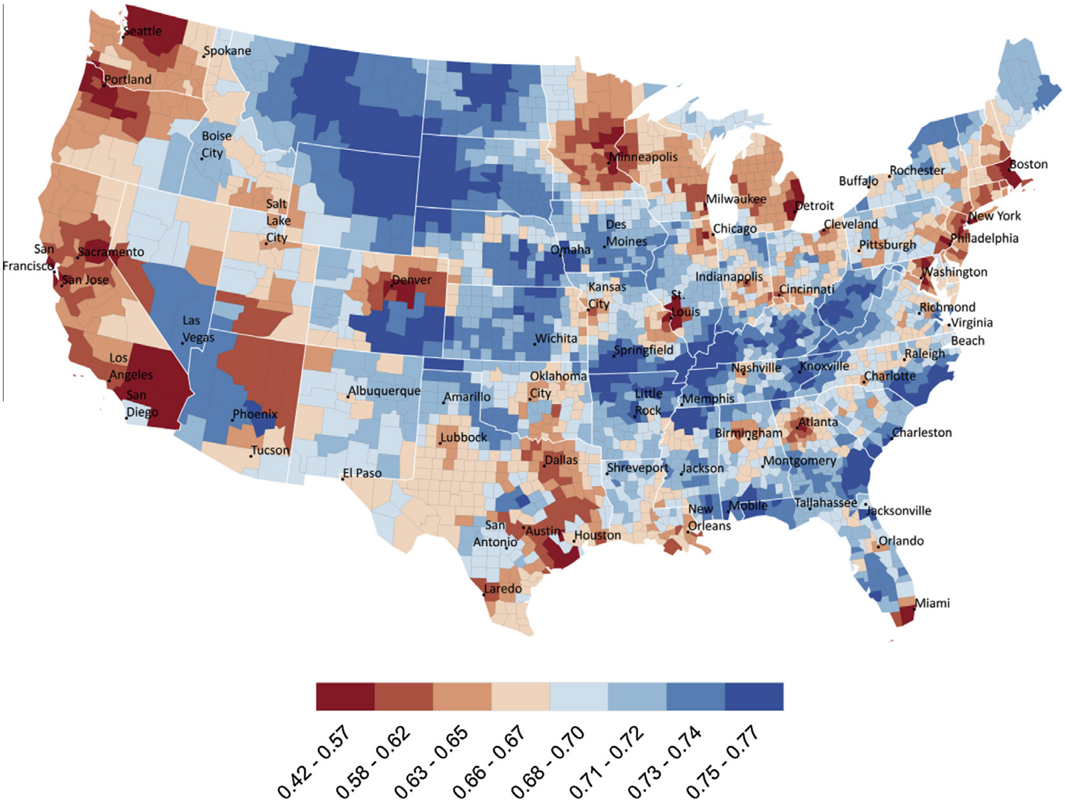

Fig. 15. Smoothed out-flow entropy.

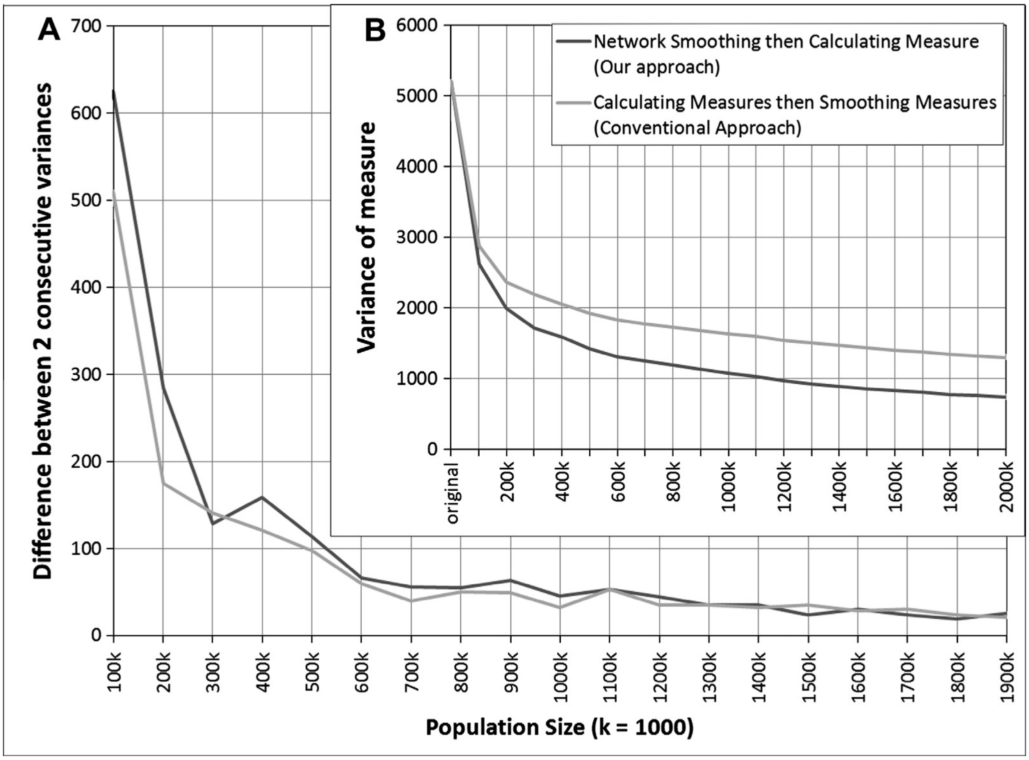

Fig. 16. The variance of smoothed net migration rates for a series of population sizes. (A) The difference in variance between two consecutive thresholds. (B) The total variance for each population size (threshold).

In this paper, we only present the results with a threshold of 1 million, which approximates the population of a medium-sized metropolitan area.

4.5. Comparison with conventional methods

In addition to analyzing the sensitivity to population thresholds, we also compare our approach to a conventional smoothing approach using net migration rate and inflow entropy measures. Fig. 17 shows the two results (conventional approach vs. our approach) for smoothed net migration rates for all population. In order to allow comparison, both methods use the same bandwidth (i.e., one million) and the same spatial kernel function (i.e., Gaussian). The overall patterns are similar in both maps. However, for the conventional approach the effect of small base populations can still be observed in many places such as the surrounding counties of Salt Lake City, UT, Albuquerque, NM and Houston, TX (Fig. 17, Left), where smoothed rates are affected by the original unstable rates (see Fig. 5) and the flows within the neighborhood. Our approach eliminates the effect of small base populations by

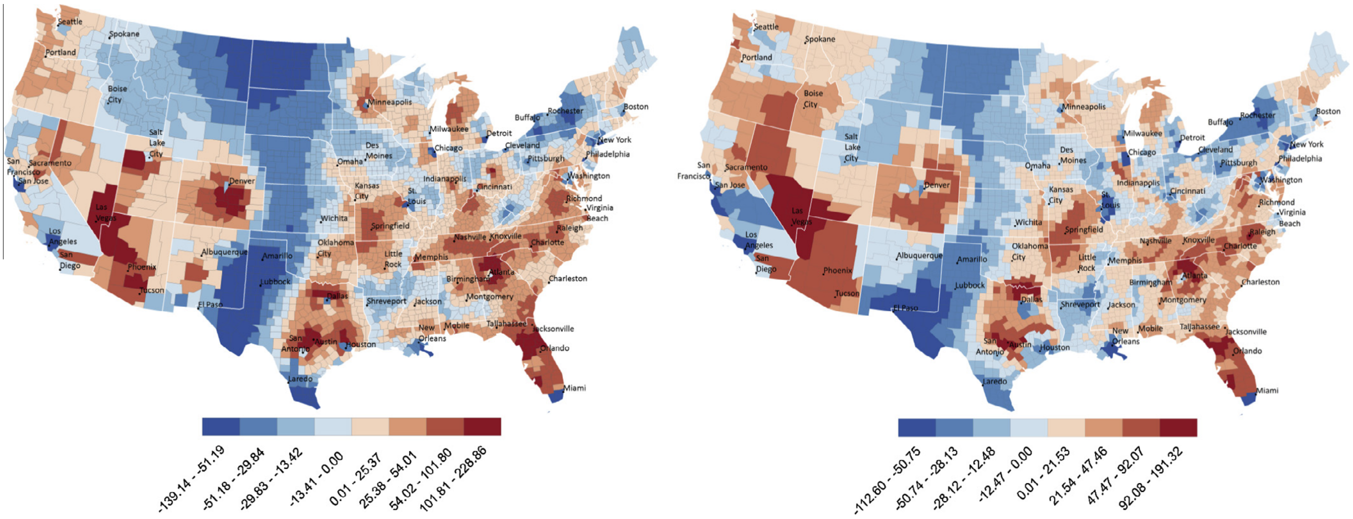

Fig. 17. Comparison of conventional smoothing result (left) and our result (right) for net migration rates. The overall patterns are similar but there are significant local differences between the two results.

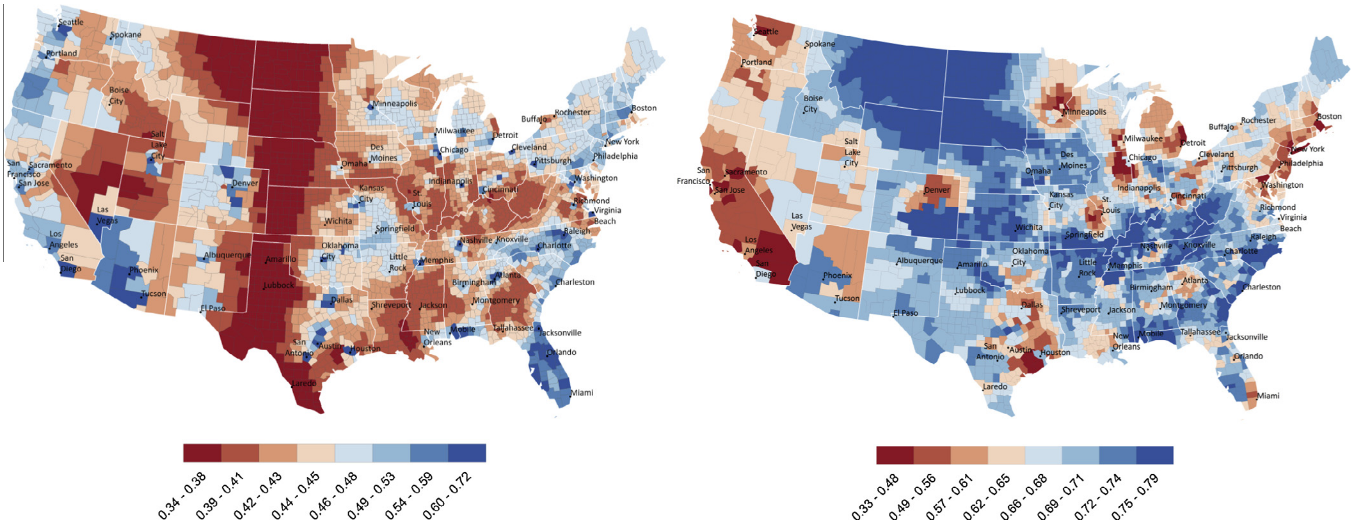

Fig. 18. Comparison of conventional smoothing result (left map) and our result (right map) for inflow entropy. The patterns are dramatically different.

treating the neighborhood as whole, removing internal flows, and calculating the measure based on smoothed network (Figs. 6 and 17 (right)).

The effect of small base populations is more dramatic for the entropy measure, causing small areas to have very small entropy values due to the sparse flows to/from those areas. This can easily be seen in the original measure result as well as the smoothing result of the conventional approach (Fig. 18, Left), which produces large clusters of low entropy values which are highly correlated with the presence of small counties and their unstable rates (see Fig. 12). Our approach, on the other hand, first smoothes the network related to a neighborhood and then calculates its entropy measure. As such, our approach reduces the effect of the small-area problem and reveals spatial clusters of low inflow entropy values, indicating places that draw focused in-migration flows (Fig. 18, Right), which is dramatically different from the result of the conventional approach. Such differences also exist for smoothing the net migration rates of different age groups as shown in Section 4.2.

5. Discussion and conclusion

Spatial interaction datasets with a relatively small number of observations for most origin-destination pairs suffer greatly from spurious data variations and as a result locational measures calculated for such datasets become unreliable. To demonstrate the usefulness of the approach, we smoothed the net migration rates for all migrants and for migrants of different age groups. We also smoothed in-flow and out-flow entropy measures (1) to show the applicability of our method to smooth network measures; and (2) to capture the variation in the magnitude of flows that each location has. The method can be used to smooth a variety of locational measures such as centrality, chi-square and flow efficiency.

It is important to note that a locational measure can only represent one aspect of the spatial and/or structural characteristics of a location in the network. More insight can be gained through analyzing the relationships between different measure results. For example, if we compare the results of smoothed net migration rate (Fig. 6) and smoothed in-flow entropy (Fig. 14), we could discover overlapping clusters such as the coincidence of high entropy clusters with clusters of high net migration rate in the east coast from Virginia to Florida in addition to the coastal areas in the east of New Orleans; and throughout most of the counties in the states of Arizona and Florida. The overlapping of these clusters could help reveal regions that attract migrants from diverse places (as opposed to places that only receive migrants from certain places). We limit our analysis scope since the goal is to present the new smoothing approach rather than carry out a comprehensive analysis of the migration dataset.

We conducted a variance-driven sensitivity analysis to evaluate a series of population thresholds to examine their effect on smoothing result. The analysis showed that smoothing results are consistent in overall patterns. A large population threshold highlights global patterns such as at the national scale while a smaller threshold can better reveal local patterns. For example, a threshold of two million population shows the Southeast as a homogeneous region of attraction, whereas a threshold of 300,000 population shows downtown Atlanta as a place of depletion and its surroundings as places of attraction.

In this paper, we employed a domain-based approach and used population or inflow/outflow to select an adaptive bandwidth. For other types of spatial interaction data where an attribute such as population doesn't exist or using a population-based threshold is inappropriate for the context of the analysis, one can employ a data-driven approach to select a bandwidth based on the properties of the network. One potential approach is to employ a graph partitioning method (Guo, 2009) to discover community structures and natural regions (groups of spatially contiguous and strongly connected units). The size of the discovered regions can provide important information for determining the size of the neighborhood (bandwidth). Our experiments with different bandwidth values showed that the smoothing results are not sensitive to small changes in bandwidth and that results with different bandwidths usually reveal patterns at different scales.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. 0748813. We greatly appreciate the constructive comments and suggestions from anonymous reviewers.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates – A square root law. *The Annals of Statistics, 10*(4), 1217–1223.

Borruso, G., & Schoier, G. (2004). Density analysis on large geographical databases. Search for an index of centrality of services at urban scale. In A. Laganá, M. Gavrilova, V. Kumar, Y. Mun, C. Tan, & O. Gervasi (Eds.). *Computational science* *and its applications – ICCSA 2004* (Vol. 3044, pp. 1009–1015). Berlin/Heidelberg: Springer.

Bors, A., & Nasios, N. (2009). Bayesian estimation of kernel bandwidth for nonparametric modelling. In C. Alippi, M. Polycarpou, C. Panayiotou, & G. Ellinas (Eds.). *Artificial neural networks – ICANN 2009* (Vol. 5769, pp. 245–254). Berlin/Heidelberg: Springer.

Carlos, H., Shi, X., Sargent, J., Tanski, S., & Berke, E. (2010). Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics, 9*(1), 1–8. http://dx.doi.org/10.1186/1476-072x-9-39.

Clark, G. L. (1982). Volatility in the geographical structure of short-run US interstate migration. *Environment and Planning A, 14*(2), 145–167.

D'Amico, M., & Ferrigno, G. (1990). Technique for the evaluation of derivatives from noisy biomechanical displacement data using a model-based bandwidth-selection procedure. *Medical and Biological Engineering and Computing, 28*(5), 407–415. http://dx.doi.org/10.1007/bf02441963.

Danese, M., Lazzari, M., & Murgante, B. (2008). Kernel density estimation methods for a geostatistical approach in seismic risk analysis: The case study of Potenza Hilltop Town (Southern Italy). In O. Gervasi, B. Murgante, A. Laganà, D. Taniar, Y. Mun, & M. Gavrilova (Eds.). *Computational science and its applications – ICCSA 2008* (Vol. 5072, pp. 415–429). Berlin/Heidelberg: Springer.

Davies, T. M., & Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine, 29*(23), 2423–2437. http://dx.doi.org/10.1002/sim.3995.

De Montis, A., Barthelemy, M., Chessa, A., & Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B: Planning and Design, 34*(5), 905–924.

Estrada, E., & Bodin, O. (2008). Using network centrality measures to manage landscape connectivity. *Ecological Applications, 18*(7), 1810–1825.

Fagiolo, G., Reyes, J., & Schiavo, S. (2009). World-trade web: Topological properties, dynamics, and evolution. *Physical Review E, 79*(3), 036115. http://dx.doi.org/10.1103/PhysRevE.79.036115.

Fischer, M. M., Essletzbichler, J., Gassler, H., & Trichtl, G. (1993). Telephone communication patterns in Austria: A comparison of the IPFP-based graph-theoretic and the intramax approaches. *Geographical Analysis, 25*(3), 224–233. http://dx.doi.org/10.1111/j.1538-4632.1993.tb00293.x.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. New York: John Wiley & Sons.

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics: IEEE, 15*, 1041–1048.

Holmes, J. (1978). Transformation of flow matrices to eliminate the effects of differing sizes of origin–destination units: A further comment. *IEEE Transactions on Systems, Man and Cybernetics, 8*(4), 325–332.

Hughes, H. L. (1993). Metropolitan structure and the suburban hierarchy. *American Sociological Review, 58*(3), 417–433.

Irwin, M. D., & Hughes, H. L. (1992). Centrality and the structure of urban interaction – Measures, concepts, and applications. *Social Forces, 71*(1), 17–51.

Kafadar, K. (1994). Choosing among two-dimensional smoothers in practice. *Computational Statistics & Data Analysis, 18*(4), 419–439.

Limtanakool, N., Schwanen, T., & Dijst, M. (2009). Developments in the Dutch urban system on the basis of flows. *Regional Studies, 43*(2), 179–196. http://dx.doi.org/10.1080/00343400701808832.

Long, L. E. (1988). *Migration and residential mobility in the United States*. New York: Russell Sage Foundation.

Pandit, K. (1994). Differentiating between subsystems and typologies in the analysis of migration regions: A US example. *The Professional Geographer, 46*(3), 331–345.

Plane, D. A., & Heins, F. (2003). Age articulation of US inter-metropolitan migration flows. *Annals of Regional Science, 37*(1), 107–130.

Plane, D. A., & Jurjevich, J. R. (2009). Ties that no longer bind? The patterns and repercussions of age-articulated migration. *The Professional Geographer, 61*(1), 4–20.

Plane, D. A., & Mulligan, G. F. (1997). Measuring spatial focusing in a migration system. *Demography, 34*(2), 251–262.

Plane, D. A., & Rogerson, P. A. (1991). Tracking the baby boom, the baby bust, and the echo generations: How age composition regulates US migration. *The Professional Geographer, 43*(4), 416–430.

Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., et al. (2009). Street centrality and densities of retail and services in Bologna, Italy. *Environment and Planning B: Planning and Design, 36*(3), 450–465.

Rogers, A. (1992). Heterogeneity, spatial population dynamics, and the migration rate. *Environment and Planning A, 24*(6), 775–791.

Rogers, A., & Raymer, J. (1998). The spatial focus of US interstate migration flows. *International Journal of Population Geography, 4*(1), 63–80.

Rogers, A., & Sweeney, S. (1998). Measuring the spatial focus of migration patterns. *The Professional Geographer, 50*(2), 232–242.

Sain, S. R., & Scott, D. W. (1996). On locally adaptive density estimation. *Journal of the American Statistical Association, 91*(436), 1525–1534.

Shi, X. (2009). A geocomputational process for characterizing the spatial pattern of lung cancer incidence in New Hampshire. *Annals of the Association of American Geographers, 99*(3), 521–533.

Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science, 24*(5), 643–660.

Shi, X., Duell, E., Demidenko, E., Onega, T., Wilson, B., & Hoftiezer, D. (2007). A polygon-based locally-weighted-average method for smoothing disease rates of small units. *Epidemiology, 18*(5), 523.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London; New York: Chapman & Hall.

Slater, P. B. (1975). Hierarchical regionalization of RSFSR administrative units using 1966–69 migration data. *Soviet Geography Review and Translation, 16*(7), 453–465.

Slater, P. B. (1976). The use of state-to-state college migration data in developing a hierarchy of higher educational regions. *Research in Higher Education, 4*(4), 305–315. http://dx.doi.org/10.1007/bf00991624.

Sohn, K., & Kim, D. (2010). Zonal centrality measures and the neighborhood effect. *Transportation Research Part A: Policy and Practice, 44*(9), 733–743. http://dx.doi.org/10.1016/j.tra.2010.07.006.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing* (1st ed.). London; New York: Chapman & Hall.

Whisler, R. L., Waldorf, B. S., Mulligan, G. F., & Plane, D. A. (2008). Quality of life and the migration of the college-educated: A life-course approach. *Growth and Change, 39*(1), 58–94. http://dx.doi.org/10.1111/j.1468-2257.2007.00405.x.

Yang, B. S., Luan, X. C., & Li, Q. Q. (2010). An adaptive method for identifying the spatial patterns in road networks. *Computers Environment and Urban Systems, 34*(1), 40–48. http://dx.doi.org/10.1016/j.compenvurbsys.2009.10.002.