Computers, Environment and Urban Systems 41 (2013) 12–25

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.