# Origin-Destination Flow Data Smoothing and Mapping

Diansheng Guo, Member, IEEE, and Xi Zhu

Abstract-This paper presents a new approach to flow mapping that extracts inherent patterns from massive geographic mobility data and constructs effective visual representations of the data for the understanding of complex flow trends. This approach involves a new method for origin-destination flow density estimation and a new method for flow map generalization, which together can remove spurious data variance, normalize flows with control population, and detect high-level patterns that are not discernable with existing approaches. The approach achieves three main objectives in addressing the challenges for analyzing and mapping massive flow data. First, it removes the effect of size differences among spatial units via kernel-based density estimation, which produces a measurement of flow volume between each pair of origin and destination. Second, it extracts major flow patterns in massive flow data through a new flow sampling method, which filters out duplicate information in the smoothed flows. Third, it enables effective flow mapping and allows intuitive perception of flow patterns among origins and destinations without bundling or altering flow paths. The approach can work with both point-based flow data (such as taxi trips with GPS locations) and area-based flow data (such as county-to-county migration). Moreover, the approach can be used to detect and compare flow patterns at different scales or in relatively sparse flow datasets, such as migration for each age group. We evaluate and demonstrate the new approach with case studies of U.S. migration data and experiments with synthetic data.

Index Terms-flow mapping, kernel smoothing, generalization, multi-resolution mapping, graph drawing, spatial data mining

# **1** INTRODUCTION

Geographic mobility data such as human daily activities, migration and vehicle movements have become increasingly available due to the wide adoption of location-aware technologies. The analysis and mapping of geographic mobility data is of great importance to advance our understanding of complex systems and their space-time dynamics in various domains such as transportation, demography, and emergency management [24, 27, 37, 38, 44]. However, it remains a challenging research problem to visualize large mobility data and understand its embedded complex patterns due to the constrained map space and massive connections.

In this paper, we focus on a specific type of geographic mobility data, the origin-destination flow data (i.e., OD data), which concerns the origin and destination of each movement but ignores the actual trajectory route. Following are two examples of such data sets:

- A taxi data set that has the origin and destination GPS points
- for millions of taxi riders (Point-based OD data);
- A U.S. migration data set that has migration flows between origin and destination counties (Area-based OD data).

Even a moderate-sized OD dataset, such as the county-to-county U.S. migration data, can easily have thousands of locations and millions of flows. Much larger datasets have also been emerging, for example, cell phone calls [8], geo-tagged social media messages [22] [12], taxi trips in metropolitan areas [14, 21], and simulation model outputs for an entire country [7, 10].

Flow map is the most common approach to present flow data, which visualize flows with straight or curved lines connecting origin and destination locations [18, 23, 26, 37, 39]. However, a flow map quickly becomes illegible as the data size increases due to the massive intersections and overlapping of flows. In a flow map, origin and destination locations have to be fixed to allow context-based interpretation. A number of new approaches have been proposed to address the cluttering problem [2, 5, 11, 41].

• Xi Zhu is with the Department of Geography, University of South Carolina. Email: zhuxiflying2012@gmail.com

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Digital Object Identifier 10.1109/TVCG.2014.2346271

Nevertheless, flow mapping remains a research challenge and several major problems remain to be addressed, including:

(1) The cluttering problem. Most existing flow mapping approaches are only effective for mapping small datasets due to the visual cluttering problem. There are a number of recent researches that aim to reduce the cluttering problem through intelligent rerouting [26], edge bundling [15], and matrices of multiple maps [42]. These approaches, however, suffer from significant information loss in the rendered map, either missing the visual connections or relying heavily on user interaction (such as selection and filtering) to interpret patterns. For example, the edge-bundling approach partially merges flow lines based on their geometric closeness, which makes it difficult to perceive the actual connection and flow volume between two locations unless the flow is selected or has a unique color.

(2) The modifiable area unit problem (MAUP). Another type of approach for flow mapping is through location aggregation, such as spatial clustering [2] and graph partitioning [11], or simply using high-level administrative units (e.g., states or provinces) to aggregate the original locations to a small set of regions, based on which flow maps are generated. These methods suffer from the modifiable area unit problem (MAUP) [25], i.e., different aggregations may present different (or even wrong) patterns. Excessive and arbitrary aggregation may also cause a severe loss in spatial resolution and missing major patterns. Moreover, such approaches do not support a smooth transition between scales, since flow maps based on different aggregations are very different and not comparable to each other.

(3) The normalization (or size-difference) problem. Existing approaches often use the default geographic units in the data (e.g., counties) to analyze the data. However, the given units or aggregations are often dramatically different in size, in terms of population or area, and therefore the flows among them are not directly comparable. For example, Los Angeles is the largest county in the U.S. with a population of over 9,000,000 while the smallest county, Loving County (Texas), has less than 100 residents. Without proper normalization, a flow map based on these units will give wrong understanding of patterns. Fig. 1 shows the original countyto-county migration data in the U.S., where the flows between large counties in metropolitan areas are inevitably larger than others. As such, this flow map (and further visual improvements based it) offers little insight on the true migration patterns. A related problem in this regard is the small-area problem, where the flow count between small areas is unstable since the involved population is very small.

<sup>•</sup> Diansheng Guo is with the Department of Geography, University of South Carolina. Email: guod@mailbox.sc.edu.



Fig. 1. Migration in the contiguous U.S (Census 2000), with more than 3000 counties and over 750,000 non-zero county-to-county flows. This map shows flows with 100 or more migrants, which represents 13% of migration in the data. Due to the dramatic size differences among counties, the main flows visible are among large metropolitan counties.

This paper presents a new approach to flow mapping that addresses the above three problems. It extracts inherent patterns from massive geographic flow data and constructs a generalized, schematic flow map that faithfully represent the major flow patterns. The approach consists of a flow-based density estimation method and a flow map generalization method to remove spurious data variance, normalize and smooth flows with controlled neighborhood size, and detect high-level patterns in the data.

The approach can work with both point-based flow data (such as taxi trips with GPS locations) and area-based flow data (such as county-to-county migration). Area-based flows will be converted to point-based flows using the centroids of areas. The approach can be used to compare flow patterns at different scales or in different data subsets, such as stratified migration for each age group. In addition to enabling effective flow mapping, the results of our approach can also be used for further visual improvements (such as intelligent flow layout), integration with other visualizations (such as matrix view or multidimensional visualization), and mobility modeling.

## 2 BACKGROUND AND RELATED WORK

Flow mapping has long been used in a wide range of applications such as human migration [36, 39], transportation [9], commodity flow [40], and commuting [4]. There are several challenges for flow mapping: the visual cluttering problem, the modifiable area unit (MAUP) problem, the normalization problem (i.e., size difference among geographic units), and the salience bias (i.e., patterns in flow maps tend to be dominated by flows over longer geographic distances). Although origin-destination flow data is a special type of graph (with locations as nodes), graph-drawing methods in nonspatial domains [3, 6, 17, 20, 29] are usually not directly applicable for flow mapping. The major difference is that nodes (locations) have to be fixed in a flow map since locations carry significant meaning and are of critical importance for interpreting patterns. With this limitation, a flow map can quickly become cluttered even for a small data set.

To address the visual cluttering problem, a number of approaches have been proposed with point clustering [1, 2], graph partitioning [11], surface generation [39], edge rerouting [26], and edge bundling [5, 16, 41]. These methods can be classified into three types: location aggregation, surface generation, and edge rerouting.

The first type aggregates locations into larger regions and then aggregates flows based on the new regions, which significantly reduces the number of flows for mapping. A review of aggregation methods for movement data can be found in [1]. However, aggregation will inevitably cause a significant loss of information, skip flow patterns at local scales, and suffer from the modifiable areal unit problem (MAUP).

The second type aims to resolve cluttering by producing a vector flow surface that only maps movements between geographically adjacent places [39]. Similarly, a recent development for mapping large mobility data [2] partitions the geographic area to small regions and only show movements between adjacent places. The limitation of these approaches is that the origin and destination information of each particular movement is lost in the map.

The third type focuses on minimizing edge crossing in flow maps through edge rerouting [26] or edge bundling [5, 16, 41], which reroute or bundle edges to improve the visual clarity of flow maps. These methods are effective in producing aesthetic representation of flow data, especially for small data sets. On the other hand, the main limitation of this type of approach is that bundled or re-routed edges make it difficult to perceive the actual connection between specific pairs of origin and destination *unless* each connection (analogous to an electronic wire) is uniquely identifiable, e.g., with a unique color or being highlighted.

There are also a variety of methods for spatial flow visualization based on *non-spatial* views, such as ordered matrices [10], or combinations of maps, matrices and other methods, such as Map<sup>2</sup> [13], interactive OD maps [42], and exploratory visualization [42, 43]. Normally, interactive visualization systems do *not* intend to summarize the entire data set in a *single* flow map. Instead, they provide a non-spatial view (such as a matrix) and rely on user interactions to select data to map, examine patterns from different perspectives, and make sense of data through an iterative process. These non-spatial approaches to a certain degree avoid the visual cluttering problem and do not have the salience bias. However, they cannot provide an overview of spatial flow patterns.

There are also methodologies for summarizing flow properties for each location using graph measures such as net migration ratio [14], centrality, and flow density (the number of flows passing a pixel) [28]. Kernel based smoothing and density estimation has long been used in analyzing geographic data [33, 34]. Recently, kernel density estimation has also been applied to spatial mobility data. For example, a flow density map is produced in [28], which is a raster data output with the cell value representing the total number of flow lines passing through the raster cell. Koylu and Guo [14] introduce a kernel-based approach to smooth locational measures of spatial mobility. Other density-based approaches for analyzing mobility data are introduced in [30-32], which are also a raster-based.

Different from the above raster-based or location-based density estimation methods in analyzing mobility data, we propose a vectorbased or flow-based density model. The key difference is that our approach does not estimate the flow density at a pixel or location. Instead, it estimates the flow for each pair of locations. In other words, the output of our approach is a set of smoothed flows (or flow densities) for pairs of origin and destination. This new approach is also different from the flow clustering method introduced in [45], which does not work with area-based flows and does not address the normalization problem.

Comparing to the dominant focus on the cluttering problem in the literature, there is relatively little attention on the *modifiable area unit problem*) and the *normalization (or size-difference) issue* in flow mapping. Existing approaches often use the default geographic units in the data (e.g., counties) or derive aggregations of varying sizes. In this research, we specifically take into account the differences in unit size (e.g., population, area, or other measures) to control the neighborhood size and normalize flows.

#### **3** OVERVIEW OF METHODOLOGY

In this section we summarize the design of our approach. Detailed introduction will be presented in Section 4.

Let  $T = \{T_f\}$  be an OD flow data set, n = |T| is the total number of flows;  $T_f = \langle X_{Of}, X_{Df} \rangle$  is a directed flow that starts at an origin location  $X_{Of}$  and ends at a destination  $X_{Df}$ . Let  $X = \{X_i\}$  be all locations involved in T, m = |X| is the total number of locations. Each location  $X_i$  has a non-negative *size* value  $s_i$ , which represents the size of the location such as population, area, or other context-dependent measures. The size field will be used to define the size of a neighborhood and subsequently to normalize the flow volume.

Location sizes are often dramatically different from each other (such as the population of US counties) or very small (such as the size of GPS points). In either case, it is neither meaningful nor reliable to directly compare flows among such locations. The new approach addresses these issues by re-estimating flows with controlled neighborhood-based smoothing, which are then used to extract patterns and render flow maps. Our approach has two steps:

(1) Kernel-based flow estimation and smoothing. First, given a neighborhood size threshold, we find the neighborhood and kernel bandwidth for each location in X. The weighted sizes for all locations inside a neighborhood should be equal to the neighborhood size threshold. For example, for the county-to-county migration data, we may use one million population as the size threshold for each county. Then each county will find a set of neighboring counties, whose weighted total of population will be exactly one million based on a chosen kernel model. Second, find the neighbors for each flow based on its origin neighborhood and destination neighborhood; re-estimate the flow value with its neighboring flows and its kernel model, which is constructed based on its origin and destination models.

(2) Flow selection and generalization. The above smoothing step produces a robust estimation of flow density for pairs of locations. However, it does not solve the cluttering problem and even creates a new problem as the smoothed flows have a significant amount of duplicate information or correlation. This is because neighborhoods overlap and each original flow may be used multiple times in related neighborhoods (although with different weights). We develop a flow selection and generalization method that can select representative flows from the smoothing result to remove duplicate information, enable effective flow mapping of large data, and discover generalized major patterns in the data.

#### 4 FLOW SMOOTHING WITH KERNEL MODELS

#### 4.1 Flow Neighborhood Definition

In traditional pixel-based density estimation, a neighborhood is defined for each pixel (or location), whose density value is estimated based on the data within its neighborhood. In our research, we will re-estimate the value for each flow  $T_f$  and thus need to define a neighborhood for  $T_f$ , which involves two locations: its origin  $(X_{Of})$  and destination  $(X_{Df})$ . Note that our approach is not to estimate the



Fig. 2: An illustration of the neighborhood of a flow AB. The two circular areas with gradient colors represent the neighborhoods of origin A and destination B (in red). Flows with origins in the neighborhood of A and destinations in the neighborhood of B will be considered neighbors of flow AB.

density of flow lines that pass through a location or pixel. Instead, we estimate the flow value for each pair of locations (i.e., an origindestination pair).

There are two types of neighborhood definition in kernel density estimation, i.e., fixed bandwidth and adaptive bandwidth. A fixed bandwidth is defined with a fixed geographic distance for all locations. This type of kernel is useful for estimating density in relation to spatial area. The other type is the adaptive bandwidth, which defines the bandwidth based on an attribute threshold (such as population or the number of units). Our approach can use either type of neighborhood. To define the size for the adaptive neighborhood, we use the size attribute  $S_i$  of each location, which can be the area, population, or any other measurement the location. By defining a neighborhood size threshold p based on the chosen size measurement  $S_i$ , we can accommodate different types of neighborhood.

Given a positive neighborhood size threshold p, we follow two steps to find the neighborhood of a flow  $T_f$ . First, construct a *p*-size neighborhood for each location  $X_i \in X$ , which is the smallest *k*nearest-neighbor neighborhood of  $X_i$  (including itself) that meets the size constraint p (see Definitions 1 and 2). The bandwidth  $\sigma_{X_i}$  of the neighborhood of  $X_i$  is the radius of the smallest circle centered on  $X_i$ that covers all points in its neighborhood (see Definition 3). Second, construct a *p*-size neighborhood for flow  $T_f$ , which is a set of flows  $\{T_q \in T\}$ , where the origin and destination of  $T_q$  are inside the *p*-size neighborhoods of the origin and destination of  $T_f$ , respectively (see Definition 4). Figure 2 illustrates the neighborhood of a flow. Note that the *p*-size neighborhood of flow  $T_f$  includes  $T_f$ .

**Definition** 1: The *k*-Nearest-Neighbor (KNN) Neighborhood of a location  $X_i$ ,  $KNN(X_i, k) = \{X_q \in X\}$ , is the nearest k locations (including  $X_i$ ) to  $X_i$  in X.

**Definition** 2: The *p*-Size Neighborhood of a location  $X_i$ , PSN( $X_i$ , p) and p > 0, is defined as the smallest  $KNN(X_i, k) = \{X_q \in X\}$  that has a total size  $\sum s_q \ge p$ . To make sure that the total weighted size is exactly p, the  $k_{ih}$  neighbor (i.e., the furthest point) will be assigned a weight  $\le 1$  so that  $\sum w_q s_q = p$ , where  $w_q = 1$  except for the  $k_{ih}$  neighbor;

**Definition** 3: The *bandwidth of the p-Size Neighborhood of* a location  $X_i$ , denoted as  $\sigma_{X_i}$ , is the Euclidean distance of the furthest point in  $PSN(X_i, p)$  to  $X_i$ . In other words, bandwidth  $\sigma_{X_i}$  is the radius of the smallest circle centered on  $X_i$  that covers all points in the neighborhood.

**Definition** 4: The *p*-Size Neighborhood of a flow  $T_f$ ,  $PSN(T_f, p) = \{T_q \in T \mid X_{Oq} \in PSN(X_{Of}, p) \text{ and } X_{Dq} \in PSN(X_{Df}, p)\}$ , where  $X_{Oq}, X_{Dq}$  are the origins and destination locations of flow  $T_q$ , and  $X_{Of}, X_{Df}$  are the origin and destination of flow  $T_f$ .  $T_f \in PSN(T_f, p)$ .

## 4.2 Flow Kernel Model

Given a flow  $T_f$ , its origin neighborhood  $PSN(X_{Of}, p)$ , destination neighborhood  $PSN(X_{Df}, p)$ , and flow neighbors  $PSN(T_f, p) = \{T_q\}$ , we define a kernel model to calculate the weight for each flow  $T_q$  in relation to  $T_f$ . Note that a flow may belong to multiple neighborhoods and its weight can be different in different neighborhoods. Commonly used models for kernel-based smoothing include the Gaussian model, Epanechnikov model and the triangular model. According to [35] and our experiments, the choice among these models do not have a significant impact on the result. In this research we extend the Gaussian kernel model for weighting and smoothing flows.

We re-estimate the flow value for  $T_f$  with three considerations: (1) the new flow value should be a stable measurement with a sufficient base population (i.e., each neighborhood should be sufficiently large to avoid the small-area problem); (2) the estimated flow value should be a normalized measurement that removes the effect of size difference among units; and (3) the estimation should give nearby flows more weights than distant ones within the flow neighborhood. The first consideration concerns the choice of the p value, which will be discussed separately in Section 4.5. The second consideration requires that each neighborhood be of the exact size p. The third consideration can be satisfied with the Gaussian model, which we explain below. The model construction involves two steps: location-based model and flow-based model.

For a given location  $X_0$ , and its neighborhood  $\{X_i\}$ , a Gaussian model is defined as follows (Eq. 1):

$$G(X_q, X_0; \sigma_{X_0}) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{d(X_q, X_0)^2}{2\sigma_{X_0}^2}}, & \text{if } d(X_q, X_0) \le \sigma_{X_0} \\ 0, & \text{if } d(X_q, X_0) > \sigma_{X_0} \end{cases}$$
(1)

Where:

 $G(X_q, X_0; \sigma_{x_0})$  is the Gaussian weight of  $X_q$  in relation to  $X_0$ ;  $X_0$  is the center location of a neighborhood;

 $\sigma_{X_0} > 0$  is the neighborhood bandwidth of location  $X_0$ ;

 $X_q$  is a location in the neighborhood;

 $d(X_q, X_0)$  is the Euclidean distance between  $X_q$  and  $X_0$ .

In Definition 2 we explained that each of the k locations  $\{X_q\}$  in a neighborhood has a size  $s_i$ , a weight  $w_q=1$  (except for the last neighbor, which has a weight  $w_k \leq 1$ ), and the total size of the neighborhood  $\sum_{1}^{k} w_q s_q$  is exactly p. These weights will be adjusted here based on the Gaussian model (Eq. 1). Let  $\theta$  be the total *Gaussian weighted* size of the neighborhood before adjustment (Eq. 2). The adjusted weight for location  $X_q$  in relation to  $X_0$ ,  $G(X_q, X_0)$ , is defined in Eq. 3, which ensures that  $\sum_{1}^{k} G(X_q, X_0) * s_q = p$ .

$$\theta = \sum_{q=1}^{n} G(X_q, X_o; \sigma_{x_0}) w_q s_q \tag{2}$$

$$G(X_q, X_0) = G(X_q, X_0; \sigma_{X_0}) w_q p / \theta$$
(3)

After the location weight  $G(X_q, X_0)$  is configured for each point in each neighborhood, we can construct a flow-based model to calculate the weight for each flow  $T_q$  in the neighborhood of flow  $T_0$ . Let  $X_{00}$  and  $X_{D0}$  be the origin and destination of  $T_0$ ;  $X_{0q}$  and  $X_{Dq}$  be the origin and destination of  $T_q$ . The weight for  $T_q$  in relation to  $T_0$  is defined in Eq. 4. Essentially, the weight is a joint probability of the two kernels, one for the origin  $X_{0q}$  and one for the destination  $X_{Dq}$ .

$$G(T_q, T_0) = G(X_{0q}, X_{00})G(X_{Dq}, X_{D0})$$
(4)

#### 4.3 Flow Density Calculation

Based on the flow model in Eq. 4 and its derived weights for flows in each neighborhood, each original flow can be smoothed or reestimated. Let  $T = \{T_f\}$  be the original flow data;  $T_f$  represents a directed flow from location  $X_{Of}$  to location  $X_{Df}$ ;  $\{T_q\}$  be the *p*-size neighborhood of  $T_f$ ,  $|\{T_q\}| = k$ ,  $T_f \in \{T_q\}$ . The smoothed flow volume for  $T_f$  is:

$$T_f' = \sum_{q=1}^{k} [G(T_q, T_f)T_q]$$
<sup>(5)</sup>

As explained in the previous section, the neighborhood size p determines the origin bandwidth  $\sigma_{X_{Of}}$  and the destination bandwidth  $\sigma_{X_{Df}}$ , which can be considered as a scale factor. The larger the two bandwidths are, the more global and general patterns we are looking for. Therefore, we only smooth flows of a longer geographic distance than  $(\sigma_{X_{Of}} + \sigma_{X_{Df}})$ . In other words, we only calculate smoothed flows between non-overlapping neighborhoods. We may also impose an additional parameter, *minDist*, to skip flows that are

shorter than *minDist*, which are mainly local flows and thus may not be necessary to include in a flow map at a large scale. For example, for the U.S. migration maps in Fig. 3 and Fig. 4, we set *minDist* = 200km, i.e., flows shorter than 20km are relatively local and thus skipped on a national map.

The smoothed flow  $T'_f$ , which is the weighted total of flows in its neighborhood, can be interpreted as the normalized flow value from  $X_{Of}$  to  $X_{Df}$ , when each of them were exactly of the same size p. For example, if p is one million of population, a smoothed flow of 10,000 migrants between two neighborhoods means that the flow density is 10,000 flows per one million of population on each side. With the smoothed flow values, instead of the original flow values, we can reliably compare and understand the magnitude of flows among locations, without being affected by the differences in size or unstable measures for small areas.

#### 4.4 Flow Selection and Mapping

The above smoothing step improves and normalizes the original flows so that the values are robust (avoiding the small-area problem), comparable, and preserves spatial resolution (with each flow retaining its original origin and destination). However, it does not solve the cluttering problem since the total number of flows remains the same. Moreover, it adds a significant amount of redundant information and correlation between neighboring flows as their neighborhoods overlap and share neighbor flows. We developed a flow selection method that can address both problems simultaneously and render a generalized flow map that is visually clear, rich in information, and accurately reveals major flow patterns within the data.

The essential idea is to find a subset of the smoothed flows that can represent the major flow patterns in the data and do not have duplicate information. This idea is in line with the map generalization concept in cartography, where only the most salient information is selected and represented on a map in a way that suits the scale of the display.

Our flow selection algorithm is presented in Algorithm 1. It takes the smoothed flows and the neighborhood of each flow as inputs. All flows are sorted to a descending order according to smoothed flow values. Following the order of flows (i.e., starting from the largest flow), the process selects one flow at a time. There are two primary selection criteria: (1) selected flows do not share flow neighbors with each other, which is to avoid duplicate information; and (2) selected flows are not too close to each other, which is to achieve a more balanced spatial representation and to avoid the cluttering problem. Note that two selected flows may have overlapping origins or overlapping destinations but not both. The selection process stops when no more flow can be selected or when a specified number (l) of flows have been selected (which returns the top l selected flows).

Algorithms 1. Flow Selection		
Input:	Smoothed flows $\{T'_f\}$ ;	
	Distance threshold <i>minSpace</i> ;	
	Desired number of flows $l$ ;	
Output	: A set of selected flows $\{T_s\},  \{T_s\}  \le l;$	
Steps:		
$\{T_s\}$	$\} = \emptyset;$	
Sor	t $\{T'_f\}$ by smoothed flow value, in descending order;	
For	each flow $T'_f = \langle X_{Of}, X_{Df} \rangle$ :	
	<i>select</i> = true;	
	FOR each flow $T_q = \langle X_{Oq}, X_{Dq} \rangle \in \{T_s\}$ :	
	IF ((EuclideanDist( $X_{Of}, X_{Oq}$ )<( $\sigma_{X_{Of}} + \sigma_{X_{Oq}}$ ) OR	
	(EuclideanDist( $X_{Df}, X_{Dq}$ ) <( $\sigma_{X_{Df}} + \sigma_{X_{Dq}}$ ) OR	
	(EuclideanDist( $X_{Of}, X_{Oq}$ ) <minspace) or<="" td=""></minspace)>	
	(EuclideanDist( $X_{Df}, X_{Da}$ ) <minspace))< td=""></minspace))<>	

<i>select</i> = false;	
END FOR	
IF (select AND $ \{T_s\}  < l$ )	
Add $T'_f$ to $\{T_s\}$ ;	
END FOR	

Comparing with existing flow mapping approaches, our approach has four advantages with the combination of flow smoothing and flow selection. First, it preserves spatial resolution and location information by searching for the pairs of origins and destinations that have the highest flow density. The selected flows maintain their original origin and destination locations. Second, it avoids the modifiable area unit problem (MAUP) by not imposing an arbitrary aggregation at the beginning. Instead, it evaluates each neighborhood and finds the "best" aggregation (if smoothing is considered a form of aggregation). Third, it normalizes flow values by origin and destination sizes so that flows are comparable and thus enables correct interpretation of flow patterns. Fourth, the resulted flow map is visually clear and yet reveals rich information. The approach can naturally support multi-resolution flow mapping and stratified flow mapping, which we will explain below.

## 4.5 Multi-Resolution Flow Mapping

The neighborhood size p parameter, which controls the *bandwidth* of the kernel, is a key parameter in the flow-smoothing step. Intuitively, one wants to set the size as small as possible so that the result patterns have a high spatial resolution. On the other hand, however, the smaller the neighborhood size is, the less robust the estimated flow value is due to the small-area problem. Therefore, it is a trade-off between resolution and robustness in determining the value of p. Another consideration in setting p is related to scale. Different p values can allow us to search for flow patterns at different scales. For example, using a one million population threshold may be better for examining migration patterns at the national level (filtering out local details) while using a half-million or even smaller threshold one can see local patterns within a selected region.

A user may start from the national level with a large neighborhood size, which reveals flow trends across the U.S. Then the user can zoom in on a local region. There can be two ways to allow a smooth transition from one scale to another. The first option is to keep the neighborhood size unchanged and show more (small) flows for local regions, where the number of flows to be mapped will be determined by visual clarity. For example, at the national level, the flow map may show the top 200 selected flows nationwide. When zoomed to Texas, the map will show the top 200 flows in the local area, meaning that flows ignored on the national map will emerge at the local level. The second option, combined with the first, is to produce a sequence of smoothing results based on different neighborhood sizes. When the map scale changes, the map will show the smoothing flows at an appropriate level. In Section 5.1.2 we will present a preliminary example of this capability.

## 4.6 Flow Mapping of Stratified Data

Another important advantage of the approach is its ability to work with sparse flow data. It is possible to stratify a data set based on a certain attribute, make a flow map for each subset of data, and compare their flow patterns. For example, it is well understood that different age groups have very distinctive migration choices and it is very important to be able to examine such differences. Therefore, instead of mapping all migration flows in one map, we can divide the migration data into subsets based on age groups and map the migration patterns for each age group using the same neighborhood size p. This is possible because our approach can reliably extract patterns for sparse flow data with smoothing and normalization, and has the ability to maintain spatial accuracy while searching for flow patterns. In Section 5.1.3 we will present results that show this capability.

# 4.7 Computational Efficiency

With spatial index, the search for the nearest k neighbors for m locations takes  $O(mk \log k)$  time. The smoothing of n flows, each having at most  $k^2$  neighbor flows, takes  $O(nk^2)$  time. The time complexity for the flow selection step is also O(nl). Since k << m < n and l is a constant value, the overall time complexity of the approach therefore is  $O(nk^2)$ , which is scalable to process large data sets when k is relatively small.

For example, the U.S. county-to-county migration data has 3075 locations and over 720,000 flows, for which k is around 15 when p is set as one million of population. On a desktop computer with a 3.2 GHz CPU, it takes less than 30 seconds to complete all steps. Therefore, it is scalable to support real-time user interaction.

For large point-based flow data, such as taxi trips with GPS points as locations, the needed k value (controlled by neighborhood size p) may be relatively large since each GPS point is unique in the data. In this case, we can perform a preliminary clustering of GPS points and select representative points (clusters) to aggregate flows first, which can dramatically reduce the number locations without significantly impacting spatial accuracy in pattern detection.

## 5 EVALUATION WITH CASE STUDIES

## 5.1 Migration Data Mapping

In this case study we analyze and map the U.S. internal migration data set from the 2000 Census, covering the time period of 1995—2000. Census 2000 asked where the person lived five years ago (i.e., April 1, 1995) and therefore the data includes migrants who moved within the five years (1995-2000). In this research, we focus on the migration within the continental U.S. including 48 states and Washington D.C., which has 3075 counties and 721,433 unique pairs of counties with a nonzero migration flow. To apply our approach, each county area is converted to point (i.e., the area centroid), which retains all the attributes of the county such as population.

## 5.1.1 Net Migration Flow Map

The neighborhood size p is set with a population of 1,000,000 (based on the U.S. 2000 census population data at the county level), which is about the size of a metropolitan area in the U.S. According to the size constraint, each county finds a neighborhood of one or more nearby counties. The background map in Fig. 3 shows the bandwidths of counties, which range from zero (where a single county has more than one million population) to about 700km, with a mean of 187km. In this case study, we particularly focus on the *net migration*, which is the inflow minus the outflow. Net migration represents the net gain or loss of migrants for each pair of counties and is often used to examine migration patterns. For each pair of counties with a non-zero migration, we re-estimate its flow value in each direction with our approach (with the parameter *minDist* = 200km) and then calculate a new net migration value with the smoothed inflow and outflow values.

From the smoothed net migration flows, we use the flow selection algorithm presented in Section 4.4 to select the top 200 net migration flows, with the parameter *minSpace* = 300km. The *minSpace* parameter is mainly to reduce the number of similar flows within local areas of high population density. The selected flows are shown in Fig. 3. The map reveals a variety of local and global migration patterns that are not visible in the original flow map (Fig. 1) or other flow maps produced with existing approaches for the same data set (e.g., [15, 16]). To verify the patterns in Fig. 3, we overlay the flow map on top of a map of smoothed net migration ratio in Fig. 4. See [19] for the method that calculates the smoothed net migration ratios. The flow patterns discovered with the new approach clearly confirm and explain the overall net migration patterns. We can see most flows originated from bluish areas and move to reddish areas (hot destinations).



Fig. 3: Top 200 smoothed net migration flows, with the size p = 1,000,000 population. Both the line width and color represent flow values. The background map shows bandwidth of each county.



Fig. 4: Top 200 smoothed net migration flows are overlayed on top of a map of smoothed net migration rate for each county. Blue colors represent places that had more out-migration than in-migration while red areas were hot destinations. The smoothed flows clearly explain the patterns in the net migration rate map. We can see where the out-migration went and where in-migration came from for the red areas.



Fig. 5: A zoom-in migration map for the mid-west region, switched to a neighbourhood size of 500,000 (half of the size as used in Fig. 4). Flow patterns are consistent with those in Fig. 4 but more local flows are present in this map. See Fig. 4 for the legend of the background.

By examining the flow map in Fig. 4, we can obtain a rich set of new information on the national trends and local patterns of migration in the U.S. For example, the net migration rate map in Fig. 4 shows that the area around El Paso (Texas) has a strong negative net migration rate, indicating a significant loss of population due to out-migration to other areas. The flow map confirms this pattern and shows that its outgoing net-migration went almost in all directions to nearby areas. A number of metropolitan areas lose population due to negative net migration, such as Houston, Chicago, Boston, Detroit, Los Angeles, San Francisco, and Seattle. From the flow map, we can understand where the out-migration went. For example, it is surprising to see a large migration flow from Chicago to the Philadelphia area. Among others, one interesting trend shown in the map is that the net migration from the big blue band (stretching from El Paso, to Lubbock and to the west of Minneapolis) mostly moved to the east.

## 5.1.2 Multi-Resolution Migration Mapping

Our approach naturally supports a smooth transition between scales and enables multi-resolution flow mapping. For example, one may zoom in on the flow map in Fig. 4 to focus on a local area such the Mid-West area around the lakes (Fig. 5), which shows the selected flows for the local area based on a smaller neighborhood size (500,000 population). While major flows are still consistent with the national map in Fig. 4, more local flows emerge.

The desirable feature of our approach in supporting multiresolution flow mapping is that the smoothing process guarantees that the flows with high densities will be detected and the selection process ensures that large flows are represented in the map, at different scales. The smoothing and selection process is efficient and can process large datasets to create flows at multiple scales either beforehand or in real-time. To ensure the computational efficiency for large point-based flows, we briefly introduce several strategies in Section 5.3.

#### 5.1.3 Stratified Migration Maps

We can also map the migration flows of different age groups separately and compare their flow patterns. We partition the original migration data into subsets, each containing the moves of migrants within a certain age group. A flow map is generated for each subset (i.e., age group). All maps use the same configuration, namely neighborhood size p = 1,000,000 population, *minDist* = 200km, and *minSpace* = 300km. Existing flow-mapping approaches do not work well in this case since the flow matrix is sparse and most flows are small for each data subset. Without smoothing, it is difficult to obtain a robust representation of patterns.

Fig. 6 and Fig. 7 show the smoothed flow maps for age group 65-69 and 25-29, respectively. It is obvious that young migrants (e.g., college graduates) and senior migrants (e.g., retirees) have very different preferences for destination. The migration flow map in Fig. 6 shows that senior migrants in the east had a strong preference for Florida (particularly West and South Florida) while those in the west were more likely to move to Arizona or its surrounding area. For the metropolitan areas in the north (such as Minneapolis, Detroit, and Boston) there is also a senior migration trend towards the rural areas further north. On the contrary, young migrants have a strong tendency to converge on large cities with relatively short-distance moves. Atlanta, Charlotte, New York City, Dallas, Houston, Denver, San Francisco, and Portland attracted a large number of inflows of young population from surrounding regions. It also appears that San Francisco is particularly attractive to the young people in the Northeast region, who migrated over a long geographic distance to the Bay Area.

With these two maps as examples, we demonstrate the flexibility and reliable power of our approach in analyzing origin-destination flow data, compare patterns in stratified data sets, which are often sparse and therefore demand smoothing to accentuate patterns.



Fig. 6: Smoothed net migration flows for age 65-69, with population threshold = 1,000,000.



Fig. 7: Smoothed net migration flows for age 25-29, with population threshold = 1,000,000. The background map shows the net migration rate for age group 25-29.



Fig. 8: A synthetic flow data set with flow clusters and random flows. From top-left to bottom-right, the six maps are (1) 7000 points; (2) 3500 flows, with 350 clustered flows and 3150 random flows; (3) top smoothed flows with p = 200; (4) top smoothed flows with p = 300; (5) smoothed flows with p = 500; and (3) smoothed flows with p = 700.

## 5.2 Evaluation with Synthetic Data

To further evaluate our approach and assess its capability (and limitations) in detecting flow patterns, we construct a synthetic data set that has both random flows and clustered flows. We generate 7000 points in a rectangular area, with two dense 'urban' areas (see the top-left map in Fig 8). Then seven flow clusters are generated and each cluster has 50 flows. To generate a flow cluster, we choose a circular area of X points as the origin and a circular area of Y points as the destination, randomly pick a point from X and a point from Y to form a flow, repeat 50 times. The clusters are as follows:

- Blue cluster, |X| = 100 and  $|Y| = 100, X \cap Y = \emptyset$
- Green cluster, |X| = 100 and  $|Y| = 100, X \cap Y = \emptyset$
- Pink cluster, |X| = 200 and |Y| = 200,  $X \cap Y = \emptyset$
- Yellow cluster, |X| = 350 and |Y| = 350,  $X \cap Y = \emptyset$
- Magenta cluster, |X| = 500 and  $|Y| = 500, X \cap Y = \emptyset$
- Cyan cluster, |X| = 500 and  $|Y| = 500, X \cap Y = \emptyset$
- Red "cluster", |X| = 500 and |Y| = 500, X = Y.

The remaining 6300 points are randomly paired to create a set of 3150 random flows. Therefore, only 11% of the 3500 flows are clustered. Note that the red "cluster" is not considered a cluster by our approach since the flows are random within a confined area (not between two separate areas). The blue and green clusters are the strongest while the magenta and cyan clusters are the weakest. They are designed to have different "ideal" scales (i.e., neighborhood sizes), ranging from 100 to 500 (in the order of the list above). See Fig. 8 (top-center map) for the clusters. The blue cluster is on top of the yellow cluster and the green cluster is on top of the pink cluster.

The neighborhood size p in this application is defined as the number of points in a neighborhood. We run our method with different p values (200, 300, 500, and 700) and their results are shown in Fig. 8, from which we can observe the following. First, with a smaller p (e.g., 200), the strongest clusters (blue and green) are detected first and then part of the magenta and pink clusters also emerge in the top flows. However, the cyan cluster is missing. By increasing the p value, the weaker clusters are all detected (including the cyan and magenta cluster) while the overlapping clusters are merged into larger clusters (e.g., blue merged with yellow, and green merged with pink). By comparing the four smoothed maps, we can see that the results smoothly transition between scales (i.e., p values), which is a desirable feature. Other than a few weak patterns out of the random flows, the top smoothed flows contain primary the true patterns with high spatial accuracy.

# 5.3 Limitations

Despite the advantages and desirable features of our approach introduced above, it also has several limitations. First, the neighborhood size p can only be configured with empirical knowledge. The approach cannot automatically find an "optimal" p value. Second, as shown in the experiments, the method cannot simultaneously discover flow clusters of different density or from different scales. One has to change the parameter to extract patterns at different scales. A small p value cannot detect patterns at higher levels while a larger p value may still capture patterns at lower levels, as the experiments show. Third, the method does not test the statistical significance of the discovered clusters (or top smoothed flows). Some of the top flows may be from a random pattern and therefore further research is needed to test the significance of flow patterns. Fourth, the generalized flow map represents the group of strong flow clusters, which may miss location patterns formed by many smaller but weaker flows.

In this paper, we primarily focus on areabased flows (although they are converted to point-

based flows using their area centroids). For very large and originally pointed-based flow data, such as millions of taxi trips with GPS locations, a few strategies (such as clustering and sampling) are needed to achieve computational scalability for real-time processing and user interactions. This is beyond the scope of this paper due to space limitation. In general, our approach is applicable for both point-based flow data (such as the synthetic data) and area-based flow data (such as the migration data).

# 6 CONCLUSION AND FUTURE DIRECTION

This paper presents a new approach for the computational analysis and flow mapping of large spatial mobility data. It extracts inherent patterns from massive flows and constructs visually legible flow maps that faithfully represent the major flow patterns. The approach consists a flow-based density estimation method and a flow selection method to normalize and smooth flows with controlled neighborhood size and detect high-level patterns in the data. The approach has three distinctive features. (1) It removes the effect of size differences among spatial units and produces normalized flow estimation. (2) It extracts major flow patterns in the smoothed flows to filter out duplicate information. (3) It enables effective flow mapping and supports multi-resolution flow mapping and stratified flow mapping.

The approach can work with both point-based flow data (such as taxi trips with GPS locations) and area-based flow data (which will be converted point-based flows with area centroids). In addition to enabling effective flow mapping, the resulted flows of our approach can also be used for further visual improvements (such as intelligent flow layout), integration with other visualizations, and spatial modeling. The neighborhood size p is an important parameter, as it determines the smoothness of the result. Methods for automatically selection neighborhood sized have been proposed in the literature[35]. For most application scenarios, the neighborhood size has context-related meaning (such as population) and often can be configured with domain knowledge.

The approach is efficient and fairly robust with the combined steps of flow smoothing and flow selection, which have been demonstrated with various applications in Section 5. Particularly, its ability to map data at different scales with smooth and consistent transition is an important feature for the mapping of large spatial flow data. Further research is needed to fully design and support possible user interactions in using the methodology, including how to automatically choose an optimal detail level or neighborhood size given a scale or chosen region; and allow users to understand the original flows represented by each selected flow.

# ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under Grant No. 0748813.

#### **R**EFERENCES:

- G. Andrienko, and N. Andrienko. "A General Framework for Using Aggregation in Visual Exploration of Movement Data", *The Cartographic Journal*, 2002, 47, pp. 22–40
- [2] N. Andrienko, and G. Andrienko. "Spatial Generalisation and Aggregation of Massive Movement Data", *IEEE Transactions on Visualization and Computer Graphics*, 2011, 17 (2), pp. 205 - 219
- [3] G.D. Battista, P. Eades, R. Tamassia, and I.G. Tollis. Graph Drawing: Algorithms for the Visualization of Graphs (Prentice-Hall, 1999. 1999)
- [4] J. Corcoran, P. Chhetri, and R. Stimson. "Using Circular Statistics to Explore the Geography of the Journey to Work", *Papers in Regional Science*, 2009, 88 (119–32)
- [5] W. Cui, H. Zhou, H. Qu, P.C. Wong, and X. Li. "Geometry-Based Edge Clustering for Graph Visualization", *Ieee Transactions on Visualization* and Computer Graphics, 2008, 14 (6)
- [6] E. Di Giacomo, W. Didimo, L. Grilli, and G. Liotta. "Graph Visualization Techniques for Web Clustering Engines", *IEEE Trans Vis Comput Graph*, 2007, 13 (2), pp. 294-304
- [7] S. Eubank, H. Guclu, V.A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. "Modeling Disease Outbreaks in Realistic Urban Social Networks", *Nature*, 2004, 429, pp. pp. 180-184
- [8] S. Gao, Y. Liu, Y. Wang, and X. Ma. "Discovering Spatial Interaction Communities from Mobile Phone Data", *Transactions in GIS*, 2013, 17 (3), pp. 463-481
- [9] P.R. Gould. "On the Geographical Interpretation of Eigenvalues", Transactions of the Institute of British Geographers, 1967, pp. 53-86
- [10] D. Guo. "Visual Analytics of Spatial Interaction Patterns for Pandemic Decision Support", *International Journal of Geographical Information Science*, 2007, 21 (8), pp. 859-877
- [11] D. Guo. "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
- [12] D. Guo, and C. Chen. "Detecting Non-Personal and Spam Users on Geo-Tagged Twitter Network", *Transactions in GIS*, 2014, 18 (3), pp. 370-384
- [13] D. Guo, J. Chen, A.M. MacEachren, and K. Liao. "A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)", *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12 (6), pp. 1461-1474
- [14] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris. "Discovering Spatial Patterns in Origin-Destination Mobility Data", *Transactions in GIS*, 2012, 16 (3), pp. 411- 429
- [15] D. Holten. "Hierarchical Edge Bundles: Visualization of Ajacency Relations in Hierarchical Data", *IEEE Transactions on Visualization* and Computer Graphics, 2006, 12 (5), pp. 741-748
- [16] D. Holten, and J.J. van Wijk. "Force-Directed Edge Bundling for Graph Visualization", *Computer Graphics Forum*, 2009, 28 (3)
- [17] B.H. Ju, and K. Han. "Complexity Management in Visualizing Protein Interaction Networks", *Bioinformatics*, 2003, 19 Suppl 1, pp. i177-179
- [18] A. Koblin. "Flight Patterns", Science, 2006, 313 (5794), pp. 1733
- [19] C. Koylu, and D. Guo. "Smoothing Locational Measures in Spatial Networks", *Computers, Environment and Urban Systems*, 2013, 41, 12-25
- [20] G. Kumar, and M. Garland. "Visual Exploration of Complex Time-Varying Graphs", *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12 (5), pp. 805-812
- [21] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian. "Understanding Intra-Urban Trip Patterns from Taxi Trajectory Data", *Journal of Geographical Systems*, 2012, 14 (4), pp. 463-483

- [22] Y. Liu, Z. Sui, C. Kang, and Y. Gao. "Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-in Data", *PLOS ONE*, 2014, 9 (1), pp. e86026
- [23] D.F. Marble, Z. Gou, and L. Liu. Recent Advances in the Exploratory Analysis of Interregional Flows in Space and Time, in Kemp, Z. (Ed.): 'Innovations in GIS 4' (Taylor & Francis, London, 1997)
- [24] W. Mitchell, and M. Watts. "Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques", *Geographical Research*, 2010, 48 (1), pp. 24-41
- [25] S. Openshaw. The Modifiable Area Unit Problem (Geo Books, 1984.)
- [26] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. "Flow Map Layout", *IEEE Symposium on information visualization*, 2005, pp. 219-224
- [27] J.P. Poon. "The Cosmopolitanization of Trade Regions: Global Trends and Implications, 1965-1990", *Economic Geography*, 1997, 73 (4), pp. 390-404
- [28] A. Rae. "From Spatial Interaction Data to Spatial Interaction Information? Geovisualisation and Spatial Structures of Migration from the 2001 Uk Census", *Computers, Environment and Urban Systems*, 2009, 33 (3), pp. 161-178
- [29] D. Rafiei, and S. Curial. "Effectively Visualizing Large Networks through Sampling", *IEEE Visualization 2005 - (VIS'05)*, MN, October 23-28, 2005, pp. 48-56
- [30] R. Scheepens, N. Willems, H. Van de Wetering, G.L. Andrienko, N.V. Andrienko, and J.J. Van Wijk. "Composite Density Maps for Multivariate Trajectories", *IEEE Transactions on Visualization and Computer Graphics*, 2011, 17 (12), pp. 2518 - 2527
- [31] R. Scheepens, N. Willems, H. van de Wetering, and J.J. Van Wijk. "Interactive Visualization of Multivariate Trajectory Data with Density Maps". Proc. Pacific Visualization Symposium (PacificVis), 2011, pp. 147-154
- [32] R. Scheepens, N. Willems, H. van de Wetering, and J.J. Van Wijk. "Interactive Density Maps for Moving Objects", *IEEE Computer Graphics and Applications*, 2012, 32 (1), pp. 56 - 66
- [33] D.E. Seaman, and R.A. Powell. "An Evaluation of the Accuracy of Kernel Density Estimators for Home Range Analysis", *Ecology*, 1996, 77 (7), pp. 2075–2085
- [34] X. Shi. "Selection of Bandwidth Type and Adjustment Side in Kernel Density Estimation over Inhomogeneous Backgrounds", *International Journal of Geographical Information Science*, 2010, 24(5), pp. 643-660
- [35] B.W. Silverman. Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC, 1986. 1986)
- [36] W. Tobler. "A Model of Geographical Movement", *Geographical* Analysis, 1981, 13 (1), pp. 1-20
- [37] W.R. Tobler. "Spatial Interaction Patterns", Journal of Environmental Systems, 1976, 6 (4), pp. 271-301
- [38] W.R. Tobler. "A Model of Geographical Movement", *Geographical Analysis*, 1981, 13, pp. 1-20
- [39] W.R. Tobler. "Experiments in Migration Mapping by Computer", *American Cartographer*, 1987, 14, pp. 155-163
- [40] E.L. Ullman, and M.F. Dacey. "The Minimum Requirements Approach to the Urban Economic Base", *Papers in Regional Science*, 1960, 6 (1), pp. 175-194
- [41] K. Verbeek, K. Buchin, and B. Speckmann. "Flow Map Layout Via Spiral Trees", *IEEE Transactions on Visualization and Computer Graphics*, 2011, 17 (12)
- [42] J. Wood, J. Dykes, and A. Slingsby. "Visualisation of Origins, Destinations and Flows with Od Maps", *The Cartographic Journal*, 2010, 47 (2), pp. 117 - 129
- [43] J. Yan, and J.-C. Thill. "Visual Data Mining in Spatial Interaction Analysis with Self-Organizing Maps", *Environment and Planning B*, 2009, 36, pp. 466-486
- [44] D.A. Young. "A New Space-Time Computer Simulation Method for Human Migration", American Anthropologist, 2002, 104 (1), pp. 138-158
- [45] X. Zhu, and D. Guo. "Mapping Large Spatial Flow Data with Hierarchical Clustering", *Transactions in GIS*, 2014, 18 (3), pp. 421-435