Computers, Environment and Urban Systems 77 (2019) 101359

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Density-based multi-scale flow mapping and generalization

Xi Zhu[a,*], Diansheng Guo[a,b], Caglar Koylu[c], Chongcheng Chen[a]

[a] Key Laboratory of Spatial Data Mining and Information Sharing of MOE, Fuzhou University, Fuzhou, China
[b] Department of Geography, University of South Carolina, Columbia, SC, USA
[c] Department of Geographical and Sustainability, University of Iowa, Iowa City, USA

ARTICLE INFO

Keywords:
OD flow
Spatial interaction
Flow map
Multi-scale
Cartographic generalization

ABSTRACT

Mapping large volume of origin-destination flow data (or spatial interactions) has long been a challenging problem because of the conflict between massive location-to-location connections and the limited map space. Current approaches for flow mapping only work with a small dataset or have to use data aggregation, which not only cause a significant loss of information but may also produce misleading maps. In this paper, we present a density-based flow map generalization approach that can extract flow patterns and facilitate the analysis and visualization of big origin-destination flow data at multiple scales. Unlike existing methods that generalize flow data by spatial unit-based aggregation, our new flow map generalization algorithm is based on flow density distribution. To demonstrate the approach and assess its effectiveness, a case study is carried out to map 829,039 taxi trips within the New York City. With parameter settings, the proposed method can discover inherent and abstract flow patterns at different map scales and generalization levels, which naturally supports interactive and multi-scale flow mapping.

1. Introduction

Origin-Destination (OD) flow (or spatial interactions) data, which record the movements or connections between locations, have become increasingly available and accurate due to the wide adoption of location-aware technologies. Examples of such data include taxi trips, county-to-county migration, cell phone calls and commuting trips. Mapping and understanding massive origin-destination flow data is fundamentally important for a wide range of research fields and domains such as demography, urban planning, transportation and epidemiology.

Flow map (Tobler, 1987, 1981) and space-time cube (Kraak, 2003; Kraak & Koussoulakou, 2005; Miller, 1991) are commonly used in visualizing OD flow data. However, these existing approaches are limited in mapping large flow data due to the problem of occlusion and cluttered display. Traditional flow map generalization methods usually aggregate large flow data by predefined areas, which suffer from the modifiable areal unit problem MAUP (Openshaw, 1984). Fig. 1 illustrates the MAUP problem by aggregating the same flow dataset (which is also used in our case study) with two different sets of spatial units: census tracts and zip codes. The two maps show very different flow patterns, and neither is close to the inherent patterns that we will show later in our case study.

In this paper, we propose a flow map generalization approach to enable multi-scale flow pattern discovery and visualization. Unlike existing methods that generalize flow data by spatial unit-based aggregation, our new flow map generalization algorithm is based on flow density distribution and consists of two steps: (1) density estimation, and (2) flow selection. In the first step, we estimate the flow density using kernel density estimation. We treat each OD flow as a four dimensional (4D) spatial point. The four dimensions refer to the x and y coordinates of both origin and destination locations. In the second step, we select flows with density values that are the greatest in their neighborhoods to represent the overall flow density distribution. In order to evaluate the utility of the proposed approach, we demonstrate a case study for analyzing and mapping big data of taxi trips in New York City.

2. Related works

2.1. Flow map generalization

Flow maps have long been used to visualize a wide range of spatial interactions such as human migration, transportation, commuting, commodity and information flows (Koylu, 2018; Phan, Xiao, Yeh, & Hanrahan, 2005; Tobler, 1981, 1987; Wood, Slingsby, & Dykes, 2011).

* Corresponding author at: No. 2, Xueyuan Rd, Fuzhou 350116, China.
E-mail address: zhuxiflying2012@gmail.com (X. Zhu).

https://doi.org/10.1016/j.compenvurbsys.2019.101359
Received 21 November 2018; Received in revised form 24 June 2019

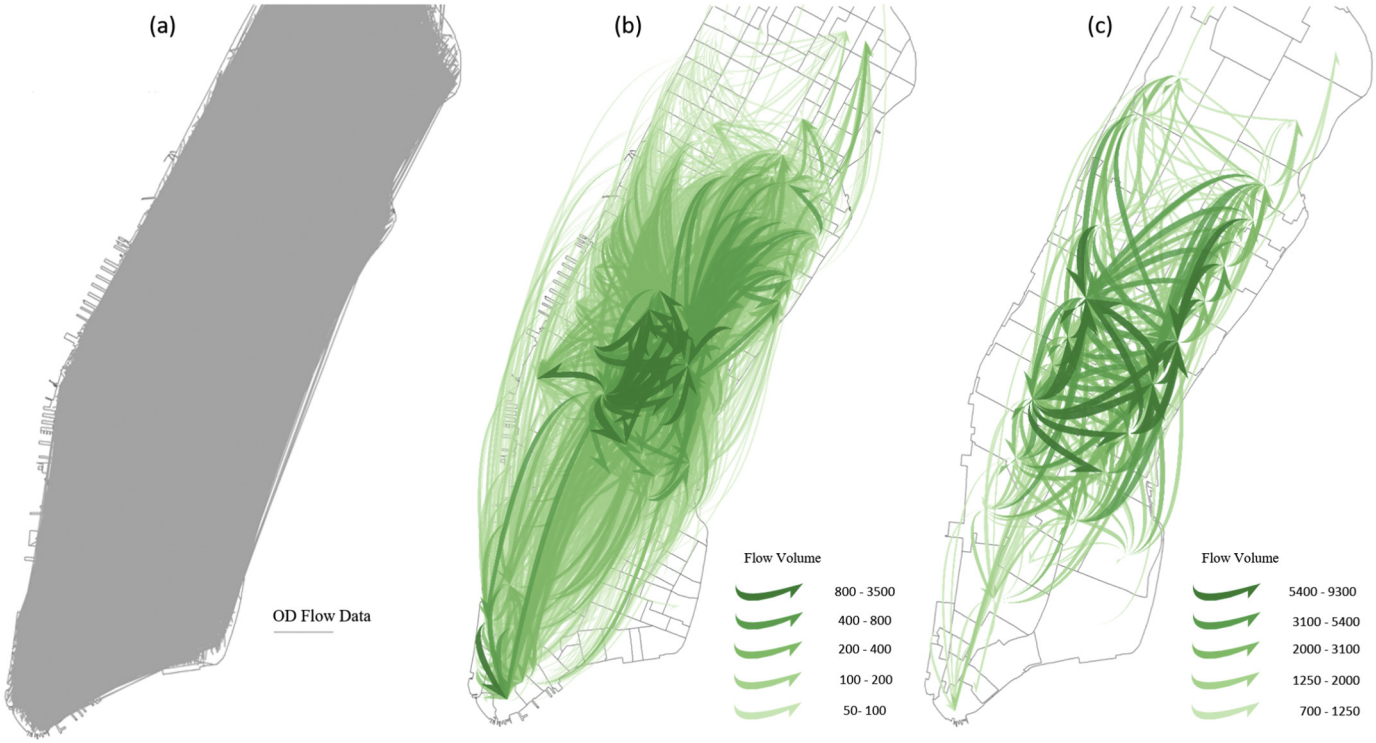

Fig. 1. An illustration of MAUP in flow data visualization. (a) A set of New York taxi trips data: it is difficult to visualize the flow data without generalization. Flows aggregated by (b) census tracts and (c) zip code boundaries demonstrate different flow patterns. Both flow maps (b) and (c) illustrate 637,116 flows (77% of the total flows).

Mapping large volumes of origin-destination flows is a challenging problem. As Andrienko, Andrienko, Dykes, Fabrikant, and Wachowicz (2008) suggested, existing visualization methods need a higher degree of abstraction to extract high-level abstract patterns from flow data. Flow map generalization methods aim to obtain a higher degree of abstraction and can be classified into two types: (1) spatial unit-based aggregation, and (2) flow-based aggregation. Spatial unit-based aggregation combines spatial units into larger regions, and aggregate flows between regions, which substantially reduces the number of flows. To discover the optimal groups of spatial units, graph-partitioning and clustering methods have been used. Guo (2009); Guo, Jin, Gao, and Zhu (2018) introduced a spatially-constrained graph partition method that can detect spatial communities as the aggregation units. Alternatively, Adrienko and Adrienko (2011) employed a clustering method to partition the territory into suitable places based on the density of trajectory points. Despite the discovery of spatial community structures, spatial unit-based aggregation methods disregard flow patterns at local scales, result in a significant loss of information and suffer from the MAUP, i.e., flow patterns may vary simply due to different aggregations. Specifically, the MAUP contains two aspects: (1) the sizes of the spatial units, which are related to the scale effect of the analysis and visualization; and (2) the shapes of the spatial units, different zoning schema may produce varies flow patterns.

Alternative to spatial unit-based aggregation, flow-based aggregation methods simplify the flow data by clustering flows. Zhu and Guo (2014) proposed a hierarchical clustering method to aggregate OD data based on the spatial similarity between nearby OD flows. Tao and Thill (2016) applied hot spot detection method on flow data to detect spatial flow clusters. Guo and Zhu (2014) extracted flow patterns by using a flow smoothing model to remove the effect of size differences between units. After the flow map generalization, a set of simplified flow data or flow data clusters are visualized on a flow map. The generalization process reduces the data size but retains the important flow patterns. In this study, our proposed method extracts high level flow patterns from the data density distribution.

2.2. Cartographic generalization and design of flow map

There is an emerging body of literature on the cartographic design principles for flow mapping based on user experiments (Jenny et al., 2018; Koylu & Guo, 2017; Yang et al., 2019; Yang, Dwyer, Goodwin, & Marriott, 2017). Similar studies have also been conducted in the graph drawing community (Alper, Bach, Riche, Isenberg, & Fekete, 2013; Dwyer et al., 2009; Xu, Rooney, Passmore, Ham, & Nguyen, 2012), which may contribute to the design of flow maps. However, flow map reading is different from graph drawing studies in that the former is focused on holistic and geographic patterns that could vary based on different spatial scales. One of the design principles for flow mapping is to minimize edge crossing to achieve visual clarity. A number of strategies have been introduced to address the visual cluttering of flows such as edge rerouting/bundling (Buchin, Speckmann, & Verbeek, 2011; Cui, Zhou, Huamin, Wong, & Li, 2008; Phan et al., 2005), line shortening (Koylu & Guo, 2017) or producing non-branching flows by adjusting the curvature of flow lines (Jenny et al., 2017). Despite these efforts for improving the design of OD flow maps, there are also a few design principles remain debatable. Xu et al. (2012) found that users prefer straight lines over curved lines while others (Jenny et al., 2018; Purchase, Hamer, Nöllenburg, & Kobourov, 2012) have found that curved lines are more effective than straight lines. For curved flow lines, Jenny et al. (2018) suggest using symmetric curvature for flows, while others recommend the use of asymmetric flow lines to encode direction (Guo, 2009; Koylu & Guo, 2017; Ware, Kelley, & Pilar, 2014). Koylu and Guo (2017) show with experiments that the choice of symbolization may depend on different tasks (e.g., the identification of the dominant flow directions or the strongest flows) and different data patterns also have a strong effect on task performance and pattern perception in flow maps. It is difficult to generalize a universal set of design guidelines to create flow maps for different tasks and across

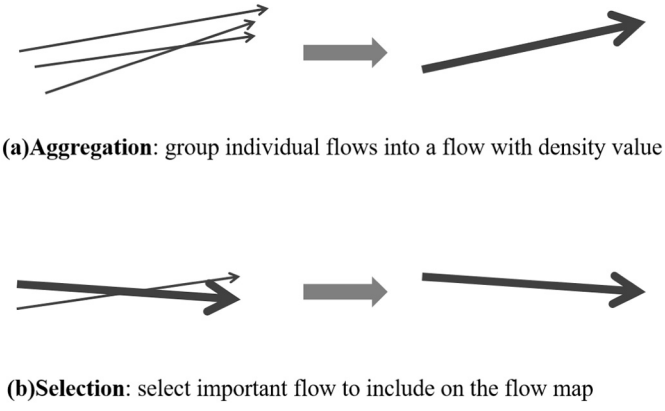

Fig. 2. Principles of flow map generalization method which consists of aggregation and selection techniques in cartographic generalization.

various datasets. Interactive techniques may provide a viable solution, within which one can customize the symbolization and layout of flow maps according to the data, task and the users' preference.

The design principles gained through experimental user tests can improve visual clarity and the design of flow maps and symbolization, but these principles are only effective for small flow datasets (e.g., up to one hundred flows). To cope with large volume of flow data, flow data abstraction or generalization has to be conducted first before applying these design principles. Map generalization methods seek to summarize detailed spatial information, reduce details, and render abstract information on maps. Meanwhile, it is critically important that the generalized map represents the original data faithfully, although with less details.

2.3. Flow data visualization and visual analytics

There are a variety of visualization and visual analytics methods for flow data such as location-based and matrix-based flow visualizations. These methods bypass the visual cluttering problem by not mapping flows directly. For example, one strategy is to visualize location measures such as net flow ratio for different time durations derived from OD flow data (Guo, Zhu, Jin, Gao, & Andris, 2012), which provides insights into the characteristic of locations in terms of flows (Koylu & Guo, 2013). The drawback of the location-based measures is that the links between locations are lost. Andrienko, Andrienko, Fuchs, and Wood (2016) introduced a spatial and temporal abstraction method to represent location measures by diagram maps instead of flow maps. The proposed method includes composite glyphs for each location to visualize the flow angle and distance to retain the links between locations.

Alternatively, OD flow data can be visualized using an OD matrix rather than plotting the OD flow data as vectors (Ghoniem, Fekete, & Castagliola, 2004). In an OD matrix, the rows represent the locations of flow origins while the columns represent the locations of destinations. Additionally, reordering and aggregation techniques (Guo & Gahegan, 2006) can enhance the utility of OD matrices to cope with large dataset. The limitation of OD matrix is the loss of spatial perception and reasoning skills due to the missing routes and the geographic context. To overcome this limitation, Wood, Dykes, and Slingsby (2010) proposed OD map, which attempts to retain the geographic context as much as possible.

Researchers have developed visual analytics methods to assist users to understand OD flow data. Ferreira, Poco, Vo, Freire, and Silva (2013) proposed a visual model that supports spatiotemporal queries of origin-destination data, which is based on users' choices of regions to aggregate flows. Boyandin, Bertini, Bak, and Lalanne (2011) proposed a view that contains two maps, and places the origin and destination separately on these two maps in order to avoid flow lines occlusion. These approaches can relieve the visual cluttering problem by selecting a subset of data, but the limitation is that they cannot provide a clear holistic overview of spatial flow patterns.

3. Methodology

In this section, we present our approach to mapping massive origin-destination flow data at multiple spatial scales and generalization levels. The key component of this approach is a density-based flow map generalization algorithm, which selects a subset of flows to represent the density distribution of an OD dataset. The design of our approach aligns with the principle of cartographic generalization. Cartographic generalization methods seek to summarize detailed spatial information into an abstract form with fewer details that can be rendered on a map, which meanwhile represent the original data faithfully with minimum information loss. Essentially, cartographic generalization methods such as *aggregation* or *selection*, aim to reduce data complexity by merging multiple features or selecting important elements. We design our flow map generalization method based on the principles and techniques of cartographic generalization.

Our density-based flow map generalization method for mapping large volumes of OD flow data consists of two steps: (1) density estimation, and (2) flow selection. In the first step, we estimate the flow density using kernel density estimation. We treat each OD flow as a four dimensional (4D) spatial point. The four dimensions refer to the x and y coordinates of both origin and destination locations. This step is similar to *aggregation*, which merges multiple individual observations into a feature. Therefore, each flow data with its flow density value represents the nearby flows. In the second step, we select flows with density values that are the greatest in its neighborhood. This step reduces data complexity by selecting the most import flows, and the density values calculated in the first step serve as the measurements of importance.(See Fig.2)

Fig. 3 is an illustration of our method using 829,039 taxi trips within Manhattan Island, New York. This dataset is a subset of the New York taxi trip data (NYC Limousine Commission, 2018) and only contains the trips start between 7 AM and 8 AM of every Monday in 2009. With this dataset, we demonstrate the efficiency of our method, and investigate mobility patterns in the morning rush hour of a workday. Fig. 3(a) illustrates the original data and each OD flow is represented as a straight line in the map. Due to the cluttering and overlapping of flow lines, this map does not provide any useful information about the data. In Fig. 3(b), the flow lines are classified and symbolized with different colors according to their density values. In Fig. 3(c), only a small portion of the original flows are selected and visualized by our selection method, which enables a high-level abstraction of the original dataset. Admittedly, this map still has the problem of visual cluttering, which can be improved with flow map design in the following sections.

In following subsections, we introduce each of the steps in detail. In Sections 3.1 and 3.2, we introduce the density estimation and flow selection steps. In Section 3.3, we provide several suggestions on flow map design. In Section 3.4, we apply the flow map generalization method at multiple spatial scales to enable multi-scale flow mapping with different parameter settings.

3.1. Flow density estimation

We adopt the following definitions to describe the parameters:

Flow: Let $T = (O_x, O_y, D_x, D_y)$ is an OD flow in which (O_x, O_y) are the coordinates of the origin point, and (D_x, D_y) are the coordinates of the destination point.

Distance between Flows: Let $T_i = (O_{xi}, O_{yi}, D_{xi}, D_{yi})$ and $T_j = (O_{xj}, O_{yj}, D_{xj}, D_{yj})$ are two OD flows, the Euclidean distance between flows is defined as:

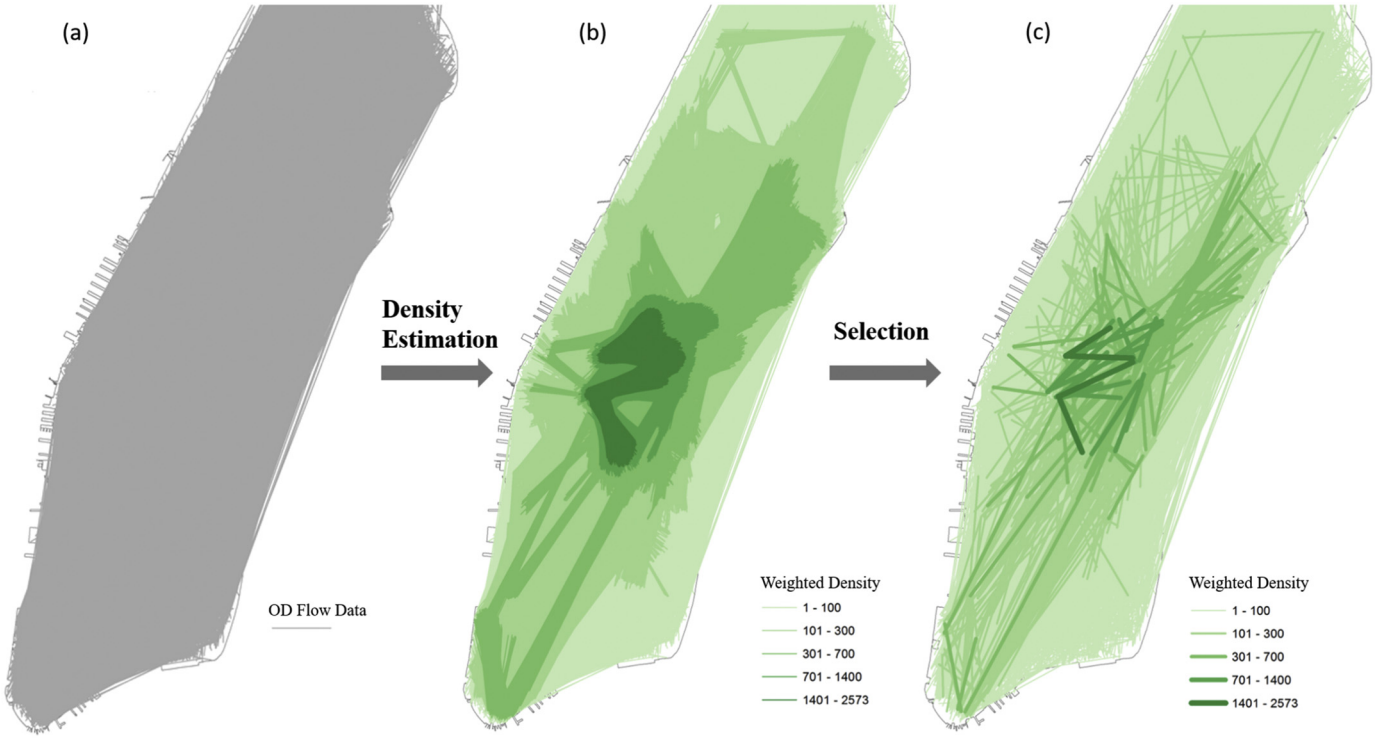

Fig. 3. An illustration of the density-based flow map generalization method. This method contains two steps: (1) density estimation, and (2) flow selection. Map (a) illustrates the original data. Map (b) classifies the flow lines according to the density values. Map (c) visualizes a set of selected flow lines to reduce the data complexity.

$$Dist(T_i, T_j) = \sqrt{(O_{xi} - O_{xj})^2 + (O_{yi} - O_{yj})^2 + (D_{xi} - D_{xj})^2 + (D_{yi} - D_{yj})^2} = \sqrt{d_O^2 + d_D^2}$$

where d_O is the Euclidean distance between two origins, and d_D is the Euclidean distance between two destinations. Intuitively, only the flows that have nearby origins and destinations considered as nearby flows, whereas lengths of these flows become irrelevant.

Flow Neighborhood: The neighborhood of a flow T_i is defined as: $N(T_i, d) = \{T_j \in T |\ \text{Dist}(T_i, T_j) < \text{R}\}$, where R is the radius of the neighborhood.

The definition of flow neighborhood is based on the definition of distance between flows. The elements whose distance to T_i is less than R are within T_i's R-size neighborhood.

Kernel density estimation is a statistical technique for removing spurious data variation and estimating a reliable density value of a feature using the observations within its neighborhood (Silverman, 1986). The kernel density estimator is defined as following:

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) Y(x_i)$$

where $x_1, x_2, \ldots, x_n \in R^d$ is a set of data points. K is the kernel function, and h is a smoothing parameter bandwidth. The kernel function can be Epanechnikov, Triangular, or Gaussian. $Y(x_i)$ is the population at points x_i.

In this step, we apply the kernel density estimation method on the OD flow dataset based on the definitions of *distance between flows* and *flow neighborhood*. In the formula, x_i - x represents the distance between flows. We use the flows within the neighborhood to estimate the data density. We use h to determine the radius of flow neighborhood in this step. Note, there is no population field for the case study, and each flow represents one individual taxi trip, so $(x_i) = 1$. However, it is practical to extend the density estimation method to data with a population field. For example, the number of migrants for each county-to-county flow can be used as the population field in a migration dataset.

The choice of bandwidth h has a strong influence on the result of density estimation. Given a dataset, a bandwidth value that is too small may cause under-smoothed estimation while one that is too large may cause over-smoothed estimation. The most common optimality criterion used to select bandwidth such as MISE (Mean Integrated Squared Error) or AMISE (Asymptotic MISE) cannot be used directly since the true density distribution of the dataset is unknown. Silverman's rule (Silverman, 1986) is an approximation method based on the assumption that the underlying density being estimated is Gaussian, which suggests $h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}}$ where σ is the standard deviation of the samples, and n is the number of the observations. The standard deviation of a flow dataset can be calculated by the following steps: (1) calculate the mean center $\overline{O_x}$, $\overline{O_y}$, $\overline{D_x}$, $\overline{D_y}$ of the dataset. (2) calculate the distance d_i to the mean center for each flow based on definition of *distance between flows, and* $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(d_i)^2}{n}}$. The standard deviation is a measure to quantify the amount of variation or dispersion in a set of data values, and the bandwidth h can be selected according to the variation in a dataset. Using the Silverman's rule, we derive the bandwidth of the taxi trip data to be approximately 100 m (98.6 m).

In this step, the bandwidth h is the parameter setting to determine the spatial scale. Larger bandwidth h uses a larger spatial neighborhood to estimate the data density and results in flow patterns among larger regions. Fig. 4 is a list of maps to illustrate the effect of varying bandwidth on the density estimation. All these maps present meaningful information about the OD flow data. In Fig. 4(a), smaller bandwidth (100 m) can capture flow patterns among small regions, while Fig. 4 (c) uses a larger bandwidth (400 m) to capture flow patterns among large regions. A detailed discussion about parameter h is provided in Section 3.4.

After density estimation step, the density distribution can be visualized by a curve or a raster image. While it is straightforward to visualize density distribution of 1D, 2D or 3D spatial points, it is difficult to visualize the density distribution of flow data. The maps in Fig. 4

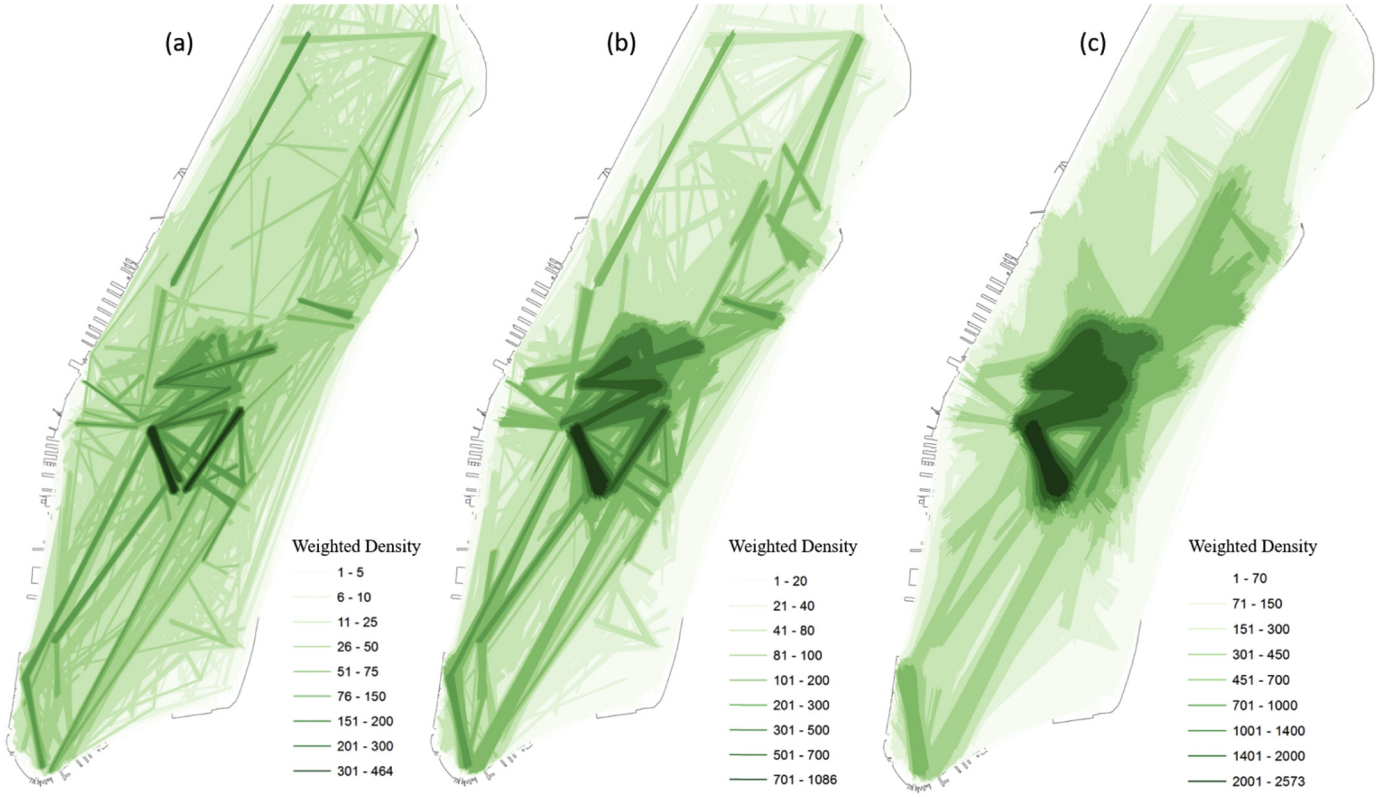

Fig. 4. An illustration of flow density estimation with different bandwidths h. (a) 100 m (b) 200 m (c) 400 m.

suffer from severe cluttering problem. Hence, we need a generalization method to visualize the density distribution.

3.2. Flow selection and generalization

To generalize the density distribution of flow data generated in the previous step, we select a subset of representative flows with local maximal density values and visualize the density distribution of flow data. Essentially, there are two objectives in the generalization process: (1) preserve overall patterns by selecting important and representative flows and (2) maintain the clarity of the flow map by only selecting a small subset of flows that are not too close to each other. For the first objective, the density values calculated in the first step serve as the measurements of importance. To achieve the second objective (i.e., avoiding cluttering), we only select flows with local maximal density values, and these local maxima are not too close to each other.

The selection step selects a flow only if its density value is the local maximum within its neighborhood. The parameter R is the radius of flow neighborhood, which is used to determine the generalization level. A larger R makes it more possible to find a stronger flow in the neighborhood, therefore it is harder for flows to get selected, since there will be less number of flows to be selected. With a larger R, we can generate a flow map at a higher generalization level. In contrast, with a smaller R, we can generate a flow map at a lower generalization level by selecting more flows.

Fig. 5 is a list of maps to illustrate the effect of R on flow map generalization. In these three maps: (1) As in Fig. 5.c, a larger search radius selects fewer flows and improves the clarity of the flow map. However, details of flow patterns are lost. (2) The selected flow set by a larger R (e.g., Fig. 5.b) is a subset of the selected flow set with a smaller R (Fig. 5.a). This is a tradeoff between information abundance and map clarity, which is controlled by the parameter R. A detailed discussion about parameter R is provided in Section 3.4.

In map generalization, the cartographer is responsible for selecting the most necessary elements and suppressing the unimportant details to reduce data complexity and achieve map clarity and balance. It is a challenging problem to automate such a selection process, which is of a critical need for dealing with a big dataset and supporting interactive data exploration across scales. In this step, we generalize the flow map by selecting a subset of flows with local maximal density values. In the next section, we will apply flow map design to further address visual cluttering problem.

3.3. Cartographic design of flow map

Generalized flow maps illustrated in previous sections still suffer from visual cluttering. Moreover, these flow maps do not include visual encoding for direction of the flows. We adopt the following cartographic design principles to increase readability and aesthetics of generalized flow maps. The overall guideline for these design principles is to emphasize flows with high density values and weaken flows with low density values:

- *Drawing order*: Draw flows with greater density values on top of flows with lower density values to avoid strong flows being obscured.
- *Color value and line thickness*: Use darker tones and thicker lines to represent strong flows and use lighter tones and thinner lines to represent weaker flows.
- *Asymmetric curvature*: Use asymmetric curves to represent direction of flows. Flows are curved at the origin and become straight at the destinations.
- *Partial arrows*: Use partial arrows to indicate direction of flows.
- *Transparency*: Use higher transparency settings for the flows with lower density values.
- *Symbol outlines*: use outlines to highlight strong flows.

In addition to mapping flows, we use location-based flow measures

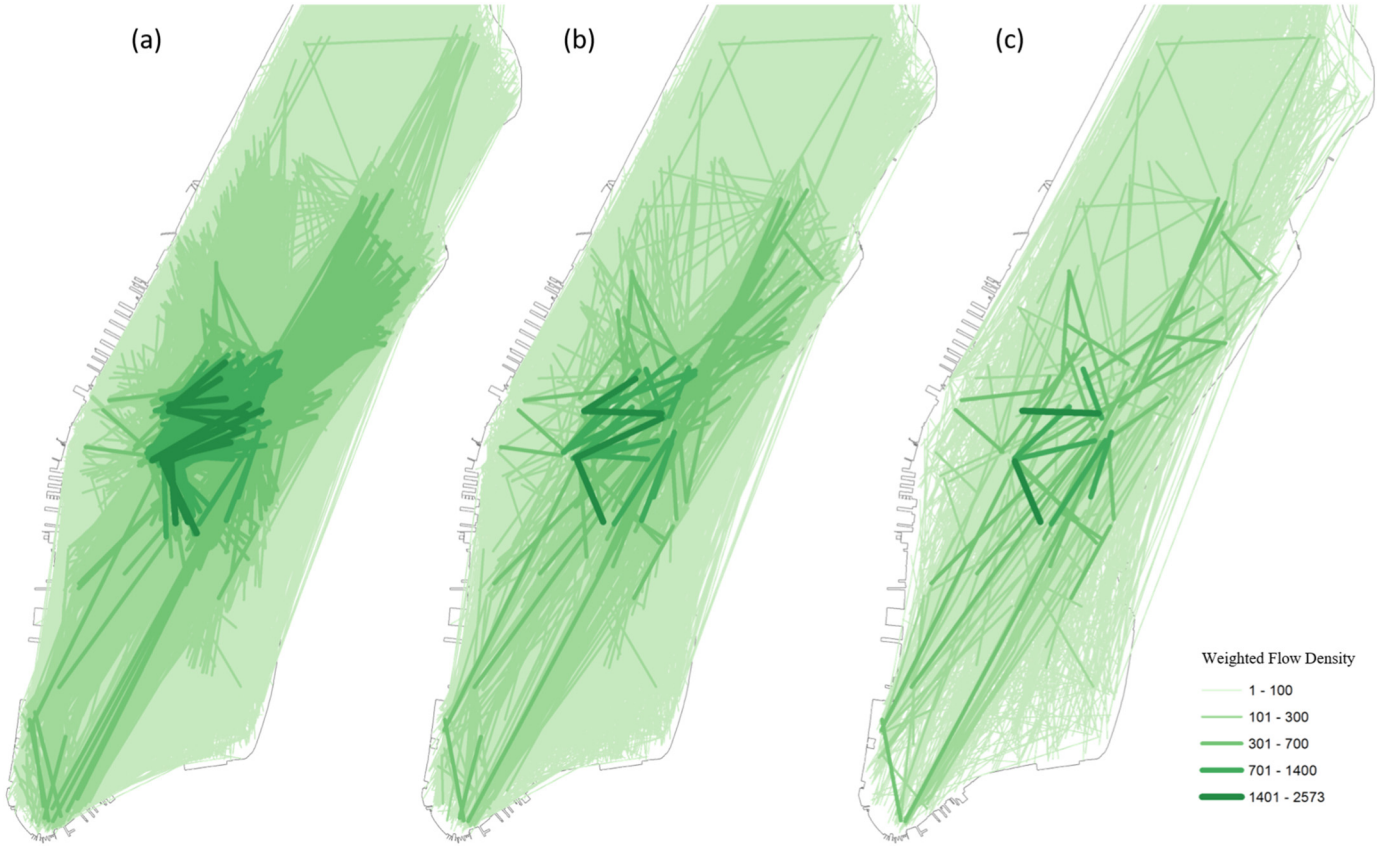

Fig. 5. An illustration of flow selection with varying neighborhood sizes. We use the same classification and symbolization for these three maps. (a) 82,541 flows are selected with R = 200 m (b) 13,253 flows are selected with R = 400 m (c) 1999 flows are selected with R = 800 m. This selection process is based on the density values calculated with h = 400 m, which is demonstrated in Fig. 4(c).

as a basemap to illustrate inflow or outflow hotspots. Overlaying of flows with locational flow measure helps users to identify hot spots of flows and identify the in and out connections of those spots. We suggest that straight lines are more appropriate for showing outflow patterns (see Fig. 9), while curved lines are better to demonstrating inflow patterns (see Fig. 10). It is flexible to switch between straight lines and curved lines (with different curvatures) in an interactive software environment.

By applying the cartographic designs principles mentioned above, we turned the flow map in Fig. 5(c) into the flow map in Fig. 6. In this flow map, we can get a holistic overview of flow patterns in the morning rush hour in Manhattan. Many taxi trips start from transportations hubs, and end at office buildings at Midtown areas and Financial District. Locations like the convention center and hospitals are also common destinations for taxi trips. We apply high transparency settings and pastel colors on the weak flows to emphasize the strong flows. These flows serve as a complement for the major flow patterns, and this design can avoid having them distract map users and blur major patterns.

3.4. Multi-scale flow mapping and parameter configuration

In this section, we present methods to configure parameter settings. We use a flow neighborhood with size h to estimate the flow density distribution. The parameter h is used to determine the spatial scale. In flow selection, we use a flow neighborhood with size R to select flows with local maximal density values. The parameter R is used to determine the generalization level.

First, we suggest the parameters should be determined based on the dataset. For the parameter h in the density estimation step, Silverman's rule is a data-driven method that recommends an optimal parameter according to the data distribution. We propose selecting a bandwidth range rather than a single optimal bandwidth. We employ a data-driven bandwidth setting for a moderate scale and modify the bandwidth above and below this value to explore scale-dependent flow patterns (see Fig. 7). This strategy enables multi-scale flow mapping based on user interactions such as zoom in and out. For parameter R in the selection step, we design a data-driven method to set the parameter in the flow selection step. In Fig. 8, we plot parameter R and the corresponding number of selected flows. As the R increases, the number of selected flows decreases. Across the range of 100 to 800 m for taxi trip data, information abundance and map clarity are traded off. This flow selection strategy with a range of values can help users select different levels of abstraction in an interactive environment. The users can select a larger neighborhood size if they prefer a more abstract map with fewer details and select a smaller size if they need more information and can tolerate some visual cluttering. We only experiment with an optimal range of the two parameters in this study, however, the control of the parameters with a full range of values can be integrated into an interactive software environment, which we plan to complete in a future study.

Second, parameter h and R should cooperate with each other. In this study, we suggest that R should be equal to or greater than h. In general, the flow map for a large spatial area (small spatial scale) needs a higher level of generalization to keep the map clear. The flow map for a local area needs a lower generalization level to make the map informative with local flow patterns. Hence, given a spatial scale, the generalization level can be determined accordingly. In other words, the parameter R should be determined by the parameter h. In this case study, we set R = 2 h to produce the following flow maps at different spatial scales in Figs. 6, 9 and 10.

In Fig. 9, the map highlights the flow patterns within an area of

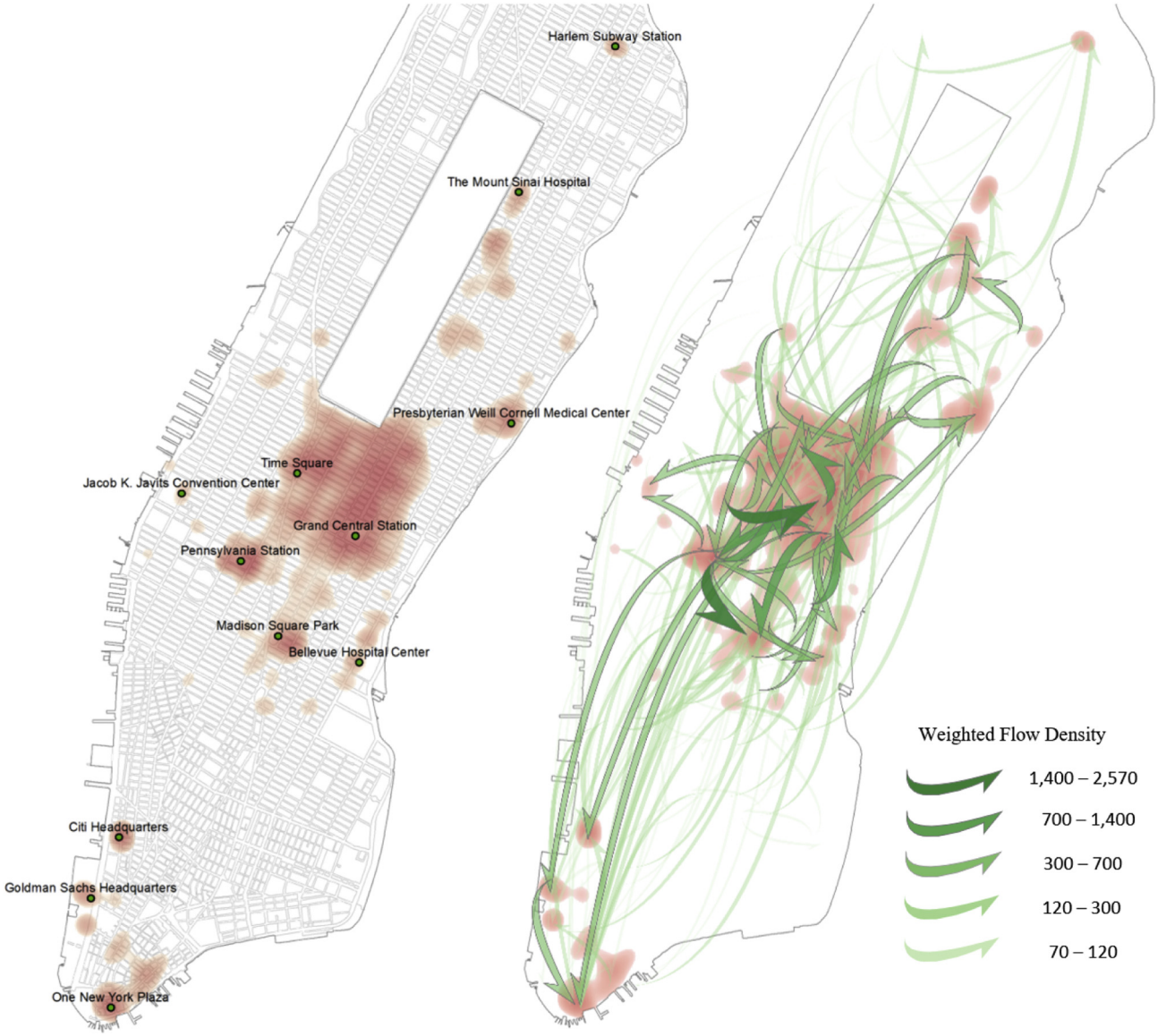

Fig. 6. The flow map demonstrates the major flow patterns in the morning rush at Manhattan, which generalized from 829,039 taxi trips. The left map with road networks and referenced locations help us understand the flow map on the right. The red hues illustrate hotspots of destination points, where points density higher than 30,000/km^2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

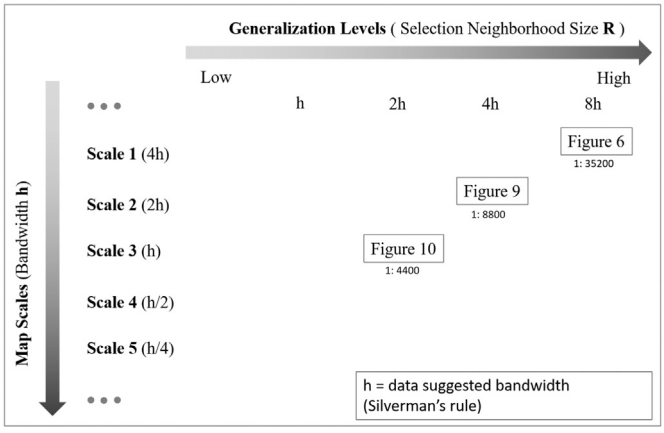

Fig. 7. A parameter setting schema for multi-scale flow mapping. The ratio numbers indicate actual map scales of the flow maps given in the corresponding figures.

6 km by 8 km in Midtown Manhattan. With h equals to 200 and R equals to 400, the top 200 flows are drawn in this map. In the morning, the taxi trips start from the major transportation hubs and end at office buildings because taxis complement public transportation.

In Fig. 10, we further scale down the map extent to a 3 km by 4 km area at East Harlem District. With h equals to 100 and R equals to 200, top 100 flows draw in this map. The flow patterns match with the destination hot spots very well.

From flow maps in Figs. 6, 9, and 10, we can gain overviews of the dominant flow patterns at multiple spatial scales. The discovered flow patterns are different from the patterns discovered by previous methods based on spatial-units aggregation, which is shown in Fig. 1. Our new method is not sensitive to MAUP, the flow maps generated without any predefined spatial units. All the flow lines in each flow maps produced by our method are comparable, which represent flow volumes in spatial units of the same size. However, in the flow maps generated by spatial-units aggregation, the stronger flows often are a result of larger spatial units. In addition, compare to the flow maps in Fig. 1, our method produces less flow lines to generalize the overall flow patterns, and the flow maps are easier to read.

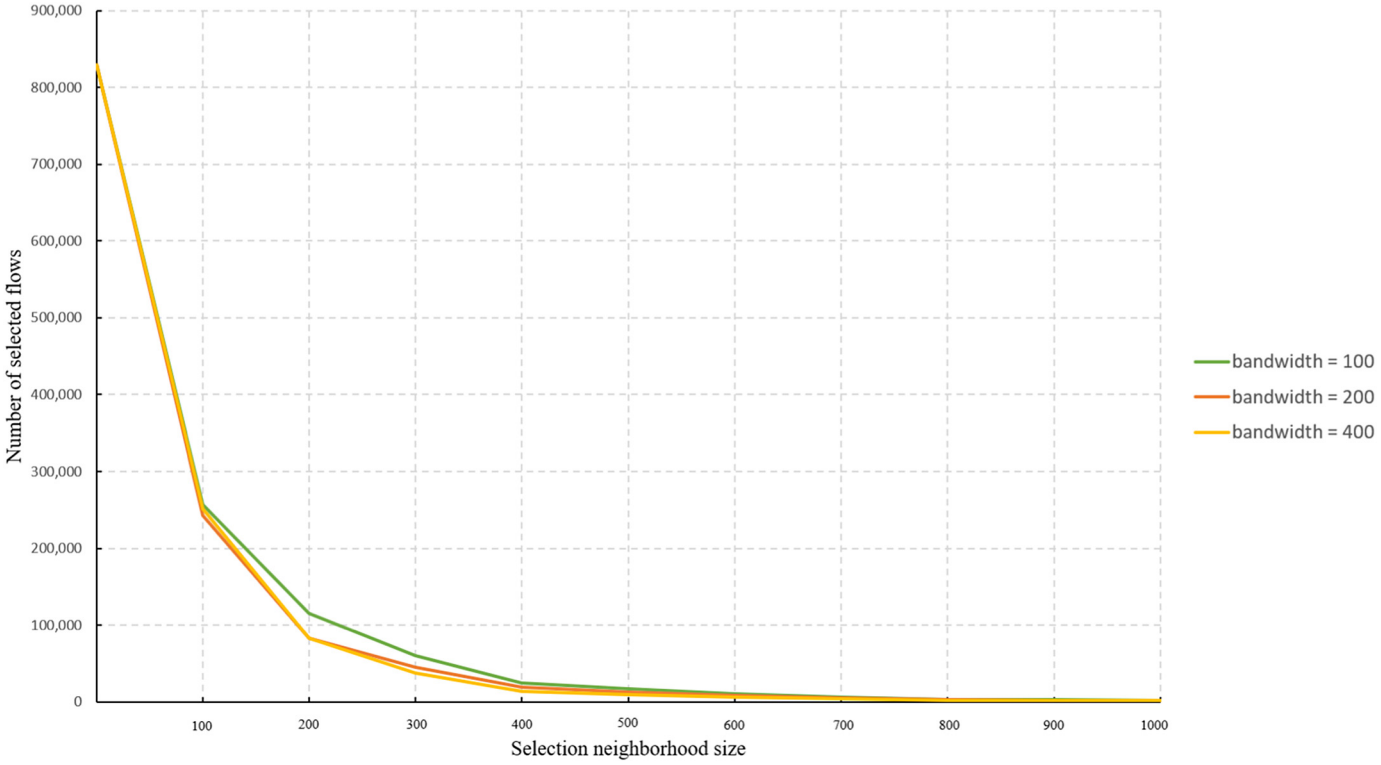

Fig. 8. The neighborhood size and the corresponding number of selected flows.

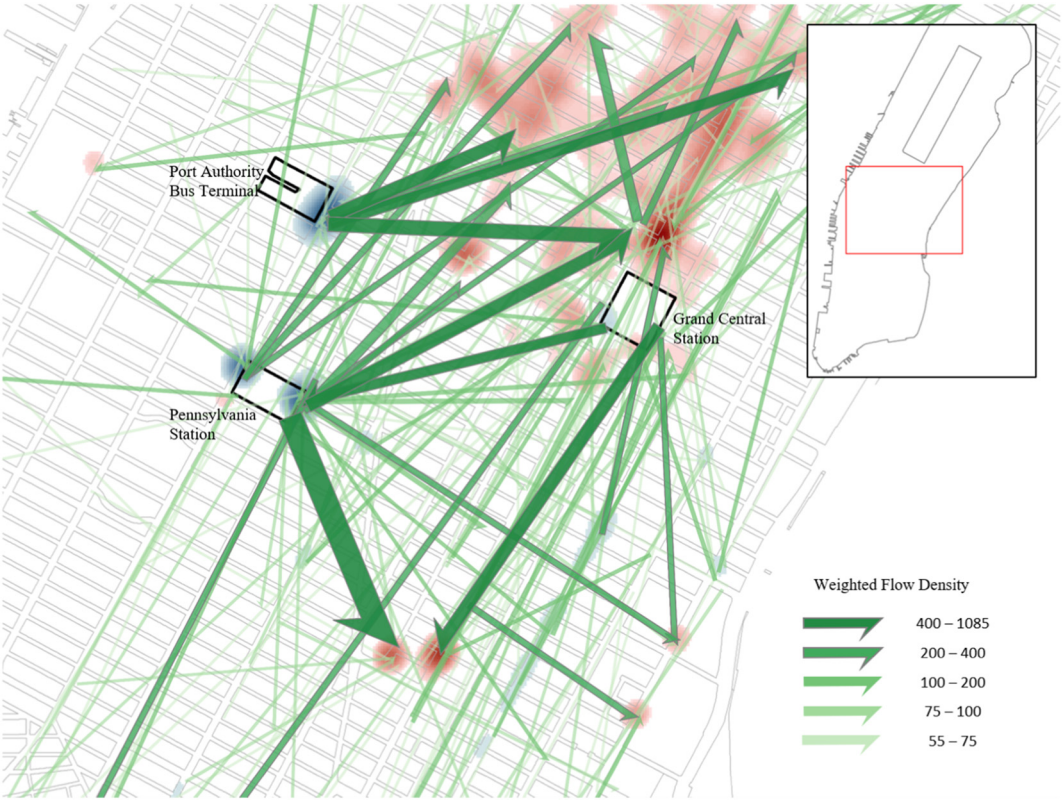

Fig. 9. The flow map demonstrates the major flow patterns in the morning rush hour at Manhattan downtown. Red hues represent high density of destination points, and blue hues represent high density of origin points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

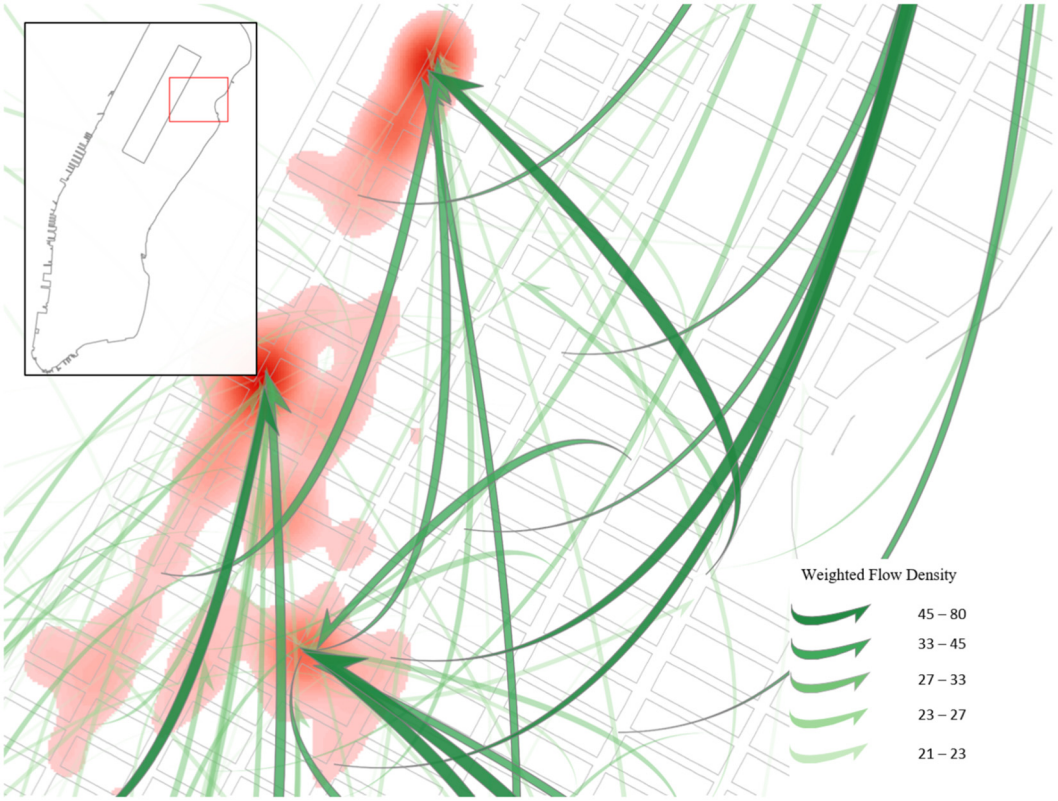

Fig. 10. A flow map demonstrates the flow patterns at local scale.

OD flow data are often used to describe and predict flow volumes in traffic network, which is essential for dynamic traffic control and intelligent traffic network management. Also, taxi OD flow data could be used to assess public transportation demand. Complementary to other analytic methods such as statistical modeling, visualization of OD flow data is a straightforward way to help urban planners to understand the data. Moreover, our method is not limited to OD commuting data in urban environment, it is also useful to analyze other types of OD data, such as migration, international trade, spatially embedded social networks, and so on.

3.5. Computational complexity

The computational complexity of the density estimation step is O (nlogn). This step is implemented by a neighborhood search, the density value of a flow equals to the weighted number of observations within its neighborhood. With the assistance of a spatial R-tree index, the time complexity for each neighborhood search is O(logn), and the overall time complexity for this step is O(nlogn). The computational complexity of the selection step is O(nlogn). The selection step identifies flows with local maximal density values. It is also implemented by a neighborhood search. For each flow, we test each flow whether it is the local maximal within its neighborhood.

With the complexity of O(nlogn), it is viable to enable real-time interactive multi-scale flow mapping with a dataset less than one million data items. However, for a larger dataset, such complexity is too high for real-time interaction. To handle large datasets, we can pre-calculate the density values in advance once the bandwidth for each map scale is selected. Such an approach is commonly used in web-mapping, in which tiles of base-maps at different scales are pre-computed and cached in the memory through web-services. For the selection step, we can pre-calculate a critical distance for each flow data T_i, that is the distance to the nearest flow which density value greater than T_i. By doing so, the flows that the critical distance are less than the selection radius R can be simply omitted. In addition to these pre-calculation solutions, we can use sampling method or fine aggregation to reduce the data size before adopting the proposed method.

3.6. Discussions

The main advantages of the presented method are twofold: (1) It is a data-driven method: the density estimation and selection steps are based on the data distribution. We also presented heuristic data-driven methods to suggest a range of parameter settings for both of the two steps. In essence, we presented a density mapping method for large volumes of OD flow data. (2) The parameter bandwidth h and selection radius R affect the flow patterns, which also help us to discover flow pattern at different map scales and generalization levels. Based upon past success of visual analytics approaches in many domains, we plan to develop an interactive framework to let the users determine these two parameters, within the constraint of our suggested range of parameter settings.

Our approach allows selecting a subset of flows to represent the overall density distribution. To achieve map clarity, we only selected the flows with local maximal density values. It is inevitable that some information is lost in such process. This case is even worse when the density distribution is over-smoothed with a large h, or when we select fewer number of flows with a large R. Local maxima can be good indicators of the density distribution, but we need to remind the map readers about the existences of unrepresented data.

The comparison between the proposed method and the spatial unit-based flow aggregation is similar as the difference between kernel density mapping and histogram. Kernel density mapping and histogram are two commonly used density estimators. However, the histogram is sensitive to the anchor position (bin origin) and the bin size (Härdle, 2012). Spatial unit-based flow aggregation is similar to histogram, where the spatial units are the bins in the histogram. Compared to the spatial unit-based flow aggregation method, our proposed method can

visualize the density distribution without such spatial units.

To verify the accuracy of the discovered flow patterns, we calculated inflow and outflow hotspots as the flow map background. We can observe the destinations of flow lines match with inflow hotspots very well. Furthermore, we conducted the same verification in different scales (Figs. 6, 9, and 10), which also demonstrated the method is not sensitive to parameter settings.

4. Conclusions

In this article, we introduced a density-based flow map generalization method that is scalable for mapping large origin-destination flow data at multiple scales. This method can be used for interactive analysis of flow patterns. Different bandwidths h reveal the patterns in different spatial scales. Larger bandwidths can discover general patterns among large areas, while smaller bandwidths can discover detailed patterns among small areas. The selection parameter R determines the levels of map generalization: larger search radius selects fewer representative flows and generates a clear flow map. In doing so, it will lose details of the flow patterns. We carried out a case study with the taxi trip data in New York City to demonstrate the usefulness of our proposed approach. Results show that the proposed method can effectively discover patterns in different scales with different levels of abstraction.

We plan to implement our proposed method in an interactive geovisual analytics environment. The interactive software will allow users to determine the scale and generalization parameters with our suggested data-driven parameter settings. By combining intelligent algorithms with human cognitive abilities, our methodology can help derive insights from large flow data. Additionally, usefulness of our proposed cartographic design choices and overlaying of location based measures point to future empirical user studies in flow mapping.

Acknowlegements

The research is supported by the national key research and development program of China under grant 2017YFB0504202.

References

Adrienko, N., & Adrienko, G. (2011). Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics, 17*(2), 205–219.

Alper, B., Bach, B., Riche, N. H., Isenberg, T., & Fekete, J.-D. (2013). Weighted graph comparison techniques for brain connectivity analysis. *Proceedings of the SIGCHI conference on human factors in computing systems*.

Andrienko, G., Andrienko, N., Dykes, J., Fabrikant, S. I., & Wachowicz, M. (2008). *Geovisualization of dynamics, movement and change: Key issues and developing approaches in visualization research.* London, England: SAGE Publications Sage UK.

Andrienko, G., Andrienko, N., Fuchs, G., & Wood, J. (2016). Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE transactions on visualization and computer graphics*.

Boyandin, I., Bertini, E., Bak, P., & Lalanne, D. (2011). Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum, 30*(3), 971–980.

Buchin, K., Speckmann, B., & Verbeek, K. (2011). Flow map layout via spiral trees. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2536–2544.

Cui, W., Zhou, H., Huamin, Q., Wong, P. C., & Li, X. (2008). Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1277–1284.

Dwyer, T., Lee, B., Fisher, D., Quinn, K. I., Isenberg, P., Robertson, G., & North, C. (2009). A comparison of user-generated and automatic graph layouts. *IEEE Transactions on Visualization and Computer Graphics, 15*(6).

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2149–2158.

Ghoniem, M., Fekete, J.-D., & Castagliola, P. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. *INFOVIS 2004. IEEE Symposium on: . 2004*.

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics, 15*(6).

Guo, D., & Gahegan, M. (2006). Spatial ordering and encoding for geographic data mining and visualization. *Journal of Intelligent Information Systems, 27*(3), 243–266.

Guo, D., Jin, H., Gao, P., & Zhu, X. (2018). Detecting spatial community structure in movements. *International Journal of Geographical Information Science, 32*(7), 1326–1347.

Guo, D., & Zhu, X. (2014). Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 2043–2052.

Guo, D., Zhu, X., Jin, H., Gao, P., & Andris, C. (2012). Discovering spatial patterns in origin-destination mobility data. *Transactions in GIS, 16*(3), 411–429.

Härdle, W. (2012). *Smoothing techniques: With implementation in S.* Springer Science & Business Media.

Jenny, B., Stephen, D. M., Muehlenhaus, I., Marston, B. E., Sharma, R., Zhang, E., & Jenny, H. (2017). Force-directed layout of origin-destination flow maps. *International Journal of Geographical Information Science,* 1–20.

Jenny, B., Stephen, D. M., Muehlenhaus, I., Marston, B. E., Sharma, R., Zhang, E., & Jenny, H. (2018). Design principles for origin-destination flow maps. *Cartography and Geographic Information Science, 45*(1), 62–75.

Koylu, C. (2018). Discovering multi-scale community structures from the interpersonal communication network on twitter. In L. Perez, E.-K. Kim, & R. Sengupta (Eds.). *Agent-based models and complexity science in the age of geospatial big data: Selected papers from a workshop on agent-based models and complexity science (GIScience 2016)* (pp. 87–102). Cham: Springer International Publishing.

Koylu, C., & Guo, D. (2013). Smoothing locational measures in spatial interaction networks. *Computers, Environment and Urban Systems, 41*, 12–25. https://doi.org/10.1016/j.compenvurbsys.2013.03.001.

Koylu, C., & Guo, D. (2017). Design and evaluation of line symbolizations for origin–destination flow maps. *Information Visualization, 16*(4), 309–331. https://doi.org/10.1177/1473871616681375.

Kraak, M.-J. (2003). *The space-time cube revisited from a geovisualization perspective.* In: *Proc. 21st international cartographic conference*.

Kraak, M.-J., & Koussoulakou, A. (2005). A visualization environment for the space-time-cube. *Developments in spatial data handling* (pp. 189–200). Springer.

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems, 5*(3), 287–301.

NYC Limousine Commission (2018). Taxi trip data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and techniques in modern geography*.

Phan, D., Xiao, L., Yeh, R., & Hanrahan, P. (2005). Flow map layout. *INFOVIS 2005. IEEE Symposium on: . 2005*.

Purchase, H. C., Hamer, J., Nöllenburg, M., & Kobourov, S. G. (2012). On the usability of Lombardi graph drawings. *International symposium on graph drawing*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis. Vol. 26*. Chapman & Hall/CRC.

Tao, R., & Thill, J.-C. (2016). Spatial cluster detection in spatial flow data. *Geographical Analysis, 48*(4), 355–372.

Tobler, W. R. (1981). A model of geographical movement. *Geographical Analysis, 13*(1), 1–20.

Tobler, W. R. (1987). Experiments in migration mapping by computer. *The American Cartographer, 14*(2), 155–163.

Ware, C., Kelley, J. G. W., & Pilar, D. (2014). Improving the display of wind patterns and ocean currents. *Bulletin of the American Meteorological Society, 95*(10), 1573–1581.

Wood, J., Dykes, J., & Slingsby, A. (2010). Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal, 47*(2), 117–129.

Wood, J., Slingsby, A., & Dykes, J. (2011). Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization, 46*(4), 239–251.

Xu, K., Rooney, C., Passmore, P., Ham, D.-H., & Nguyen, P. H. (2012). A user study on curved edges in graph visualization. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2449–2456.

Yang, Y., Dwyer, T., Goodwin, S., & Marriott, K. (2017). Many-to-many geographically-embedded flow visualisation: An evaluation. *IEEE Transactions on Visualization and Computer Graphics, 23*(1), 411–420.

Yang, Y., Dwyer, T., Jenny, B., Marriott, K., Cordeil, M., & Chen, H. (2019). Origin-destination flow maps in immersive environments. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 693–703.

Zhu, X., & Guo, D. (2014). Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS, 18*(3), 421–435.