(Article)

A Method for Analysis and Visualization of Similar Hotspot Flow Patterns between Different Regional Groups

Haiping Zhang [1,2,3#], Xingxing Zhou [1,2,3#], Xin Gu [5], Genlin Ji [1,2,4] and Guoan Tang [1,2,3,*]

[1] Key laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University; Nanjing 210023, China;gissuifeng@163.com
[2] State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China;
[3] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China;
[4] College of Computer Science and Technology, Nanjing Normal University; Nanjing 210023, China
[5] Department of Geography Sciences, University of Maryland College Park, MD 20742, USA; xgu12347@terpmail.umd.edu
* Correspondence: tangguoan@njnu.edu.cn; Tel.: +8613 776623891

#The contribution to the article is the same

Abstract: The interaction between different regions normally is reflected by the form of the stream. For example, the interaction of the flow of people and flow of information between different regions can reflect the structure of cities' network, and also can reflect how the cities function and connect to each other. Since big data has become increasingly popular, it is much easier to acquire flow data for various types of individuals. Currently, it is a hot research topic to apply the regional interaction model, which is based on the summary level of individual flow data mining. So far, previous research on spatial interaction methods focused on point-to-point and area-to-area interaction patterns. However, there are a few scholars who study the hotspot interaction pattern between two regional groups with some predefined neighborhood relationship by starting with two regions. In this paper, a method for identifying a similar hotspot interaction pattern between two regional groups has been proposed, and the Geo-Information-Tupu methods are applied to visualize the interaction patterns. For an example of an empirical analysis, we discuss China's air traffic flow data, so this method can be used to find and analyze any hotspot interaction patterns between regional groups with adjoining relationships across China. Our research results indicate that this method is efficient in identifying hotspot interaction flow patterns between regional groups. Moreover, it can be applied to any analysis of flow space that is used to excavate regional group hotspot interaction patterns.

Keywords: regional group interaction; similar hotspot flow patterns; spatial interaction; visual analytics; Geo-Information-Tupo; GIS.

1. Introduction

Our society is built based on mobility, such as the flow of people, the flow of goods and the flow of information technology. And these elements of flow form a flow space[1]. Compared to traditional local space, the flow space pays more attention on the interaction of elements and their interaction relationship[2,3]. In the past, geographers were focus on physical space[4-6]. Nowadays, with the increasing development of economic globalization and Internet technology, geography researchers transfer their sights to flow space[7-10]. On one hand, the outstanding change of economic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

globalization is: people have strengthened their exchanges in tourism, trade and technology from all over the world, thus directly leading the advanced enhancement on the flow of people, logistics and technology; On the other hand, information flow has further weakened the distance between places with the development of internet technology. One apparently fact is that: the distance is no longer suitable to apply for the metric space when the time required to transmit information for 1 kilometer is almost the same as the time required to transmit information for 100,000 kilometers, that is to say, the connection of Internet has realized the change on transmission of spatial information. In fact, for geographers, the main points not only should be the flow space itself, but it is also about how these flow elements reconstruct the spatial organization structure, how to make the organization works, and what kind of flow patterns emerge[11]. Based on all these facts above, it is important for us to use the quantitative analysis method to excavate the interaction patterns and define the interaction patterns, because they are the basis methods to solve spatial relationships between two regional groups.

Over the past few decades, many methods have already been proposed to find out interaction patterns of flow space. In terms of spatial interaction model, some scholars have built many spatio-temporal interaction pattern mining algorithms from a summary sight[12-16]. However, the spatial dependence of interactive nodes among all these methods are lacked. Some of them apply complex network methods to discover the spatial interaction patterns[17-20]. The conception of interaction regions model based on the idea of complex networks has been proposed by some other scholars[21]. To some extent, the dependencies and similarities between flowing nodes are considered, especially the method of interaction relation which is proposed by Kira are able to identify areas with strong interactivity. But the limitation is that it only recognizes all the individual regions that are similar in interaction, rather than the interaction modes between different regions. Kwanho Kim et al. has proposed a regional mobile pattern recognition algorithm(MZP) based on the aggregation of metro nodes recently. Based on his idea[22], Chen et al. expanded the proximity relationship and realized the mobile pattern recognition based on taxi OD data, and MPFZ algorithm is proposed[23].All these methods are mainly focus on point data and its adjacency relationship, the main disadvantage is that their algorithm is inefficient and have not given a visually well-resolved solution to excavate interaction model. So in terms of visualizing the interaction patterns results, none of the above methods can solve the interaction pattern between two regional groups.

A basic characteristic of the existing models is: it lacks a more interactive flow pattern recognition method that may exist between one regional group and another, and it's hard to define the adjacency relationship between two regional groups. We can refer to the adjacency matrix related literature for the definition of regional adjacency relations. For example, there is a strong interaction between region A and region B (B does not have a predefined proximity relationship). Not only that, but also between several regions around A, and regions around B. We assume there is a strong interaction relationship, and a pre-defined adjacency relationship is satisfied between region A and its surroundings, and also a pre-defined adjacency relationship between region B and its surroundings, then regional group A and regional group B are located. So we can conclude that there is a strong interaction between A or B and surroundings, more importantly, a regional group interaction flow pattern is formed between regional group A and regional group B.

This paper presents an advanced method for discovering, analyzing and visualizing the interaction hotspot flow patterns between two different regional groups. During the next section, a review of related work has been written, and the expected results of the method will also be expressed. After the second section, a new method which is used to mine regional group flow patterns has also been proposed. During that part, we mainly include the definition of regional adjacent relationship, the structuration of flow pattern mining algorithm, the introduction of flow pattern visualization and methodological issues. In the end, an air flow volume data will be shown during the case study part by using the section three methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

2. Related work

In most cases, individual flow data will be modeled as a flow pattern from node to node [24, 25]. It is also for the above reasons that many of the macro mode summary or interactive mode discovery methods for individual flow data are based on node flow data [26-28]. In contrast, there is few flow data modeling and analysis between regions, moreover, the interactions between regions can also be abstracted as point-to-point interactions. It is easy to use basic spatial analysis methods to achieve the goal even if the flow data from point to point is aggregated to the region-to-region flow data. However, interaction modeling and analysis between regional groups will involve many issues such as how to identify and determine the regional adjacency relationship, and better visualize the expression. Most of the existing researches are based on the first two cases. Today we will have a brief introduction to the existing related research below. In order to better understand the limitations of the research objectives and the existing methods, we will also discuss point-to-point flow pattern, area-to-area flow pattern and flow pattern of two different regional groups, but we only mention the one having a strong relationship.

As we mentioned above, most of the flow data exists in the form of point-to-point with arrow. Related interaction analysis methods mainly include point-to-point interaction pattern mining [29-33], interactive pattern mining in between multiple points, and a model analysis of adjacent points in a same community [34-37].We can see clearly from Figure 1(a) that the interaction between the three nodes in the northwest corner and the two nodes in the southeast corner are significantly stronger than other flow data. A similar situation exists between the neighboring points in the southwest and northeast corner. As is shown on Figure 1(b), the MZP algorithm proposed by Publication et al. can discover a strong interaction pattern between a set of adjacent nodes in a network structure data to another set of adjacent nodes. Then, the two modes shown on Figure 1(c) could be identified. The MZP algorithm mainly represents a prefect method for solving such problems, and provides valuable reference value for related researches. However, the time complexity of this algorithm process is too high, and the visualization of the analysis results has not yet proposed a good solution. Based on that, Chen et al. proposed the MPFZ method, but Chen's method only extended the data which was applied by the network MZP algorithm from the network node flow to other analysis of the arbitrary node flow data. No other major changes have been improved in other areas..

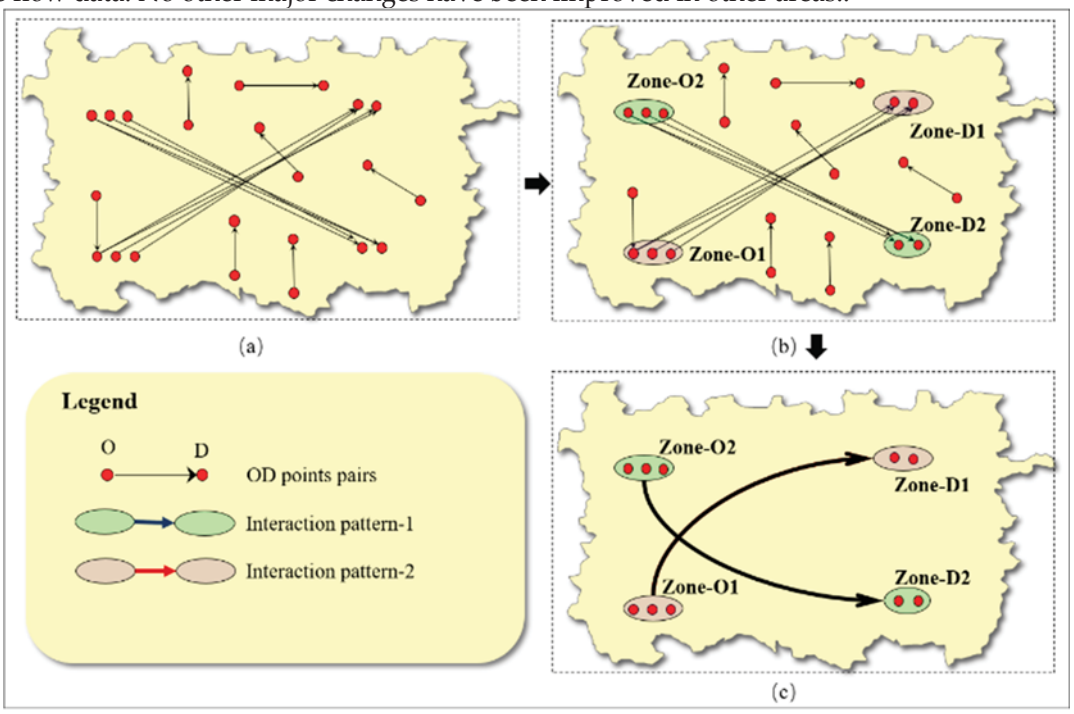


Figure 1. An example for point-to-point flow data and its analysis methods.(a)point-to-point flow data;(b)points-to-points flow data;(c)points-to-points flow patterns.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

We may pay more attention to the interaction pattern between different areas for flow data in some cases. For example, for the point-to-point with arrow data shown in Figure 1(a), we can easily obtain the area based area-to-area flow data through the basic spatial superposition and statistical analysis methods. As is shown in Figure 2(a), the arrow must also contain an attribute to indicate the size of the interaction value for each area-to-area flow data. Based on the results shown on Figure 2(a), we can easily identify the regional interaction shown on Figure 2(b), and thus obtaining the region interaction pattern shown on Figure. 2(d).

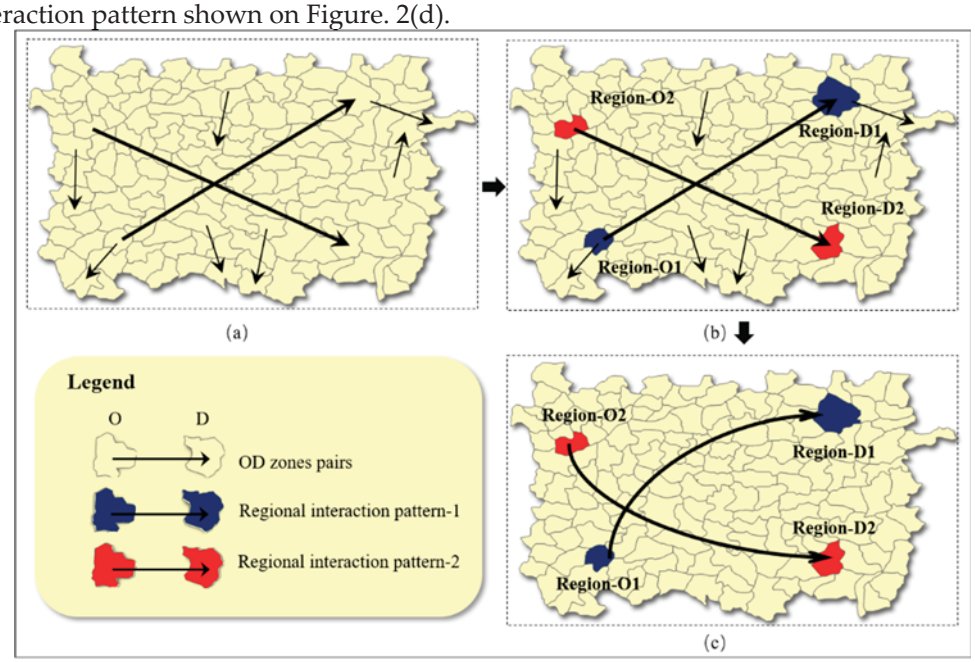


Figure 2. An example for region flow data and its analysis methods.(a)area-to-area flow data with high interaction values;(b);(c)area-to-area flow patterns.

Concerning area-to-area model , the obvious disadvantage is that each area interaction mode does not consider the correlation characteristics of the starting and ending area with other existing adjacent areas, which means the spatial autocorrelation of any area interaction mode and the surrounding area in interaction directions and sizes. As shown in Figure 3(a), it is much more significant that the interaction between several adjacent areas in the northwest and southeast among the area-to-area flow data. Also similar patterns are applied on the southwest and northeast sides. As shown in Figure 3(b), the goal of this paper is to identify the existence of regional group interaction flow patterns by defining specific area adjacency relationships, Figure 3(c) shows the results and visualization of the flow pattern that is expected. And then do further research on the interaction strength, value size and significant level of each regional group based on the results of the analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

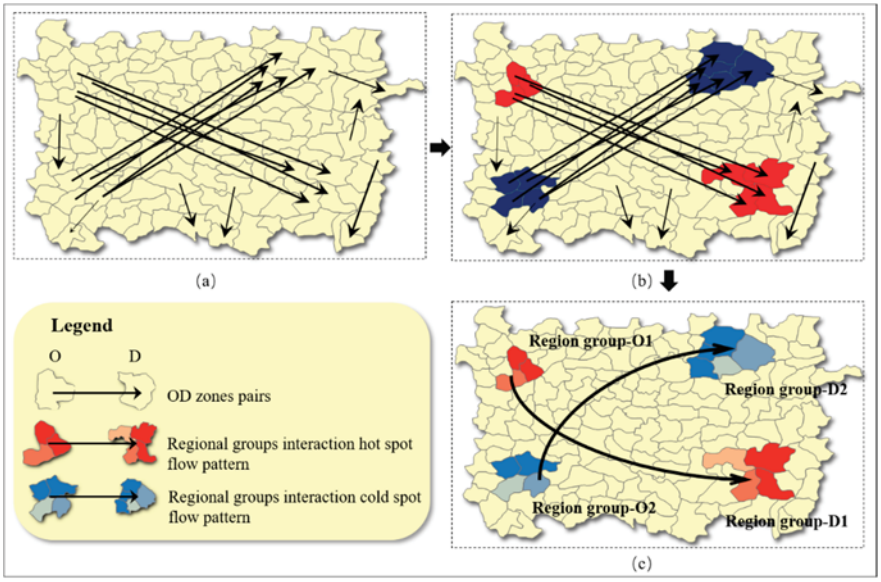


Figure 3. similar interaction hotspot flow pattern and it's visualizaiton among regional groups.(a)area-to-area flow data;(b)areas-to-areas flow data;(c)similar hotspot flow pattern between regional groups.

3. Methodology

The entire research framework includes the input of node-based flow data, data processing, and mining of regional group flow patterns, flow pattern output, and visualization. Since most of the flow data is counted and then stored by nodes, this study supports node-based flow data input during the design process. Firstly, the input node-to-node flow data is converted according to a certain regional unit and then converted into regional-to-regional flow data. This process can be realized by using the common GIS overlay and statistics functions. Then determine the adjacency relationship of the regional units (Section 3.1.1), and based on this adjacency relationship, merge the adjacent areas where the interaction value reaches a certain threshold before being constructed into regional groups. After that, we are able to identify all similar hotspot flow patterns among different regional groups (Section 3.1.2). In the end, the Geo-Information-Tupu visualization method is used to present the regional groups with similar hotspot flow patterns and visual variables are used to visualize the evaluation results of their own characteristics in each flow pattern. The rest of the writing, we refer to the similar hotspot flow pattern between regional groups as RG-Flow-Pattern.

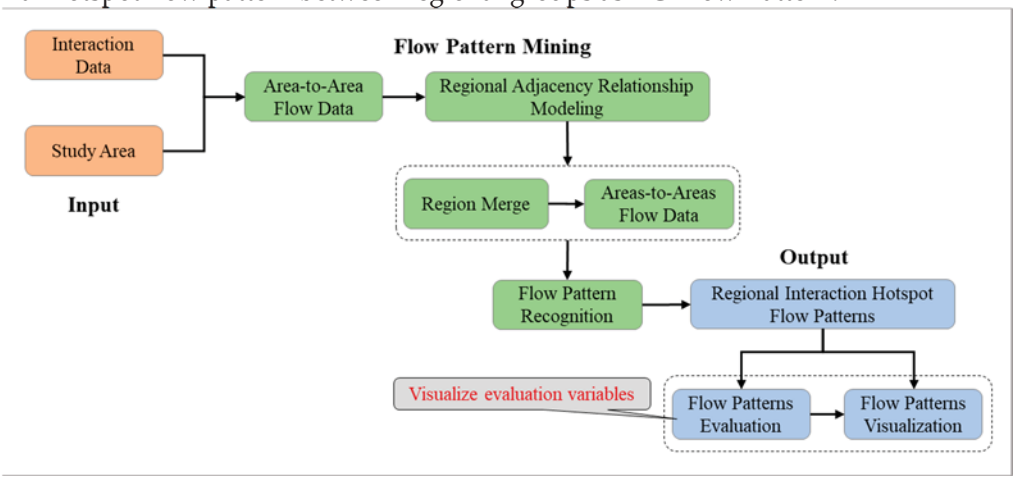


Figure 4 Overview of the framework for analysis and visualization the similar hotspot flow patterns between regional groups.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

3.1 Building algorithm for similar hotspot pattern between regional groups

In this study, the regional hotspot interaction model algorithm mainly includes three aspects. They are 1) Defining the regional neighborhood relationship. 2) Constructing a regional hotspot interaction pattern recognition algorithm based on the defined neighborhood relationships. 3) Multiple test parameters are used to evaluate the results of the identified area hotspot interaction models.

3.1.1 Regional adjacency relationship modeling

In order to identify the hotspot interaction pattern, we must clearly define the regional adjacency relationship and its merger principle. In this method, four ways are defined for determining the adjacency relationship of the area. As shown in figure 3, if each grid is used as a region, the adjacency relationship between regions can be expressed as the following four ways: figure3 (b), 3(c), 3(d) and 3(e). As shown in figure 3(a), if we assume the target area is the red one, the specific meanings of the four adjacency relationships are briefly described as below.

1. Adjacent edges

In figure 3(b), there are four areas have common edges with target area, and these four areas are defined as adjacent areas of the target area. The adjoining relationship in this manner is called an edge-adjacent relationship. In an actual partition, under this rule, a target area may have more adjacent areas or less than four adjacent areas.

2. Adjacent edges and corners

Figure 3(c) shows a similar adjacency relationship to figure 3(b). However, except that the area having a common edge with the target area belongs to the adjacent edge of the target area, it also includes an area having a common node with the target area. This kind of adjoining relationship is called edge-corner adjoining.

3. Customized adjacent range

In figure 3(d), a circular buffer area is defined with the center of mass of the target area as the origin. When other areas are within or intersect the buffer area, they are defined as the adjacent areas of the target area. In this method, the adjacency relationship is called the adjoining relationship of customized adjacent range.

4. Logical adjacent relationship

In addition to the above three methods to define the adjacency relationship, we can also determine whether the target area and the other areas are adjacent by customizing the logical relationship that is independent of the spatial position. In figure 3(e), there are some logical relations between the three blue areas and the target area. Therefore, even though these areas do not coincide with the target area or coincide with the vertices, these three areas are defined as the adjacent areas of the target area.

Basically, the above four are the typical modeling methods for the spatial relationship of surface features. Other adjacencies include k-nearest, custom based on spatial adjacency matrix, etc.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

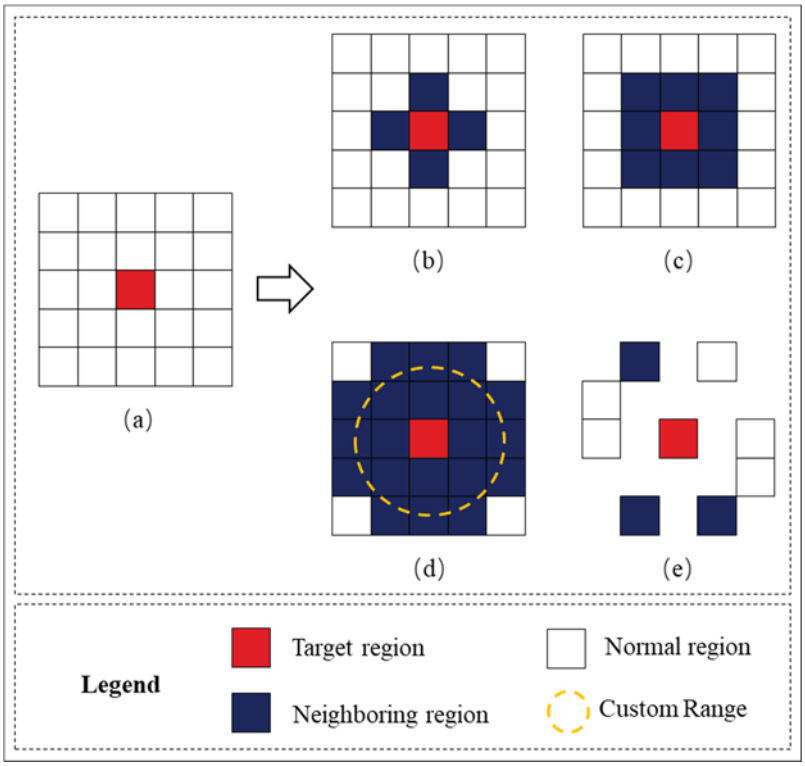


Figure 4. similar interaction hotspot flow pattern and it's visualizaiton among regional groups.(a);(b);(c);(d);(e).

3.1.2 Region merge and similar hotspot flow pattern recognition

1. Definitions of similar hotspot flow pattern between regional groups

In this research, a set of data sets containing n planar area units were given, Rset = $\{R_1, R_2, \dots, R_n\}$(i = 1,2, … , n), $R_i$represents the nth region. In a regional group interactive hotspot flow pattern, the origin area group is defined as $RGOset = \{R_1, R_2, \dots, R_u\}$, and the destination area group is defined as RGDset = $\{R_1, R_2, \dots, R_v\}$. In addition, a pair of origin areas and destination areas that have interactions in a regional group interaction hotspot pattern are called regional flow. A data set RFset was given to store all the regional flow in Regional group interaction hotspot flow pattern,（RIH-FP）, $RFset = \{RF_1, RF_2, \dots, RF_m\}(j = 1,2, \dots, m)$, The jth regional flow can be represented as $RF_j = RF_{j\_o} \rightarrow RF_{j\_d}$, $RF_{j\_o} \epsilon RGOset$,it indicates the origin area of the regional flow. $RF_{j\_d} \epsilon RGDset$,it represents the destination area of this regional flow. In some situations, for ease of exposition, we use the term flow pattern instead of RIH-FP in the remainder of the paper.There are some definitions about the flow pattern.

Definition1: A regional group interactive hotspot flow model consists of three parts. They are the starting regional group RGOset, the destination regional group RGDset, and the interaction direction indicating the interaction relationship. A regional group hotspot interaction pattern has the same direction as any $RF_j = RF_{j\_o} \rightarrow RF_{j\_d}$ in the RFset.

Definition2: Given an area-adjacent relationship defined in 3.1.1, a single region in the RGOset of the origin region group must satisfy such a adjacent relationship, and a single region in the destination region group RGDset must also satisfy this adjacent relationship.

Definition3: The number of regions in the origin and destination regional group of a regional group interactive hot spot flow mode cannot be 1 at the same time, that is, at least more than one region is included in the start or termination regional group.

Definition4: The interaction value of the regional flow refers to the interaction value from one region to another, which is represented by InterVal. This value has different meanings in different applications, but the following conditions must be required:

Given a threshold θ, the interaction strength value P(RF_j) of the j-th regional stream RF_j must satisfy the following conditions:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

$$\mathrm{P}(RF_j) = \frac{InterVal(RF_{j\_o} \rightarrow RF_{j\_d})}{InterVal(RF_{j\_o} \rightarrow RF_{*\_d}) * InterVal(RF_{*\_o} \rightarrow RF_{j\_d})} (\mathrm{P}(RF_j) \geq \theta) \quad (1)$$

$InterVal(RF_{j_o} \rightarrow RF_{j_d})$ represents the interaction value which is from origin area RF_{j_o} to destination area RF_{j_d}. $InterVal(RF_{j\_o} \rightarrow RF_{*\_d})$ represents the sum of the interaction values of the origin region RF_{j_o} to all other destination regions. $InterVal(RF_{*\_o} \rightarrow RF_{j\_d})$ represents the sum of the interaction values of all the origin regions to the destination region RF_{j_d}.

Definition5: The RFset, which contains all regional flow in the same flow pattern, is no pre-defined adjacent relationship from the starting region(s) to the ending region(s) in any regional flow RF.

2. Region merge

Firstly, we randomly select a group of regional flow data that satisfy: $\mathrm{P}(RF_j) \geq \theta$, $RF_j = RF_{j\_o} \rightarrow RF_{j\_d}$, and make $RF_j = RF_{j\_o} \rightarrow RF_{j\_d}$ as the first region flow of a new regional interactive hotspot flow pattern, and the interaction value size is expressed as Inter Val$(RF_{j_o} \rightarrow RF_{j_d})$. Then using RF_{j_o} as the starting regional group elements of the new region interactive hotspot flow mode, and satisfy $RF_{j\_o} \epsilon RGOset$. Using RF_{j_d} as the new regional interactive hotspot flow mode, which is the termination elements of regional group, it should satisfy $RF_{j\_d} \epsilon RGDset$. Search for all regions adjacent to RF_{j_o}, whose set is defined as ARGOset $= \{RF_{j\_o\_1}, R_{j\_o\_2}, \dots, R_{j\_o\_u}\}$(m = 1,2, … , u), the mth adjacent region of RF_{j_o} is $RF_{j\_o\_m}$; all regions adjacent to RF_{j_d} are searched in the same way, and the set is defined as ARGDset $= \{RF_{j\_d\_1}, R_{j\_d\_2}, \dots, R_{j\_d\_v}\}$(n = 1,2, … , v),. The nth adjacent of RF_{j_d} is $RF_{j\_d\_n}$. For $RF_{j\_o\_m}$ in any ARGOset, if $RF_{j\_o\_m}$ interacts with the area $RF_{j\_d\_n}$ in the ARGDset, it constitutes the regional flow $RF_{j\_m\_n} = RF_{j_{o_m}} \rightarrow RF_{j\_d\_n}$, then:

$$\mathrm{P(RF)} = \frac{InterVal(RF_{j\_o} \rightarrow RF_{j\_d}) + InterVal(RF_{j\_o\_m} \rightarrow RF_{j\_d\_n})}{(InterVal(RF_{j\_o} \rightarrow RF_{*\_d}) + InterVal(RF_{j\_o\_m} \rightarrow RF_{*\_d\_n})) * (InterVal(RF_{*\_o} \rightarrow RF_{j\_d}) + InterVal(RF_{*\_o\_m} \rightarrow RF_{j\_d\_n}))} \quad (2)$$

Among them, InterVal$(RF_{j_o} \rightarrow RF_{j_d})$ is the interaction value of the regional flow RF_j, InterVal$(RF_{j_o} \rightarrow RF_{*_d})$ indicates the sum of the interaction values of the starting area RF_{j_o} to all other termination areas RF_{*_d}, InterVal$(RF_{*_o} \rightarrow RF_{j_d})$ represents the sum of the interaction values of all other starting regions RF_{*_o} to the ending region RF_{j_d}). Similarly, InterVal $InterVal(RF_{j\_o\_m} \rightarrow RF_{j\_d\_n})$ is the interaction value of the regional flow $RF_{j\_m\_n}$, $InterVal(RF_{j\_o\_m} \rightarrow RF_{*\_d\_n})$ indicates the sum of the interaction values of the starting area $RF_{j\_o\_m}$ to all other areas. $InterVal(RF_{*\_o\_m} \rightarrow RF_{j\_d\_n})$ indicates all other areas to the interaction value of $RF_{j\_d\_n}$.

After calculating the P(RF) value, if $\mathrm{P}(RF) \geq \theta$, then $RF_{j\_o\_m}$ is also included at the origin regional group of the regional group interaction mode, $RF_{j\_o\_m} \epsilon RGOset$ is satisfied, The $RF_{j\_d\_n}$ is included in the termination zone group of the regional group interaction mode, $RF_{j_d_n} \epsilon$RGDset is satisfied. After all the above is completed, statistical analysis is performed on other adjacent areas by the same method, and it is known that an area does not meet the merge threshold and the merge operation is ended. The newly included start and end regions are then searched for their adjacent regions, and the above operations are iterated until no region satisfies the merge threshold. Finally, a complete regional interaction hotspot flow mode start zone group and termination zone group are obtained.

3. Regional interaction hotspot flow pattern recognition

Through the merging of the upper part of the region, a starting zone group and an ending zone group of several regional interactive hotspot modes are formed. For an area interaction hotspot flow pattern RIH-FP if the set of start area groups is defined as: $RGOset = \{R_1, R_2, \dots, R_u\}$, the ending regional group is defined as $RGDset = \{R_1, R_2, \dots, R_v\}$, the set of regional flow is defined as $RFset = \{RF_1, RF_2, \dots, RF_m\}(j = 1,2, \dots, m)$. RF_p represents the p-th region flow, and RF_q represents the q-th region flow. The initial regional group RGOset, the termination area group RGDset, and the interaction stream set RFset between the two regional groups constitute a complete regional hotspot

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

interaction flow mode. The direction of interaction between the regional groups is indicated by the directional arrows. Thus, the start region group, the termination region group, and the direction arrow constitute a basic visualization element of an area hotspot interaction flow pattern and form a feature structure of the flow pattern. Based on a complete regional interaction hotspot flow pattern, in addition to the visual elements and feature structure, some evaluation values are needed to distinguish the strength of each flow pattern. If the variable P is used to indicate the strength of a certain RIH-FP, then:

$$P = \sum_{j=1}^{m} P(RF_j) \tag{3}$$

$P(RF_j)$ represents the interaction strength value of the jth regional flow in the regional flow set RFset. The interaction strength of the entire RIH-FP is the sum of the values of all the regional flow interaction strengths in the RFset.

If the V denote the size of the interaction value of a certain RIH-FP, then V should satisfy the following formula:

$$V = \sum_{j=1}^{m} Interval(RF_j) \tag{4}$$

$Interval(RF_j)$ represents the interaction value of the jth region flow in the regional flow set RFset. The interaction value of the entire RIH-FP is the sum of all the regional flow interaction value in the RFset.

In addition, it is also necessary to separately calibrate the contribution of each of the start regional group and the termination regional group to the current flow mode interaction value in a complete mode. For the i-th region R_j in the starting regional group RGOset:

$$DO(R_j) = \frac{\sum_{j=1}^{m} Interval(RF_j)}{InterVal(RF_{j\_0} \rightarrow RF_{*\_d})} (R_j \in \text{RGOset}, RF_j \in \text{RFset}, RF_{j\_o} = R_i) \tag{5}$$

For the i-th region R_j in the termination regional group RGDset:

$$DD(R_j) = \frac{\sum_{j=1}^{m} Interval(RF_j)}{InterVal(RF_{*\_o} \rightarrow RF_{j\_d})} (R_j \in \text{RGOset}, RF_j \in \text{RFset}, RF_{j\_d} = R_i) \tag{6}$$

3.2 SHFP-RG visualization method based on Geo-information Tupo theory

3.2.1 Visualization of single RG-Flow-Pattern

In the RG-Flow-Pattern method of this paper, the analysis results are evaluated and investigated by using different flow pattern variables. These variables have both an assessment of the starting and ending regional groups as well as an overall assessment of the interaction model. Viewing these evaluation variables that match a particular pattern in a table is not desirable for spatial pattern analysis. It also loses the advantage of visualizing spatial data analysis results based on maps and further visual analysis. Therefore, it is very important to design a scientific and reasonable RG-Flow-Pattern visualization method. So based on the above facts, the RG-Flow-Pattern visualization method is designed as shown in Figure6 (a) and (b). Figure6 (a) and (b) are two basic examples of RG-Flow-Pattern visualization. The basic meanings and expression purposes of the two model examples are described in detail below.

As we mentioned earlier, a complete RG-Flow-Pattern contains three basic constructs, namely the start regional group, the termination regional group, and the directional arrows. In order to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

visualize the results of each RG-Flow-Pattern, the interaction value size, and the contribution rate of each RG-Flow-Pattern in each of the start and termination regional group, some Visual variables such as color and size are expressed. As shown in Figure. 6(a) and Figure.6(b), if one proceeds from the basic definition, it is obvious that the basic requirements of the RG-Flow-Pattern structure are satisfied.

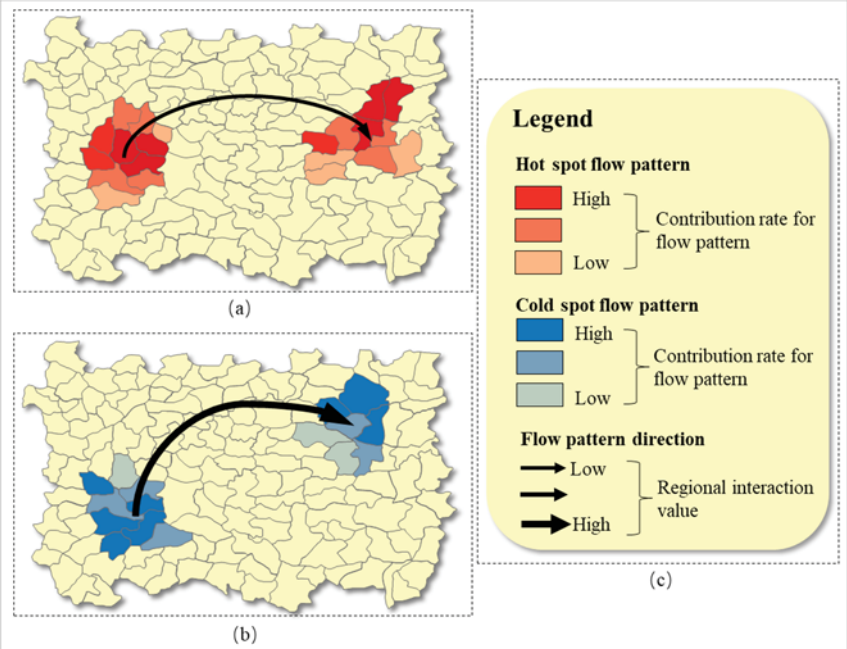


Figure 6. Two simple example for single RG-Flow-Pattern visualization and instrument its meaning.(a) A regional interaction hot spot flow pattern with low interaction value. (b)A regional interaction cold spot flow pattern with high interaction value.

Comparing the two findings, there are significant differences in the overall color design of the regional group. Figure 6(a) shows a warm tone, while Figure 6(b) shows a cool tone. The purpose of this design is to express the strength of each RG-Flow-Pattern by means of cool and warm colors. The warm tone indicates that the RG-Flow-Pattern behaves in a strong interactive mode, and the cool color represents the performance of the RG-Flow-Pattern behaves in a weak interactive mode. The degree of strength is measured by the P value in equation (2). The critical value of strength is divided according to the overall distribution of P values of all models by using natural discontinuity method, quantile method, etc., and the user of the model can definite it by themselves. Obviously in the two examples given in this paper, Figure 6(a) belongs to the strong regional interaction flow mode, where the hot spot flow mode is further defined. Figure 6(b) belongs to the weaker interactive flow mode, which is further defined as the cold spot flow mode. In addition to the differences in the cool and warm tones of the regional groups as a whole, there are also differences among the inner regions of each RG-Flow-Pattern. This represents the contribution rate of a single region to the current RG-Flow-Pattern interaction value. The darker the color, the greater the contribution rate of the region to the RG-Flow-Pattern interaction value, and vice versa. The contribution rate is measured by Equation (4) and Equation (5). The former measures each mode. The contribution of a single zone in the starting regional group, which is used to measure the contribution rate of a single region in the termination regional group for each flow mode. This rule can be applied to both hot flow pattern and cold spot flow pattern. The first two parts of the legend shown in Figure 6(c) illustrate the specific meanings and corresponding relationships between the expression flow pattern strength and the contribution rate of interaction values in each region in the visualization results.

In addition, RG-Flow-Pattern also needs to evaluate the value of the overall model interaction value through the value of V, so as to make up for the inadequacy of the interaction value that can be used to evaluate the strength of the interaction model. In the visualization, the size of the V value is expressed by the thickness of the arrow, which indicates the current RG-Flow-Pattern interaction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

value. Comparing Figure. 6(a) and Figure. 6(b), although RG-Flow-Pattern in Figure. 6(a) shows a strong flow pattern, the interaction value is smaller than that in Figure. 6(b). The flow pattern direction portion of Figure. 6(c) is a legend of the interaction value size relationship.

We can conclude that in addition to the directional arrows including the starting regional group and the ending regional group, the group of cooling and heating tone variables representing the strong and weak P value of the interaction mode, a saturation vision variable of a single region contribution rate V value to the current mode interaction value, and an arrow size vision variable representing the size of the flow pattern interaction value are also included in a complete visualization result of RG-Flow-Pattern.

3.2.2 Visualization and classification of multiple RG-Flow-Patterns based on Geo-information Tupu

In the traditional spatial data distribution and visualization mode, the distribution pattern of the same topic and region can be presented on a map. For example, the classic analysis method Local moran's I and General G index for analyzing the local spatial autocorrelation, the analysis of the models are easy to present on the same map. However, it is difficult to present on the same map for the regional group interactive hotspot flow mode. As shown in Figure. 6, pattern-01 and pattern-02 belong to two different flow patterns in the same region, but both patterns have a single repeating unit in both the real regional group and the termination regional group, which means, it is difficult for such situations to express two modes on the same map.

In the 90 decade of the 20th century, the theory and method of Geo-Information-Tupu put forward by Chen can be used to solve this problem[38]. In Chen's Geo-Information-Tupu theory, it emphasizes the structuring, abstraction, type, and relevance features of geographic laws, and uses these principles in a map sequence. Since in many cases, it is difficult to present multiple RG-Flow-Patterns in the same map, and different RG-Flow-Patterns of the same topic can also be type-divided, the map sequence can be adopted by the Geo-information-Tupu method. The RG-Flow-Pattern map sequence can be arranged according to types, and can also be arranged according to interaction strength, interaction value size. Since the interaction strength and interaction values can be directly organized by P value and Z value, thus only the type division of the RG-Flow-Pattern map is introduced in this paper.

In fact, for RG-Flow-Patterns, the type division is also a relatively simple task. In this paper, RG-Flow-Patterns is divided into two basic types and complex types. The basic types mainly include the five types shown in figure 8.

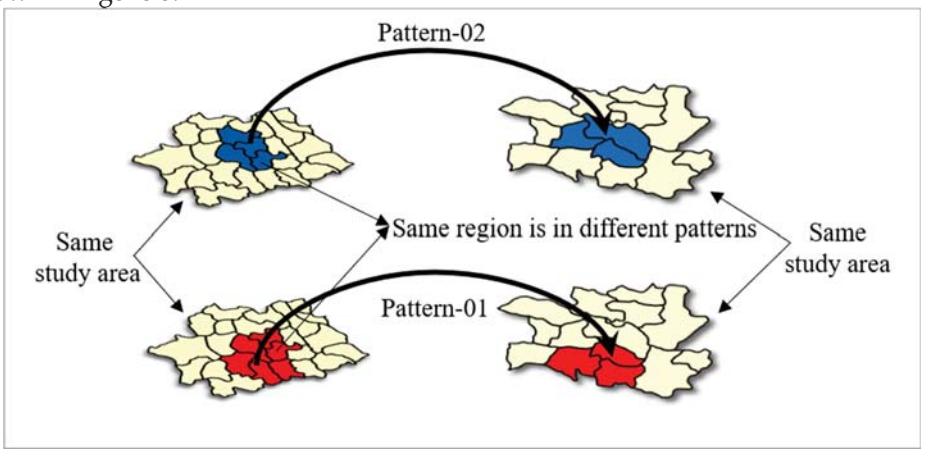


Figure 7. An example of the same region belong to different patterns

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

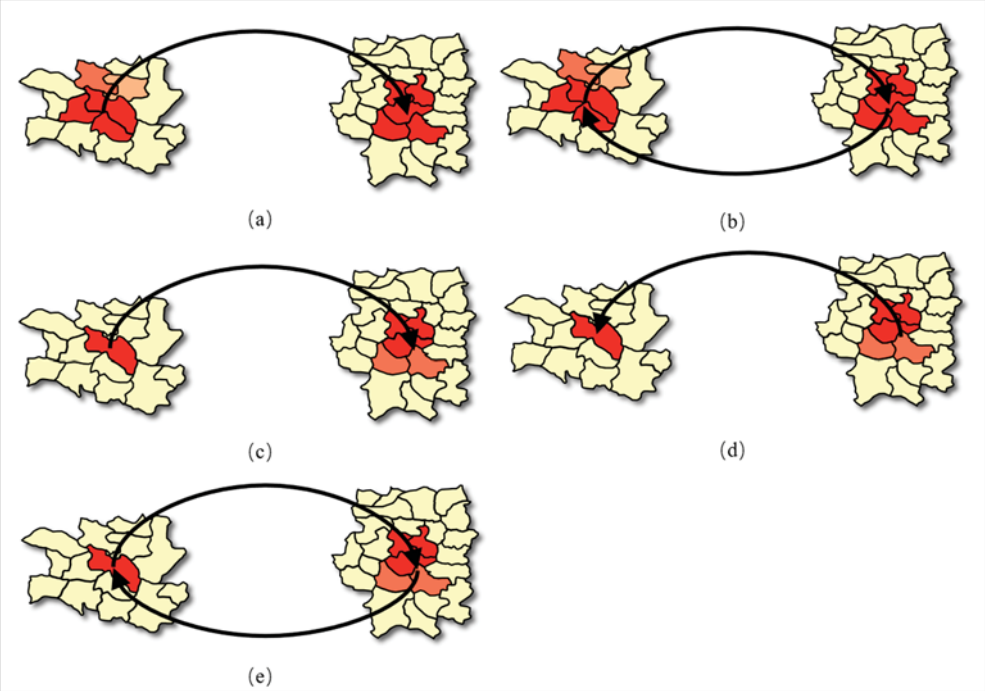


Figure 8. Basic categories of RG-Flow-Pattern based on Geo-information-Tupu.(a) many-to-many regions and single direction RG-Flow-Pattern.(2) many-to-many regions and double directions RG-Flow-Pattern.(c) one-to-many single direction RG-Flow-Pattern. (d) many-to-one single direction RG-FP.(e) one and many double direction RG-Flow-Pattern.

4. Case Study：the national scale migration flow data of China

4.1 Study area and data descriptions

Due to work, leisure travel and other purposes, a large number of people travel from one place to another every day. Human mobility can reflect lots of issues, such as urban attractiveness, tourism resources and so on. China has a population of 1.3 billion and there are significant differences in economic, political, cultural and resource characteristics in different regions. The huge imbalance in population size and regional disparities further promotes population movements. In terms of transportation, China's national-wide cross-regional transportation includes three types of transportation: automobiles, trains, and aircraft. A car is more suitable for short trips, the train is more suitable for people with short-to-medium-distance or low- and middle-income groups, while the airplane is mainly for long-distance or high-income travel. Because the method proposed in this paper is more effective in the analysis of flow data across regions, this paper uses the migratory flow data of the Chinese mainland as the main data source, and the prefecture-level city as the smallest research unit. We adopt the RG-Flow-Pattern method to develop the empirical analysis. Figure 8 shows the distribution of the population migration routes (by airplane) for the main study area on April 1, 2017. It should be noted that only the top ten data inflows and relocations from each prefecture-level city are used here.

The demographic data provided by the Tencent location big data platform was used in this research. Tencent is a major Internet company in China that provides nationwide location-based real-time migration big data services. On this platform, daily migration data from mainland China are provided. The migration types include aircraft, trains, and automobiles. Also, the top ten regions by rank of flow data was included, and the degree of hotspot flow value of moving in and out was calculated. Among the three modes of transportation migration data, the flight data has the longest distance, and the RG-Flow-Pattern method is better for this analysis. Therefore, the population migration data of flights was analyzed in this paper. The data used in this study involves 315 cities,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

and the total number of data points for all of the cities is approximately 6300, including flow data with original city, destination city, and hot value as the main attributes.

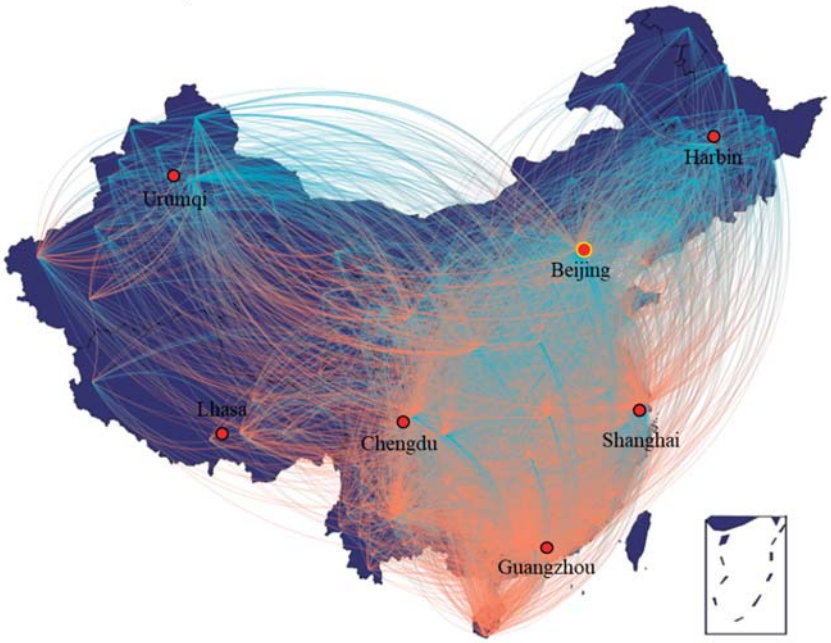
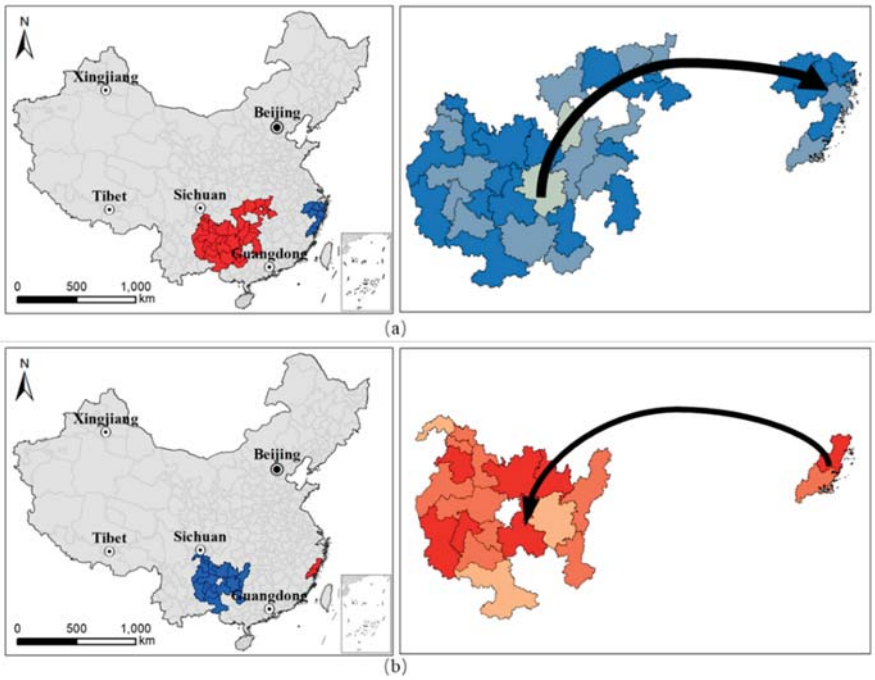


Figure 9. Study area and visualization of flow data

4.2 Result

The RG-Flow-Pattern method proposed in this paper was adopted, and set the prefecture-level city was a regional unit and the modal method of spatial relationship shown in Figure 5(c) was used, then set the θ value of $P(RF_j) \geq \theta$ to 0.00001. The partial patterns in the analysis result are shown in figure 10 below:



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

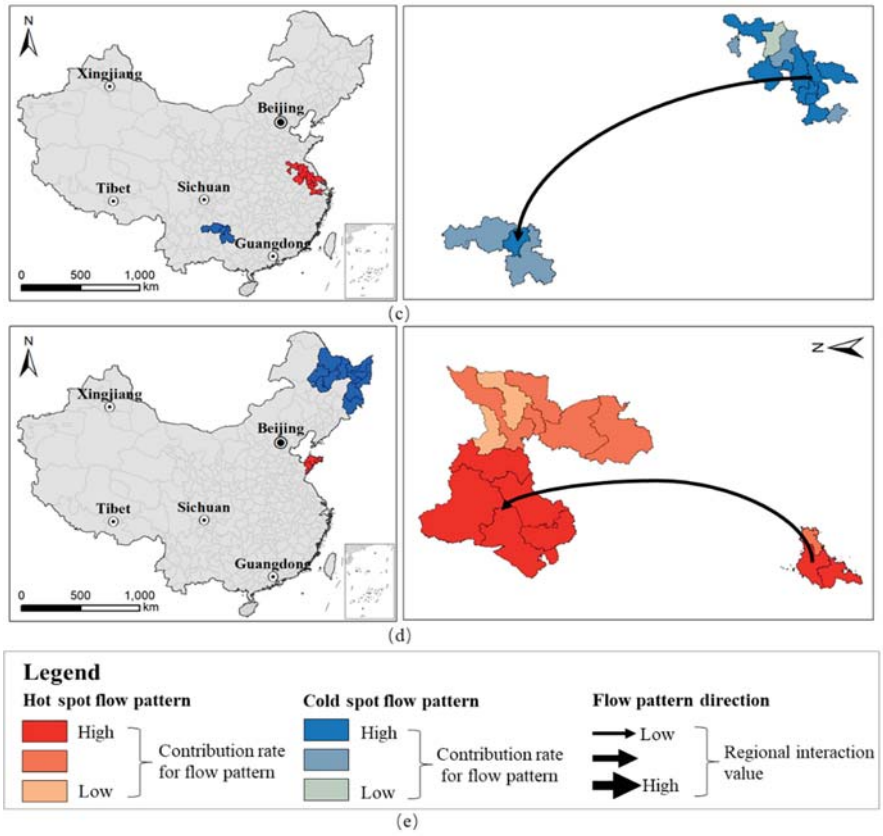


Figure 10. Four examples of RG-Flow-Patterns Geo-Infomaion-Tupu by threshold of 0.00001.(a) a coldspot RG-Flow-Pattern.(b) a hotspot RG-Flow-Pattern. (c) a cold spot RG-Flow-Pattern. (d) a hotspot RG-Flow-Pattern. (e) legend for RG-Flow-Patterns.

As shown in Figure 10(a), through the RG-Flow-Pattern algorithm analysis, it found that some regions in the southwestern part of China (the red part) and the eastern part of the coastal area (the blue part) form the regional group interaction flow model. Figure 10(a) shows the geographical distribution of the flow pattern on the left, and figure 10(a) shows the pattern representation of this pattern on the right. As can be seen from the latter, the pattern belongs to the cold spot flow pattern, and the direction of flow pattern is from the southwest area to the eastern coastal area. The color of a single area represents the contribution of the flow of that area to the entire pattern. The southwest area used as the starting regional group of the flow pattern, in which the color depth of each area represents the contribution of the sum of the values of the area flowing out to the termination area group to the outflow value (also called the outdegree) of the entire model; The coastal area is the most frequent end-of-flow model, in which the color depth of each area indicates the contribution rate of the inflow value of the area to the inflow value of the entire model. The darker the color, the greater the contribution rate. Refer to the legend shown in Figure 10(e) for details.

Figure 10(b) and figure 10(d) are interactive cold spot flow patterns recognized by the RG-Flow-Pattern algorithm. Figure. 10(c) is another set of identified regional group interaction hotspot flow patterns.

5. Discussion and conclusions

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2018 doi:10.20944/preprints201807.0063.v1

5.1 Discussion

5.1.1 Selection pricinple of region ajacency relationship and region merge threshold

In this case, the adjacent edges and corners approach is used for the adjacency of the area, which means that this approach is considered as the adjoining area of the target area as long as there is an edge or corner adjacent to the target area. When we model the area's adjacency, other methods mentioned in 3.2.1 section can be chosen. However, based on RG-Flow-Pattern analysis, using different regional adjacency relationships, there may also be differences in the models. This is the impact of regional adjoining relationships on the model. Among the specific issues, it is recommended to refer to the selection principles of regional spatial relationships in spatial statistical methods such as Moran's I, the Geary index, and Geographically Weighted Regression (GWR). Another problem is that when the value of θ in $P(RF_j) \geq \theta$ is different, the resulting flow pattern may also be different. The larger the value of θ, the smaller the number of flow patterns to be formed. The number of areas in the flow pattern that make up the start area group and the termination area group also decreases. To solve this problem, the recommended practice is to first obtain the $P(RF_j) \geq \theta$ values for all regional flows, and then use the bar histogram to evaluate the distribution of all regional flow $P(RF_j) \geq \theta$ values and select them according to the analysis target. A reasonable threshold is taken as the value of θ. This method can control the number and strength of flow patterns to a certain extent. So in this case study, figure 11 shows the plot distribution of P values for all regional flows.

5.1.2 result evaluation

In a complete flow pattern, both the basic elements of the flow pattern (starting regional group, termination regional group, and interaction arrows) are included, as well as the interaction strength, interaction value size, each individual flow pattern and the rate of contribution of the area's traffic to the interaction value of the entire flow pattern. Although this design makes it possible for each flow pattern to contain enough information to evaluation itself, the disadvantages are also obvious. First of all, these assessments are for a single flow model and lack the assessment of the overall characteristics of all models. For a single flow mode, starting from the strength of the mode and the size of the interaction value, there are four situations: firstly, a strong interaction mode with a large interaction value; secondly, a weak interaction mode with a small interaction value; thirdly, a strong interaction mode with a small interaction value; and fourthly, a weak interaction mode with a large interaction value. For all the overall characteristics of the model, it is obviously very useful for subsequent analysis to understand these four scenarios. If the strength and interaction values of each flow pattern can be described by XY coordinate system, the four cases can be expressed clearly and transparently through the four-quadrant diagram.

5.1.3 shortcomings and future improvements

The RG-Flow-Pattern method realizes that all flow patterns satisfying a certain intensity are recognized from the mass flow data, and a plurality of visual variables are used to better express the pattern and the related evaluation amount. However, there still exist some deficiencies. First of all, although the goal of this method is to analyze any type of flow data such as people flow, logistics, and traffic flow, for some flow data with a shorter interaction distance, it is difficult to find two cross-regional regional groups by this algorithm. This means that this method is more suitable for the mining of regional group interaction patterns between regions with long interaction distances. Although one can solve this problem by setting smaller partitions, more often than not, the interactive areas used for analysis are predefined and show some geographic significance, and cannot be customized for their size. In subsequent studies, we will try to build a flow data model mining model that is suitable for short interaction distances based on this method. Secondly, in a complete regional group interaction flow model, a strong self-interactive mode may exist between a single region of a starting regional group and a single region of an ending regional group, and the RG-Flow-Pattern method cannot recognize their self-interactive mode in this case Although it is not considered in the

 doi:10.20944/preprints201807.0063.v1

RG-Flow-Pattern method, this self-interactive pattern mining method is relatively simple. Its main challenge is how this kind of self-interactive mode plays a role in the flow model of this article and how it can be improved. The expression is performed in a visual manner to facilitate subsequent visual analysis. These are the tasks that need to be further improved.

5.2 Conclusion

With globalization and the development of the Internet, geographers have turned their attention from physical space to flow space. Spatial analysis methods also have been extended from spatial pattern analysis to spatial interaction pattern discovery. Although spatial interaction has always been the focus of the GIS field, with the advent of big data technologies, spatial interactions and even space-time interactions have successfully attract the attention of scholars nowadays. Many researchers mainly focus on point-to-point, area-to-area, or interaction-based research on regional convergence or diffusion. Few people consider the interaction patterns that may exist between regional groups that have some sort of adjoining relationship. In fact, the interaction of most flow data does not only exist between two separate areas, but the interaction always happens between a group of areas and another regional group.

We assume that two different regions, the relationship of one area to another is formed since an imbalance in certain resources. Furthermore, since this kind of imbalance, the surrounding area of one certain region has similarity demand of this resource, so it leads the target area and its surroundings with limited sources (we call it regional groups) interacts with other regional groups having abundant resources. The area and the surrounding areas that also have such resources interact with the surrounding areas that require such resources but lack them, forming an interaction between the two regional groups. In this paper, the RG-Flow-Pattern analysis and visualization method proposed can effectively mine the possible interaction patterns between two regional groups under such scenarios. In this analysis method, not only can all the regional groups having such interaction relationships which satisfy a specific traffic threshold be identified, but also the level of the strength of each group of interaction flow modes, the size of the interaction of the modes, and each of the interaction variables and the extent to which the area contributes to the overall model interaction volume can be measured by some outcome variables.

The first law of geography is the basic principle of the GIS spatial analysis model, that is, the spatial unit has spatial correlation characteristics. In the past, spatially-distributed characteristics tent to be considered in analytical models in spatial distribution models and spatial relationship modeling. Concomitant with the "interactive" turn of the GIS analysis model, and under the perspective of flow space, the spatial flow model or spatial interaction model should also be considered as spatial correlation. However, describing the spatial flow model is more complex than the spatial distribution model and the spatial relationship modeling, and it is difficult to visualize all the patterns through a single map. In this paper, based on the consideration of the relevance of neighboring regional units, we proposed a spatial group interaction model analysis method, and at the same time, geo-information maps was used to express the analysis results model and to deal with the difficulty of single diagram visualization. This analysis method can be extended to mine regional data interaction relationship in any other flow data forms.

6. Patents

Acknowledgments: This work is supported by National Natural Science Foundation of China under Grants No. 41471371, and supported by National Natural Science Foundation of China, No.41671389 We would like to express our sincere appreciation to the anonymous reviewers for their insightful comments that have greatly aided us in improving the quality of this paper.

References

1. Marty, P.F. An introduction to digital convergence: Libraries, archives, and museums in the information age. *Libr Quart* **2010**, *80*, 1-5.

2. Andris, C.; Liu, X.; Ferreira, J. Challenges for social flows. *Computers, Environment and Urban Systems* **2018**, *70*, 197-207.
3. Andris, C. Integrating social network data into gisystems. *Int J Geogr Inf Sci* **2016**, *30*, 2009-2031.
4. Midler, J.C. Non-euclidean geographic spaces: Mapping functional distances. *Geogr Anal* **2010**, *14*, 189-203.
5. Alamri, S.; Taniar, D.; Safar, M.; Al-Khalidi, H. A connectivity index for moving objects in an indoor cellular space. *Pers Ubiquit Comput* **2014**, *18*, 287-301.
6. Wang, J.F.; Li, X.H.; Christakos, G.; Liao, Y.L.; Zhang, T.; Gu, X.; Zheng, X.Y. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the heshun region, china. *Int J Geogr Inf Sci* **2010**, *24*, 107-127.
7. Limtanakool, N.; Schwanen, T.; Dijst, M. Developments in the dutch urban system on the basis of flows. *Reg Stud* **2009**, *43*, 179-196.
8. McKenzie, G.; Janowicz, K.; Gao, S.; Gong, L. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Comput Environ Urban* **2015**, *54*, 336-346.
9. Tao, R.; Thill, J.C. Spatial cluster detection in spatial flow data. *Geogr Anal* **2016**, *48*, 355-372.
10. Seto, K.C.; Reenberg, A.; Boone, C.G.; Fragkias, M.; Haase, D.; Langanke, T.; Marcotullio, P.; Munroe, D.K.; Olah, B.; Simon, D. Urban land teleconnections and sustainability. *P Natl Acad Sci USA* **2012**, *109*, 7687-7692.
11. Zhu, X.; Guo, D.S. Mapping large spatial flow data with hierarchical clustering. *Transactions In Gis* **2014**, *18*, 421-435.
12. Adams, P.C. A taxonomy for communication geography. *Prog Hum Geog* **2011**, *35*, 37-57.
13. Mesbah, M.; Currie, G.; Lennon, C.; Northcott, T. Spatial and temporal visualization of transit operations performance data at a network level. *Journal Of Transport Geography* **2012**, *25*, 15-26.
14. Fonte, C.C.; Fontes, D.; Cardoso, A. A web gis-based platform to harvest georeferenced data from social networks: Examples of data collection regarding disaster events. *Int J Online Eng* **2018**, *14*, 165-172.
15. Hale, M.L.; Ellis, D.; Gamble, R.; Walter, C.; Lin, J. Secuwear: An open source, multi-component hardware/software platform for exploring wearable security. *Ieee Int Conf Mo* **2015**, 97-104.
16. Li, M.; Sun, Y.R.; Fan, H.C. Contextualized relevance evaluation of geographic information for mobile users in location-based social networks. *Isprs Int J Geo-Inf* **2015**, *4*, 799-814.
17. Li, J.W.; Ye, Q.Q.; Deng, X.K.; Liu, Y.L.; Liu, Y.F. Spatial-temporal analysis on spring festival travel rush in china based on multisource big data. *Sustainability-Basel* **2016**, *8*.
18. Rosvall, M.; Bergstrom, C.T. Maps of random walks on complex networks reveal community structure. *P Natl Acad Sci USA* **2008**, *105*, 1118-1123.
19. Esquivel, A.V.; Rosvall, M. Compression of flow can reveal overlapping-module organization in networks. *Phys Rev X* **2011**, *1*, 1668-1678.
20. Zhou, M.; Yue, Y.; Li, Q.Q.; Wang, D.G. Portraying temporal dynamics of urban spatial divisions with mobile phone positioning data: A complex network approach. *Isprs Int J Geo-Inf* **2016**, *5*.
21. Kempinska, K.; Longley, P.; Shawe-Taylor, J. Interactional regions in cities: Making sense of flows across networked systems. *Int J Geogr Inf Sci* **2018**, *32*, 1348-1367.
22. Kim, K.; Oh, K.; Lee, Y.K.; Kim, S.; Jung, J.Y. An analysis on movement patterns between zones using smart card data in subway networks. *Int J Geogr Inf Sci* **2014**, *28*, 1781-1801.
23. Chen, Z.L.; Gong, X.; Xie, Z. An analysis of movement patterns between zones using taxi gps data. *Transactions In Gis* **2017**, *21*, 1341-1363.
24. Liu, L.A.; Hou, A.Y.; Biderman, A.; Ratti, C.; Chen, J. Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen. *2009 12th International Ieee Conference on Intelligent Transportation Systems (Itsc 2009)* **2009**, 1-6.
25. Munizaga, M.A.; Palma, C. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from santiago, chile. *Transport Res C-Emer* **2012**, *24*, 9-18.
26. Ghasemzadeh, M.; Fung, B.C.M.; Chen, R.; Awasthi, A. Anonymizing trajectory data for passenger flow analysis. *Transport Res C-Emer* **2014**, *39*, 63-79.
27. Zhang, Y.P.; Martens, K.; Long, Y. Revealing group travel behavior patterns with public transit smart card data. *Travel Behav Soc* **2018**, *10*, 42-52.
28. Chu, K.K.A.; Chapleau, R. Enriching archived smart card transaction data for transit demand modeling. *Transp Res Record* **2008**, 63-72.

29. Higuchi, T.; Shimamoto, H.; Uno, N.; Shiomi, Y. A trip-chain based combined mode and route choice network equilibrium model considering common lines problem in transit assignment model. *State Of the Art In the European Quantitative Oriented Transportation And Logistics Research* **2011**, *20*.
30. Concas, S.; DeSalvo, J.S. The effect of density and trip-chaining on the interaction between urban form and transit demand. *Journal Of Public Transportation* **2014**, *17*, 16-38.
31. Zhou, L.; Ji, Y.X.; Wang, Y.Z. Analysis of public transit trip chain of commuters based on mobile phone data and gps data. *2017 4th International Conference on Transportation Information And Safety (Ictis)* **2017**, 635-639.
32. Blythe, P.T. Improving public transport ticketing through smart cards. *P I Civil Eng-Munic* **2004**, *157*, 47-54.
33. Pelletier, M.P.; Trepanier, M.; Morency, C. Smart card data use in public transit: A literature review. *Transport Res C-Emer* **2011**, *19*, 557-568.
34. Wang, Y.; Lim, E.P.; Hwang, S.Y. Efficient algorithms for mining maximal valid groups. *Vldb Journal* **2008**, *17*, 515-535.
35. Aung, H.H.; Tan, K.L. Discovery of evolving convoys. *Scientific And Statistical Database Management* **2010**, *6187*, 196-213.
36. Li, Y.X.; Bailey, J.; Kulik, L. Efficient mining of platoon patterns in trajectory databases. *Data & Knowledge Engineering* **2015**, *100*, 167-187.
37. Williams, H.J.; Holton, M.D.; Shepard, E.L.C.; Largey, N.; Norman, B.; Ryan, P.G.; Duriez, O.; Scantlebury, M.; Quintana, F.; Magowan, E.A., *et al.* Identification of animal movement patterns using tri-axial magnetometry. *Mov Ecol* **2017**, *5*.
38. Ye, Q.; Tian, G.; Liu, G.; Ye, J.; Yao, X.; Liu, Q.; Lou, W.; Wu, S. Tupu methods of spatial-temporal pattern on land use change. Journal of Geographical Sciences **2004**, 14, 131-142.