DOI: 10.1111/tgis.12772

RESEARCH ARTICLE

Transactions in GIS WILEY

A spatially aware method for mapping movement-based and place-based regions from spatial flow networks

Sebastijan Sekulić[1] | **Jed Long**[2] | **Urška Demšar**[1]

[1]School of Geography and Sustainable Development, University of St Andrews, St Andrews, UK

[2]Department of Geography and Environment, Western University, London, Ontario, Canada

Correspondence
Sebastijan Sekulić, School of Geography and Sustainable Development, University of St Andrews, St Andrews KY16 9AJ, UK.
Email: ss372@st-andrews.ac.uk

Abstract

Community detection (CD) is a frequent method for analysing flow networks in geography. It allows us to partition the network into a set of densely interconnected regions, called communities. We introduce a new technique for including geographical weighting in existing methods for detecting spatially coherent communities. We take a link-based CD algorithm and adjust it to incorporate geographical weighting. We call this approach geographically weighted community detection (GWCD). Our method is demonstrated on two case studies of commonly encountered flow networks: commuter flows and taxi pick-up/drop-off flows. Further, we test different measures of distance for geographic weighting and compare our results with the unmodified CD algorithm. Our results show that GWCD can capture the geographical nature of flow regions, generating spatially smaller and more compact areas than if geography is omitted, and that it can be used to distinguish between different types of movement-type communities.

1 | INTRODUCTION

Spatially referenced origin–destination flow data are common across many application areas in geography and are used to model different types of movement processes, for example commuting (Farmer & Fotheringham, 2011), international migrations (Tranos, Gheasi, & Nijkamp, 2015), or urban transportation (Yang, Heppenstall, Turner,

& Comber, 2019). Movement is modelled using spatial flow networks, where locations of origins and destinations are represented by the vertices of the spatial network and flows between these locations are represented by the weighted edges. Two vertices in a flow network are connected by an edge if there is a record of movement between the two locations. Every connection can be directed (i.e., a flow between A and B is not the same as a flow between B and A) and weighted by a measure of the flow (e.g., the number of individuals) along the edge.

Using networks to represent movement is not new, for example, Goddard (1970) used taxi flow networks to partition London into traffic zones and Illeris and Pedersen (1968) investigated communication flow networks. More recently, the analysis of spatial flow networks is being rediscovered throughout a wide range of disciplines interested in analysing the structure of spatial networks, such as mobility flows inferred from contemporary mobile phone data (Song, Qu, Blumm, & Barabasi, 2010).

Spatial networks are any network where the nodes are located in a real space that has a metric property (Barthélemy, 2011). In the geographical context, that metric is usually a geographical (e.g., Euclidean) distance, but it can also be time or network distance. For example, for commuting, travel time might be more important than distance travelled, or when looking at distances in cities, network distance could be more interesting than the straight-line distance. The importance of geographical distance for studying regional interactions and movement flows is one of the foundations of quantitative geography (Eldridge & Jones, 1991; Fotheringham, 1981; Taylor, 1975). In the context of spatial flow networks, these processes commonly follow rules of the distance decay process. Distance decay means that interactions that are closer to each other in geographic space are stronger than those that are further apart. In terms of flows, this means that flows that are close in the geographical space have a higher chance of being more similar to each other. This is because distance is one of the constraints in movement decision-making—for example, it is more likely that someone will decide to commute to a place nearby than to a place further away (Halás, Klapka, & Kladivo, 2014). Cheng and Bertolini (2013) incorporate distance decay, competition, and diversity to measure urban job accessibility in Amsterdam. Distance decay therefore has both economic and cultural importance.

Distance decay is also important in spatial interaction, a common term for any movement or communication over space that results from a decision-making process (Roy & Thill, 2003). The early spatial interaction models were called gravity models, as they were inferred from Newton's law of universal gravitation (Zipf, 1946) and highly dependent on the distance between locations. Spatial interaction models have since been used for a variety of purposes, from retail (Siła-Nowicka & Fotheringham, 2019) and commuting (McArthur, Kleppe, Thorsen, & Ubøe, 2011), to large studies of human mobility (McCulloch, Golding, McVernon, Goodwin, & Tomko, 2021). While these models can be used for a statistical representation of mobility, they do not provide information on the underlying structure of movement, which is what spatial networks capture and what we study in this article.

Many networks show inherent structures, which means that their links or vertices can be organized into meaningful groups termed *communities* or clusters or modules (Fortunato & Hric, 2016). One particular class of network methods for studying spatial flows identifies these groups—these are the so-called community detection (CD) methods (Comber, Brunsdon, & Farmer, 2012; De Meo, Ferrara, Fiumara, & Provetti, 2014; De Montis, Caschili, & Chessa, 2013). CD methods are used to partition the network into sub-networks that are internally dense (i.e., strongly connected) with relatively weak connections to the other parts of the network (Newman, 2004). This allows researchers to find tightly connected groups of nodes or edges, identify central nodes, and find important shared connections (Ahajjam, El Haddad, & Badir, 2018). The identified groups of dense connections are called *communities* in the complex network literature in physics, and in spatial networks communities can be more broadly referred to as *regions*.

CD is used in various disciplines: it has been used to detect communities within a football teams network (He, Li, Soundarajan, & Hopcroft, 2018; Newman, Watts, & Strogatz, 2002; Wu, Huang, Hao, & Chen, 2012), to find groups of connected topics on the Internet (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), or to identify the most important individuals in a social network (Ahajjam et al., 2018). CD is widely used in biology: to detect protein interaction and find functional modules in yeast or gene networks (Fortunato, 2010; Guimerà & Nunes Amaral, 2005; Palla, Derényi, Farkas, & Vicsek, 2005; Pavlopoulos et al., 2011).

The study of large networks (and spatial networks) has become popular in social physics, which seeks to identify generalizable laws about social systems. However, the geographical context specific to different analytical methods and applications is often overlooked in social physics (O'Sullivan & Manson, 2015). In terms of spatial flow networks, this means that geographical context is commonly omitted from the analysis (Pavlopoulos et al., 2011), or even discarded on purpose (Expert, Evans, Blondel, & Lambiotte, 2011). A recent review of CD (Javed, Younis, Latif, Qadir, & Baig, 2018) lists the use of these methods in social networks, communication networks, e-commerce, scientometrics, biological networks, networks in healthcare and economics, with no mention of geographical or movement-related networks. Conceptualizing geographical context within spatial flow network methods is, however, crucial to understanding relevant movement patterns. In particular, geographical distance is strongly related to travel time and therefore affects the nature of movement flows and the resulting structure of the network (Cerina, De Leo, Barthelemy, & Chessa, 2012). There is therefore a need for new methods that can incorporate geographical structure within flow networks into the CD process.

In spatial statistics, the classical way geographers incorporate geographical structure into an analytical procedure is through geographical weighting by some measure of distance. Examples are geographically weighted regression (Fotheringham, Brunsdon, & Charlton, 2003) and geographically weighted principal components analysis (Harris, Brunsdon, & Charlton, 2011). These methods build local models that incorporate the concept of spatial autocorrelation and similarity of proximal data into the model. That is, for each model, data are weighted based on their geographical distance from an individual model point (observation).

We propose a similar approach for CD in flow networks to inherently represent the geographical similarity of nearby flows in the partitioning algorithm. Existing CD methods that do not use spatial information generally create numerically large clusters that cover large spatial areas and are not aligned in any spatial direction. Moreover, as traditional algorithms take only connectivity into account, we lose local information. We expect that incorporating geographical similarity into a CD algorithm will lead to communities that are smaller in area, more tightly clustered in geographic space, and more directionally aligned. For example, in commuting flows, this will allow us to identify flow patterns within versus between cities, which traditional CD methods are unable to do. Geographical weighting will therefore allow us to better explore local patterns and processes by explicitly capturing local information in the derivation of network communities.

We take an existing CD algorithm (Ahn, Bagrow, & Lehmann, 2010) and modify it to incorporate geographical weighting. Using two case studies, each using a different distance measure (Euclidean distance and travel time) and a different movement phenomenon (commuting and taxi traffic), we demonstrate how using geographically weighted community detection (GWCD) allows us to detect more meaningful patterns in geographic flow networks when compared to the traditional non-geographically weighted models. We use a hierarchical clustering algorithm (HLC) because of its ability to cluster edges instead of nodes, as our interest is in detecting similar movement flows represented by edges in the network. The edge-based HLC method enables us to use both topological and geographical information to identify clusters of interconnected movement flows that have similar travel lengths and directions.

The remainder of the article is structured as follows. In Section 2 we give an overview of related work. Section 3 describes the edge-based HLC CD method and the geographically weighted extension. Section 4 introduces the two case studies and presents results, where we compare GWCD against the non-geographically weighted method. In Section 5 we discuss the results, place our work in the context of the existing literature, and present ideas for future work. Section 6 concludes.

2 | BACKGROUND

Communities within a network are defined as a set of network entities that are closer, in a network sense, to each other relative to other elements of the network (Newman, 2004). Elements contained within a community

share similar properties, roles, or behaviour (Coscia, Giannotti, & Pedreschi, 2011). CD is a class of methods used to detect communities in the network by using the properties of the network and the network structure (Newman, 2004).

Multiple approaches have been developed to identify different types of communities, for example the Louvain method (Blondel et al., 2008), the "betweenness" calculation (Girvan & Newman, 2002), Infomap (Rosvall & Bergstrom, 2008), hierarchical link clustering (HLC) (Ahn et al., 2010), and others (see Coscia et al., 2011 for a review).

Physicists focus on network properties in CD, for example, a frequent test is if the network is *scale-free*, meaning that its degree distribution follows (asymptotically) a power law (Guimerà & Nunes Amaral, 2005; Palla et al., 2005; Pavlopoulos et al., 2011). However, physics typically ignores any geographical component in the networks, even if the analysis of geographic networks would benefit from explicitly utilizing this component (Farmer & Fotheringham, 2011). For example, Comber et al. (2012) use a selection of CD algorithms to infer land use from land cover and show how a popular CD algorithm cannot universally be applied to geographical networks. Hannigan, Hernandez, Medina, Roos, and Shakarian (2013) use the spatial location of nodes and introduce a new metric to measure the quality of a community partition in a geolocated social network called "spatially near modularity" (Hannigan et al., 2013). They also compare a non-spatial CD algorithm (Blondel et al., 2008) against their algorithm.

One of the major rationales for performing CD on spatial networks is to generate a *regionalization*. It is well known in geographic and regional studies literature that using standard administrative areas for policymaking, research, and resource distribution may not be useful in terms of area coherence (Ball, 1980; Casado-Díaz, 2000). One way to overcome this obstacle is to replace the standard administrative areas with functional regions, which are regions where aggregated supply and demand meet within a spatially near region (Van Der Laan & Schalke, 2001). A data-driven way to define these regions is to use CD on networks that represent geographic flows. For example, Farmer and Fotheringham (2011) use CD on travel-to-work data to detect functional regions in Ireland. They identify a partition of the travel-to-work network where interaction within the regions is maximized and interaction between the regions is minimized. Other examples include the use of commuter flows to identify commuting regions within Sardinia (De Montis et al., 2013) and the use of taxi flows to identify functional regions within London (Demšar, Reades, Manley, & Batty, 2014).

CD algorithms can be classified as one of two main types: algorithms for disjoint communities and algorithms for overlapping communities (Javed et al., 2018). In a disjoint CD algorithm, every node belongs to a single community, which is why these methods are referred to as node-based CD methods (Fortunato, 2010; Newman & Girvan, 2004; Yang, McAuley, & Leskovec, 2013). In these methods, there is no overlap between communities. For spatial networks, this means that the detected communities are clusters of places representing *disjoint geographic regions* that have similar properties or express a similar behaviour, which then correspond to functional regions. In an overlapping CD algorithm, however, each node can belong to several different communities (Palla et al., 2005). In the context of geographic regionalization, this means that regions can overlap, which is often not desirable if the communities (i.e., regions) are to be used as an alternative to standard administrative areas (which typically do not overlap). For this reason, overlapping CD algorithms are less frequently applied to geographical networks. However, there is significant potential in using such algorithms in the context of studying spatial flow networks that represent movement. For example, overlapping methods could be used to understand polycentric flows (Zhong, Arisona, Huang, Batty, & Schmitt, 2014), where several peripheral regions feed into a central region but not into each other.

Overlapping CD algorithms classify edges of the network into communities instead of nodes: these methods are termed link-based (or edge-based) CD methods. In the context of spatial flow networks, this means that link-based CD methods group flows between locations rather than the locations (or geographical units) themselves. This also means that every node can be assigned to multiple communities, as every node can have many links that enter and exit the node (Ahn et al., 2010; Kim & Kim, 2015).

As we are interested in grouping movement flows with similar geometrical and geographical properties, we use a link-based CD method. This means that instead of grouping nodes in connected regions and ignoring movement between these regions (Figure 1a), we group the flows between regions, which allows us to identify sub-networks representing similar movement properties within a particular place (e.g., the two regions that are the same in panels A and B of Figure 1), as well as sub-networks capturing flows between places (Figure 1b). We have therefore chosen a link-based approach for GWCD in spatial flow networks.

3 | METHODS

In this section, we describe our modifications of an existing CD algorithm to incorporate geographical weighting (using distance). We first define flows and flow networks and introduce the terminology. After this, we explain the original algorithm and how we included geographical weighting into the process. We further hypothesize that statistical and spatial metrics, together with the inherent mechanism of the algorithm, will enable us to distinguish between two different types of movement communities. We expect to be able to identify scattered movement within a place (within a town, for example) and directed movement connecting two different places (e.g., movement between a satellite town and a major city centre).

3.1 | Spatial flow networks

A flow is defined as a record of movement between a recorded origin and destination. A collection of interconnected flows is called a flow network, which in its mathematical form is a graph where origin and destination locations are the nodes of the network and flows are the edges (or links) between nodes. Specifically, a

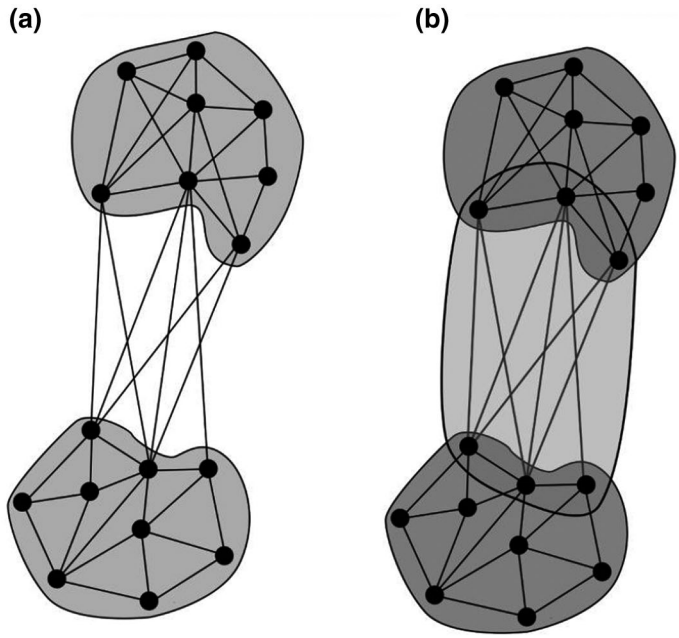

FIGURE 1 Difference between the two approaches to community detection: (a) an example of node-based communities; and (b) an example of link-based communities in the same network. The link-based approach identifies the community of flows between the two regions that are the same in both approaches

flow f_{ij} is a record of movement (of people or goods) between an origin node (n_i) and a destination node (n_j). Flows are often associated with a value (or measurement) that can be related to the importance or magnitude of the flow. Flow magnitudes can be used to define the weight w_{ij} of the flow where the weight can be representative of many things, most commonly being the number of travellers or the amount of goods moving between two nodes. Weighted networks offer the opportunity to capture more meaningful patterns about the relationships between spatial locations, as we can infer measures of how strong a connection is from network weights.

If by definition $f_{ij} = f_{ji}$, then the network is an undirected flow network. Likewise, if the network is constructed in a way that flows $f_{ij} \neq f_{ji}$, then the network is a directed flow network (Cormen, Stein, Rivest, & Leiserson, 2001). Depending on the nature of the application, the direction of the flows can be crucial for understanding the spatial patterns of flows in the network. In most practical examples in geographical research, flow networks will have implied directionality, as movement is inherently directional. For example, when exploring migration, undirected networks could be useful for capturing total (or gross) movement, but directed networks are required to capture immigration versus emigration flows, or net flows of goods or people.

Directed networks have rarely been used in CD because directionality adds an additional level of mathematical complexity (Leicht & Newman, 2008). While Leicht and Newman (2008) say that computational efficiency is almost identical when using algorithms for directed as for undirected networks, this applies to node-based CD methods only. Using information about directionality increases the number of total flows (i.e., each flow can have two directions) and therefore increases the computational cost and adds additional complexity when analysing directional networks. Therefore, we first choose to introduce edge-based geographical weights for undirected flow networks.

Flow networks in geography are often closely connected to their underlying geography and therefore capture spatial information about flows (Cerina et al., 2012). To define this mathematically, let $N = \{n_1, n_2, n_3, \ldots, n_k\}$ be the set of all nodes in a spatial network. Then, every $\boldsymbol{n}$ in N is associated with geographical coordinates (x_n, y_n) (e.g., longitude and latitude, or projected coordinates) used to capture node locations. Likewise, flows between coordinates can be associated with geographical measurements; for example, each edge can be characterized by the distance between its respective nodes. The distance can be defined as Euclidean, network, cost, or another conceptual measurement of distance. In a spatial network, we typically associate *both* the distance and the magnitude of the flow between two nodes in the network as the weight (importance) of said flow (Cerina et al., 2012).

3.2 | Geographical weigthing in community detection

Geographically weighted methods have a strong theoretical foundation in a variety of spatial models (Brunsdon, Fotheringham, & Charlton, 1996) because they are able to capture both the implicit spatial relationships between locations and the spatial heterogeneity in the processes they seek to understand. Geographical weighting in flow networks will modify the behaviour and sensitivity of CD algorithms to favour the grouping of geographically similar flows (Figure 2). Figure 2a shows communities where the algorithm considers only the number of connections each node has (Newman, 2004), while in Figure 2b the distance between points has been taken into consideration, along with the number of connections. In this example, we consider that all the connections are of the same importance. A spatially sensitive algorithm would prefer an object that is closer and with fewer connections rather than an object that is distant but with stronger connections. This means that a non-spatial approach, such as in Figure 2a, identifies a long community (grey nodes) and disregards that within-community distances in this cluster are large. A spatial version of the same algorithm (Figure 2b), however, adds an emphasis on nodes being geographically proximate and therefore identifies two clusters with short within-community distances.

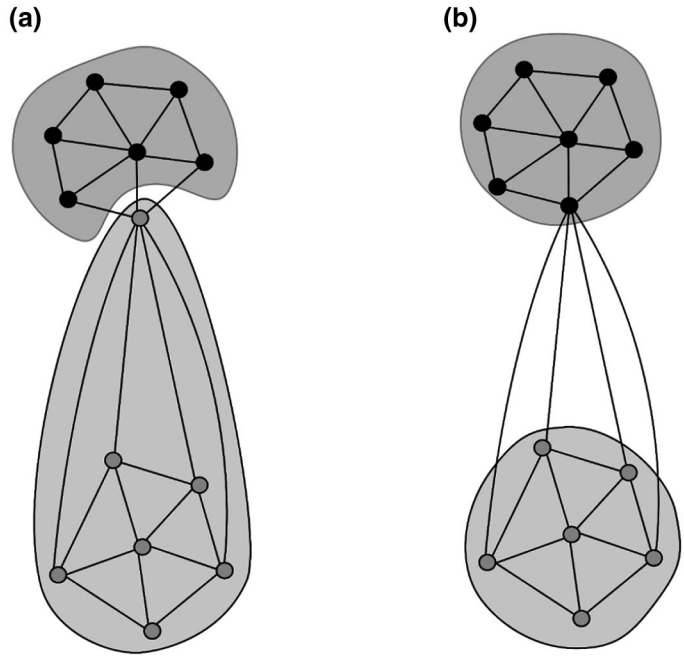

FIGURE 2 Difference between traditional community detection and the geographically weighted version. Grey areas show which nodes get grouped together in the same community. (a) Standard CD approach; and (b) approach using spatial weighting. In the geographically weighted example, nearby nodes are grouped together even if they share more connections with the nodes that are far away, thus prioritizing geography

3.3 | Community detection with HLC method

The HLC algorithm has a unique property that allows us to detect overlapping communities, while using link partitioning (Javed et al., 2018). Keeping and classifying edges allows us to use both topological and geographical information to identify clusters of interconnected movement flows that have similar travel lengths and directions.

To perform CD on movement flows using the HLC method we must define similarity between flows. Similarity measures for CD are typically based on network topology, that is, they use only the structure of the network and the number of connections between nodes. In the HLC algorithm, the similarity of two links is calculated using the so-called Jaccard similarity, which is computed using the number of secondary connections between two links that share the same node (called the keystone node; Figure 3). Secondary connections are links that originate from the non-shared nodes of the two links that we are comparing and end in the same node (where the end node is not the keystone node). The greater the number of secondary connections between two links, the more similar they are to each other (Lancichinetti & Fortunato, 2009).

For clarity, we list the node types used to calculate Jaccard's similarity between two links here:

- keystone node, the node that is common for two links we are comparing;
- impost nodes, the other two nodes that are not in common we are comparing;
- neighbouring nodes, a collection of all the nodes that belong to links that originate from the impost nodes;
- shared nodes, neighbouring nodes that are the end node of the links that originate in both impost nodes.

The Jaccard similarity measure is calculated as follows (Ahn et al., 2010). Starting with a pair of flows f_{ik} and f_{jk}, which share a node k (the keystone node) and end in nodes i and j (the impost nodes), we define $n(i)$ to be a set of

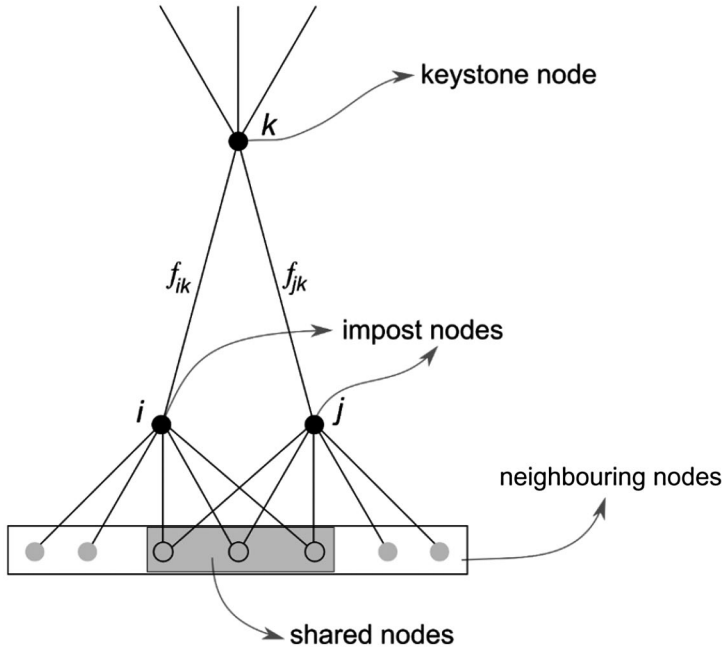

FIGURE 3 Visualization of keystone node and the surrounding edges and nodes that are used for similarity calculation. If f_{ik} and f_{jk} represent flows for which we would like to calculate similarity, then the secondary connections are the ones between the impost nodes and the shared nodes. The higher is the ratio between the number of shared nodes and the number of neighbouring nodes, the more similar flows f_{ik} and f_{jk} are

the given keystone node k and all the neighbouring nodes of i; $n(i) = \{k \cup \text{neighbours of } i\}$. The Jaccard similarity S_J between a pair of flows f_{ik} and f_{jk} is then calculated using the Jaccard index:

$$S_J(f_{ij}, f_{jk}) = \frac{|n(i) \cap n(j)|}{|n(i) \cup n(j)|} \tag{1}$$

This similarity is therefore equal to the ratio between the number of shared secondary nodes that two links have in common and all their secondary nodes (shared and neighbouring, Figure 3).

For weighted networks, the Jaccard index can be generalized to the Tanimoto coefficient (Tanimoto, 1958). For each node i we define the vector $\mathbf{a_i}$ as a length-n array that contains weights of links from node i to all the other nodes in the network (where n is the number of nodes in the network). Where no flow occurs between two nodes, the corresponding element in $\mathbf{a_i}$ is 0, that is, we set the weight as 0. If a node is connected to itself, we use that weight for the ith value in $\mathbf{a_i}$, otherwise, in networks where there is no value for self-connecting flows (f_{ii}) (i.e., connection of the node i with itself), we set the corresponding element of $\mathbf{a_i}$ to be the average flow weight of all flows stemming from node i.

To calculate the Tanimoto similarity between two edges f_{ik} and f_{jk} that share a node, we take corresponding vectors $\mathbf{a_i}$ and $\mathbf{a_j}$ and use them in the following formula:

$$S(f_{ki}, f_{kj}) = \frac{\mathbf{a_i} \cdot \mathbf{a_j}}{\left|\mathbf{a_i}^2\right| + \left|\mathbf{a_j}^2\right| - \mathbf{a_i} \cdot \mathbf{a_j}} \tag{2}$$

Here, $S(f_{ki}, f_{kj})$ is the similarity between edges f_{ki} and f_{kj},k is the shared node of two flows (the keystone node, Figure 3), and arrays $\mathbf{a_i}$ and $\mathbf{a_j}$ represent weights associated with flows originating from nodes i and j, respectively.

Once we have calculated the similarity values between every pair of flows, these are arranged in a similarity matrix of size $N \times N$, where F is the number of all flows. To identify communities based on this similarity, the HLC algorithm then uses hierarchical clustering on the similarity matrix. Hierarchical clustering starts merging flows by their similarity to generate communities and merges groups until all the elements are contained in a single community. Thus, the HLC method produces a dendrogram which contains the hierarchical structure of the network. A dendrogram is a tree representing the process of adding nodes, where the leaves represent individual elements (in our case flows) that are then merged at each step and where at each step the two nodes of the tree that were merged create a new node that is connected with the two nodes that represented the original data items (or a data item and a group if that is what was merged). This process is repeated until the top node which contains all the groups and represents the full dataset is created (Jain, Murty, & Flynn, 1999).

To find communities in hierarchical clustering, we need to cut the dendrogram at a given threshold, which can then be used to generate communities dependent on that threshold level. To find the optimal cut threshold, the algorithm uses the partition density (Ahn et al., 2010), which measures the quality of a link partition. Partition density is calculated for each step (merge) of the hierarchical clustering and the unique maximum of the partition density (Ahn et al., 2010) is chosen as the threshold for cutting the dendrogram, representing the optimal partition with resulting communities separated at the optimal level, that is, they are most connected within each group and least connected between groups.

3.4 | Geographically weighted community detection

To add geographical weights to the CD algorithm, we include geographical information in the similarity calculation. Specifically, a modified geographical weight considers the distance between two locations alongside the magnitude of the flow. This is done by multiplying the weight of the network link by a chosen distance function $g(d)$.

Drawing on previous literature, a distance function $g(d)$ can be chosen in numerous ways, but most commonly, distance functions are chosen to be variations of the exponential or power function (Chen, 2015; Taylor, 1975). Here we propose an exponential function that considers the ratio of the distance of each flow to a cutoff distance, where the chosen cutoff distance is a distance at which the effect of geographical distance starts to fade (this is conceptually similar to the estimate of the range from a semi-variogram). In this article we apply weighting in such a way that shorter travel distances are more strongly weighted but, depending on the research question, other definitions would be warranted. We use the average length of all flows in the network as a cutoff distance, as we expect that the average distance will capture communities of average travel length, and define $g\left(d_{ij}\right)$ as:

$$g\left(d_{ij}\right) = \exp\left(-\beta * d_{ij}\right) \tag{3}$$

where d_{ij} is the length of a flow with end nodes i and j, $\beta = 1/\overline{d}$, and $\overline{d}$ is the mean length of all flows in the network.

For each flow, $g\left(d_{ij}\right)$ is multiplied by the magnitude of the flow (which corresponds to the original weight W_{ij}) to create the final geographical weight of the flow:

$$\boldsymbol{G}_{ij} = \boldsymbol{W}_{ij} * g\left(d_{ij}\right) \tag{4}$$

Here, $\boldsymbol{G}_{ij}$ is the geographical weight of the flow ij of the matrix $\boldsymbol{G}$ and $\boldsymbol{W}_{ij}$ is the measure of flow (magnitude of flow, weight of flow) between nodes i and j weights matrix $\boldsymbol{W}$. Matrix $\boldsymbol{W}_{ij}$ is predefined by the properties of the network (the number of travellers or the number of trips). Geographical weights build a matrix $\boldsymbol{G}$, which is then used as the weights in the Tanimoto similarity in the HLC CD algorithm.

3.5 | Characterizing movement patterns

For each flow, we use the coordinates of its origin and destination, the distance between them, the bearing, and the number of travellers (i.e., the magnitude of the flow). For each community from the adapted HLC algorithm, we then calculate the following summaries:

- the circular standard deviation of all flow bearings (direction) in a community;
- the kurtosis of flow bearings;
- the convex hull of all flows in the community.

The circular standard deviation of the bearings measures the directional similarity of flows in a network, which can be used to help identify the level of homogeneity (in direction) of flows within each community (e.g., homogeneous vs. heterogeneous distribution flow directions). The kurtosis of the flow directions identifies concentrated movement in a single direction in a community versus a community with more varied distribution of direction. The convex hull of the flow community captures the geographical extent of the community and will also typically be related to general flow length (defined by the parameter β). The area of the convex hull can further be used to capture the magnitude of movement (which is related to total distance), and the spatial shape of the convex hull (i.e., the elongation, which is the ratio between the longest and the shortest side of the minimum rectangle that contains the whole convex hull) differentiates between different patterns of movement. Further, convex hulls can be used to identify where communities overlap in geographic space (even if they do not share nodes) using basic spatial overlay analysis.

We hypothesize that we can use these metrics to distinguish two types of flow communities: movement-based communities and place-based communities (Figure 4). That is, we expect to find communities where most of the movement will be in a particular direction between two areas (movement-based communities) and communities where movement will be predominantly contained within a particular area (place-based communities). Movement-based communities will be associated with larger convex-hull areas and a low circular standard deviation of flow bearings. Place-based communities will be associated with smaller convex-hull areas and a larger circular standard deviation of flow bearings. Further, the kurtosis of bearings, which describes the form of the distribution

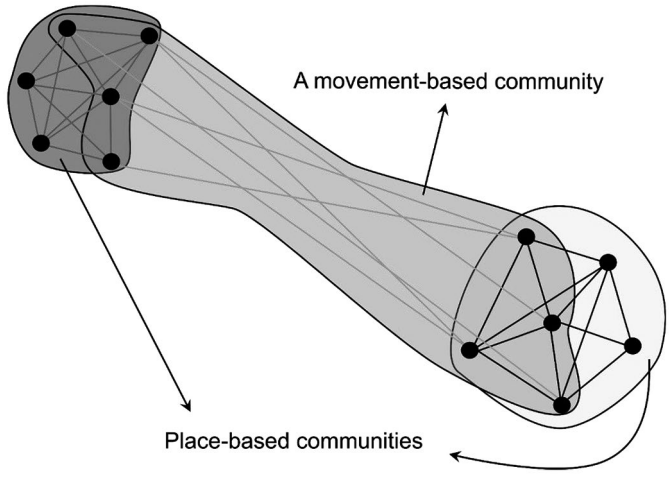

FIGURE 4 The difference between the movement-based and the place-based communities. A community is termed a movement-based community if flows are highly concentrated along a single directional bearing and are associated with larger areas, indicating directed movement between two—more distant—regions. In a place-based community, flows are distributed along many bearings and contained within a specific area

of bearing values, should be small in movement-based communities, since most of the flows are in one particular direction and large in flow-based communities, where the distribution of bearings is wider.

4 | CASE STUDIES

To show how spatial context changes the way CD algorithms work, we present two case studies for two different distance types. For the first case study, distance is calculated as Euclidean distance between centroids of areas of interest and in the second case study, the distance used is the actual recorded network travel distance. In the first case study we explore commuting patterns by using commuting flow data openly available from the Scottish Census 2011 (UK Data Service, 2015). In the second case study we focus on travel in New York by using free taxi traffic data from the New York Taxi and Limousine Commission (2019). In both cases we compare the results between using the HLC method with only flow weights (travel counts), which we call the classic approach, and using HLC with geographically weighted flows.

Datasets were chosen in such a way that they are publicly available and free to download from the respective offices' network sites.

4.1 | Commuting in Scotland

In this study we use commuting flows to identify so-called travel-to-work (TTW) areas (Ball, 1980), which partition a larger region into smaller areas that correspond to daily commuting zones connecting main corridors where people live and work. TTW areas are highly influenced by distance and are essential in urban planning, marketing, and labour market optimization (Coombes & Bond, 2008; Coombes & Openshaw, 1982). We propose that GWCD will identify a more geographically disaggregated set of TTW areas, which are at the same time more internally coherent in terms of movement patterns.

4.1.1 | Data and methods

In this case study we use commuting data from the Scottish Census 2011 (UK Data Service, 2015). We generate our flow network by using the spatially smallest level at which commuting data are collected (the output areas; OAs). In Scotland, an OA is a statistical unit that contains approximately 100 inhabitants (UK Data Service, 2015). Our data contain 1,339,578 flows between 46,352 OAs across the whole of Scotland. As the origin and destination areas are relatively small, we can expect very detailed commuting patterns and a lot of overlapping information.

To generate the spatial flow network, first we calculated the node locations as the centroid of each OA polygon. We assigned the flow magnitudes between nodes in the network from the census flows; that is, each link weight between two nodes representing two OAs is the aggregated number of people who live in one and work in the other area. We define flows as undirected, and therefore the number of commuters along an edge is the sum of commuters in both directions. The Euclidean distance between two OA nodes is used as the geographical measure associated with each flow.

4.1.2 | Results

To show the difference between the geographically weighted and the classical approach, we compare statistical and spatial summaries from each method. Using the geographically weighted HLC method the optimal partition of

the communities resulted in 7,366 unique communities, and in the classical (non-spatial) HLC method the optimal partition resulted in 2,429 communities (Figure 5). The threshold used for detecting communities (cutting the dendrogram) was automatic and inherent to the HLC method by calculating partition density and identifying the cutoff level, where the partition density was highest. We compared the geographically weighted and classical approaches by using the Kolmogorov–Smirnov test of the similarity of the distributions of the statistical summaries for communities in both partitions. We found statistically significant differences between the distribution of convex hull areas ($p < .001$) and the elongation measure of the convex hull ($p < .001$) (Figure 6). This suggests that the geographically weighted model is identifying fundamentally different spatial structures in the resulting flow communities.

To separate different types of movement, we studied the relationship between the different flow-based community metrics as proposed earlier. We found a relationship between circular standard deviation and kurtosis of bearing distribution (Figure 7a). We used this scatterplot to define different community types, and categorized data above and below the first quartile of circular standard deviation and the first quartile of kurtosis. The four resulting quadrants determine three different community types, and the lower left quadrant is empty. In Figure 7a we show a representative community in each quadrant and show the distributions and bearings of flow orientations in panels B, C, and D. We also show the flows in each respective community in the map. We call a community

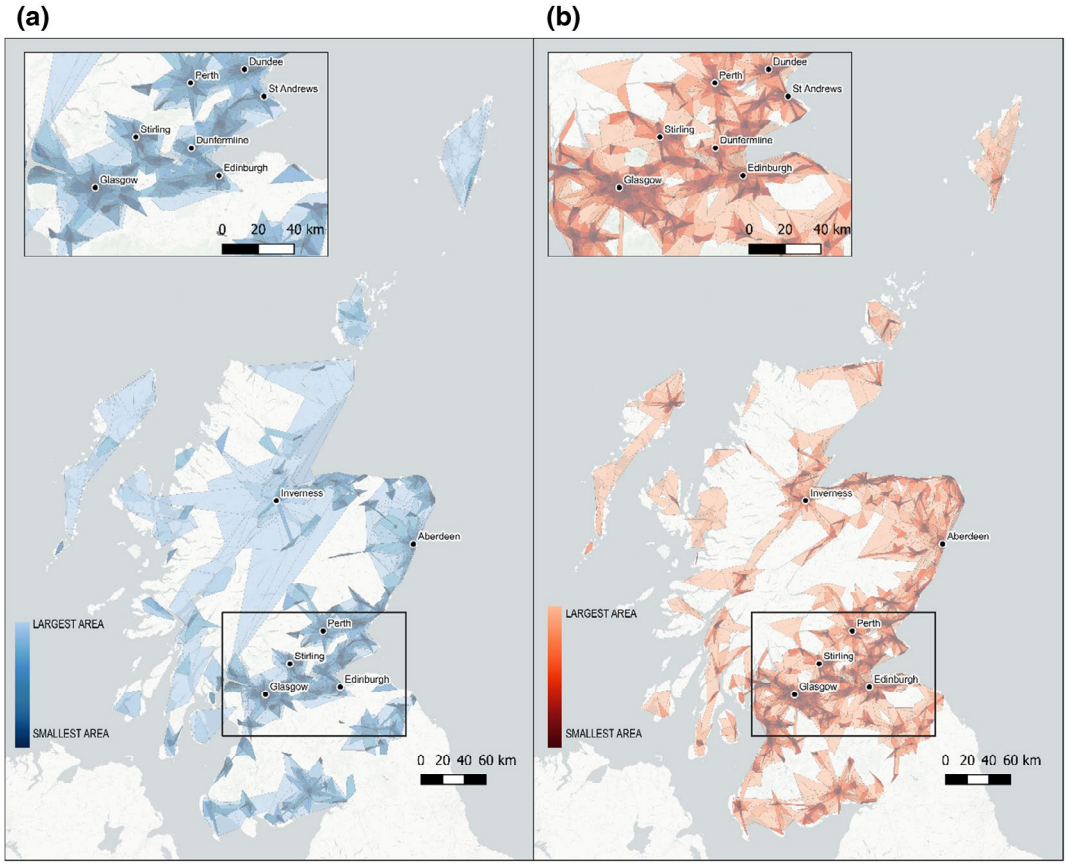

FIGURE 5 Spatial difference in classical and geographic community detection results: (a) Communities detected using the traditional CD method; and (b) communities from the GWCD model. The colour scheme is adjusted so that darker saturations represent communities with smaller convex hull area and communities with lighter saturation, larger convex hull area. Some large cities are marked with a black dot to show community concentration around populated places. This figure illustrates the difference between the spatial layout and the size of the communities; the reader is not expected to identify the individual community. Additional information on spatial distribution can be seen in Figure 6

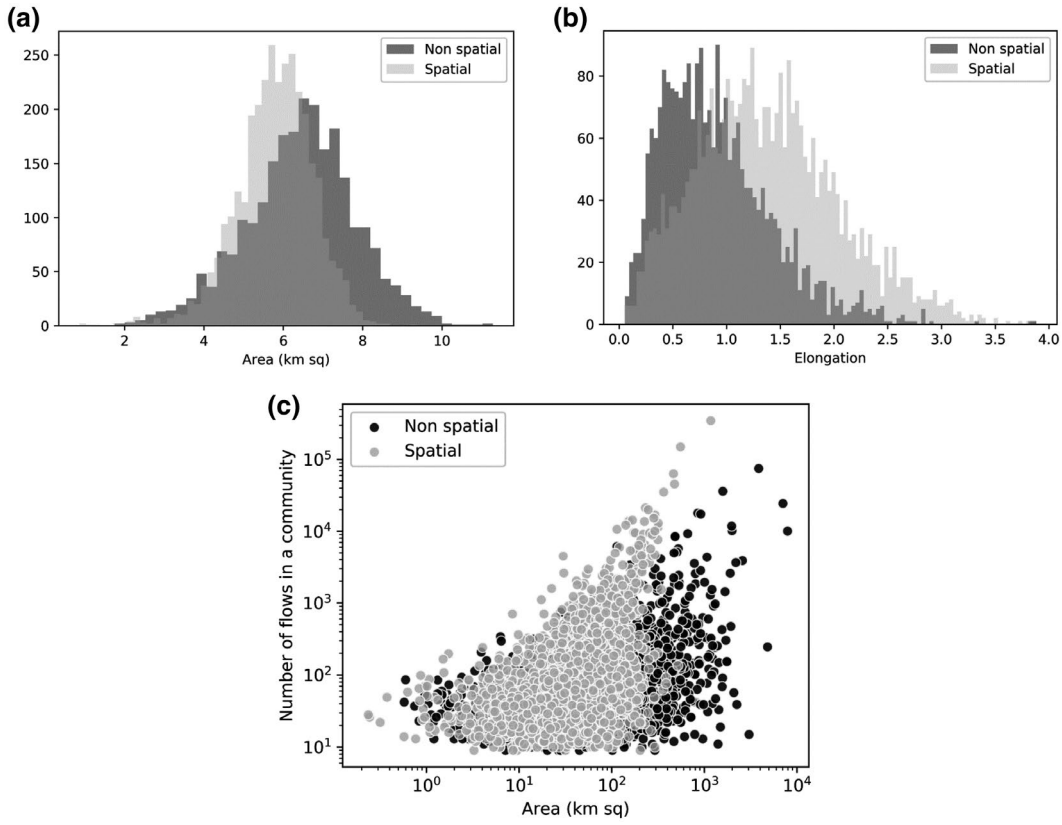

FIGURE 6 Distribution of spatial characteristics of communities obtained by using geographically weighted and traditional methods. Spatial communities are shown in light grey and non-spatial communities are shown in dark grey. (a) Distribution of communities by the area covered by their convex hull; (b) difference in the distribution of elongation of the minimum bounding rectangle of the communities; and (c) scatterplot of the relationship between the number of flows contained inside a community and the area of the convex hull of the said community

movement-based if kurtosis is high, which means that the distribution is quite narrow and the standard deviation of bearings is low (Figure 7b). This combination means that the community is strongly directed in one way, as visible on the map. For a place-based community the values are opposite; kurtosis is low and the standard deviation is high. This means that the community is widely distributed and that there is no specific direction to the movements (Figure 7c). Some communities don't fit either of these two types (Figure 7d), resulting in a mixed type between movement-based and place-based.

4.2 | New York taxi flows

4.2.1 | Data

New York taxi movement data are openly available from the New York City OpenData portal and provided by the New York Taxi and Limousine Commission (2019) at the level of taxi zones. A taxi zone is a spatial unit defined by the taxi service. While these data cover an area that is more densely populated and smaller in size than in our first case study, one interesting difference is that they include network travel distance, which we use in this

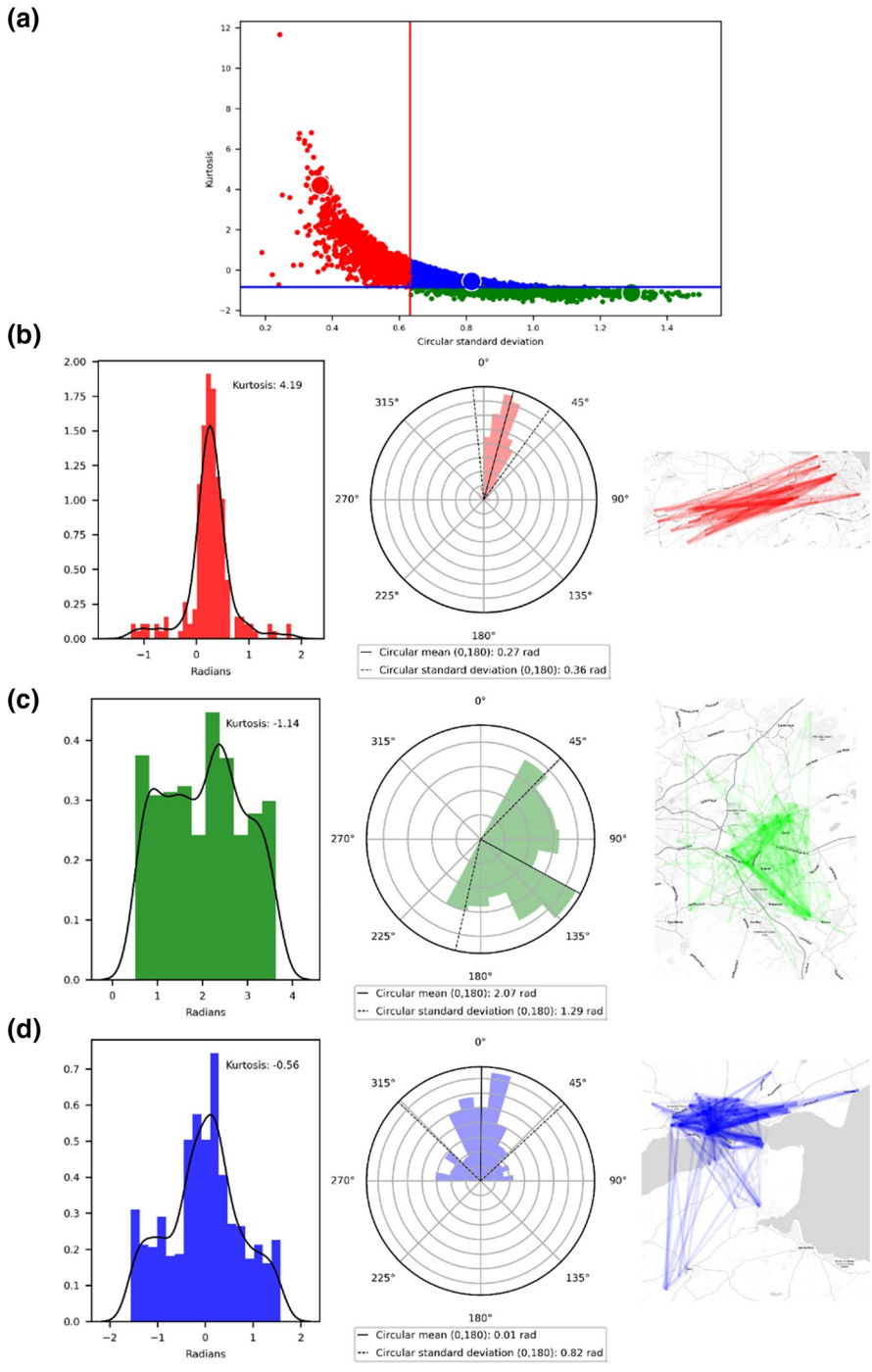

FIGURE 7 Identification of three community types for commuting in Scotland. The three large circles in panel a show three selected communities shown in panels b, c, and d in the respective colours. Each panel shows results for a different community type: the distribution of direction of flows, the radial distribution of the direction of flows, and the community on the map. The community shown in panel b has a narrow distribution of flow directions, with highly uni-directional flows, which is characteristic of a movement-based community. Panel c shows the opposite, the bearing distribution is wide and the directions are scattered, which corresponds to a place-based community. Panel d shows a community that does not have any distinct characteristics and cannot be classified into a community type

second case study instead of Euclidean distance. Dataset records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts (New York Taxi and Limousine Commission, 2019). We used a dataset comprising 25,594 flows between 263 taxi zones.

4.2.2 | Results

As the dataset is an order of magnitude smaller than the one used in the first case study, we get a smaller number of communities: using GWCD yields 33 communities and the classical (non-spatial) implementation of the HLC algorithm produces 29 communities (Figure 8). The criterion for deciding where to cut the dendrogram and choose the total number of communities was partition density from the HLC algorithm. Using the Kolmogorov–Smirnov test, we found a statistically significant difference between the distributions of the sizes of the communities ($p < .001$) and the elongation measure of the convex hulls ($p < .001$) for geographically weighted and aspatial approaches.

This example exhibits a similar relationship between kurtosis and circular deviation of the flow direction (Figure 9a) as in the previous case study, even though the number of data points is much smaller. As previously demonstrated, we can identify differences in community types if we split the scatterplot into four quadrants (Figure 9a). Figure 9a shows the three different types of communities, where we select a representative community from each quadrant and show how its direction, circular standard deviation, and kurtosis correspond to the flows on the map. Figure 9b shows a movement-based community, Figure 9d is a place-based community, and Figure 9c is an example of a mixed community between the two extremes.

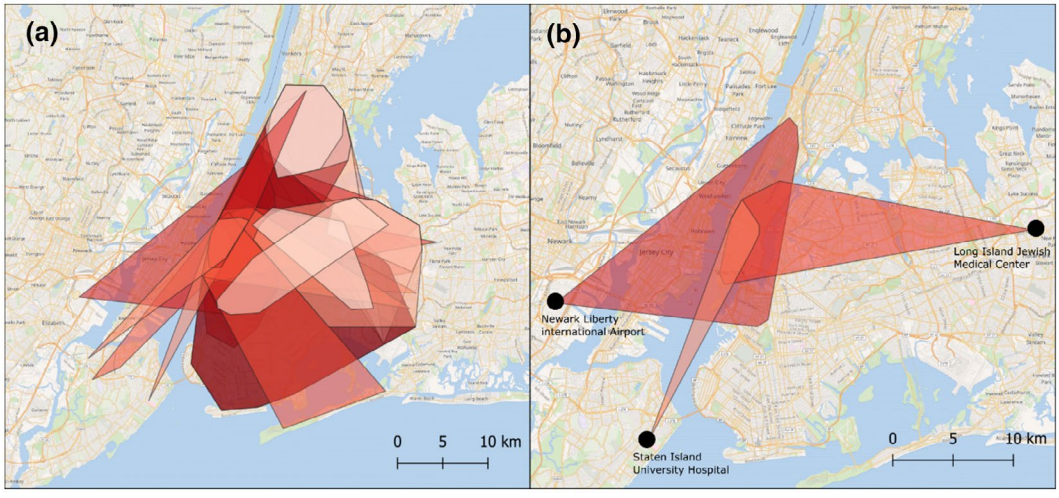

FIGURE 8 Convex hulls of the geographically weighted communities over the study area of New York City form a highly overlapping structure. Note that here we only show communities from the geographically weighted algorithm and not those obtained with the classic algorithm. (a) All overlapping communities; and (b) convex hulls that represent communities with the highest number of trips. At the extreme corners of the three selected communities, we have taxi zones that contain an airport and two hospitals

5 | DISCUSSION

Using spatial information for detecting communities in spatial networks facilitates the identification of underlying patterns that would be overlooked if we used only topological information. Specifically, spatially small communities, with a greater than average number of connections, would commonly be encompassed within another community. Additionally, we showed that using link-based CD enables us to detect overlapping communities defined by nodes. This allows us to better represent the underlying processes of movement and interactions between areas within the network. We demonstrated the importance of analysing spatial networks with their spatial properties and showed how they behave differently from networks without spatial properties. We further identified different community types by investigating the shape, size, and directional properties of flow communities using isotropy of flows, that is by studying the properties of the statistical distribution of flow orientation. We identified two extremes, which correspond to two different types of movement within a community. We found that if a community has a wide distribution of directions of flows and a high circular standard deviation, it most likely represents movement within a town, city, or region, termed a *place-based community*. However, if the directions of flow in the community are highly uni-directional and have a low circular standard deviation, we identify these communities as representing directed travel between two places, which we term a *movement-based community*. This suggests that using link-based GWCD can support the definition and mapping of functionally different movement regions. Comparing with regionalization work done by Farmer and Fotheringham (2011), in addition to partitioning the study area into functional regions, we are able to identify the type of movement process (place-based, movement-based, or a mixture of the two) that defines interactions between the regions.

One of the advantages of our approach is the flexibility with which different distance weighting functions can be implemented to characterize the bandwidth (or reach) of the distance weights, the value of the β parameter. In this way we can control the spatial size of the communities and investigate the phenomena on the scale that is required. Spatial weighting can go both ways, and we can do CD putting more importance on shorter flows with less traffic, or we can consider long flows more important by inverting the β parameter.

In the two case studies we are able to identify two extremes in community types (movement-based and place-based communities) straightforwardly by partitioning the respective distributions of circular standard deviation and kurtosis. However, communities could further be classified using additional information. As the method keeps the original flow identifiers, it is possible to join flows with additional tabular data (i.e., demographic census data and socio-economic data), which would allow us to link flows (and communities) to external information or to filter the communities depending on additional attributes.

Our method could potentially be applied to any spatial flow dataset with little or no pre-processing—for example, bike-share data (Médard De Chardon & Caruso, 2015), ship travel logs (Kaluza, Kölzsch, Gastner, & Blasius, 2010), or airport connections (Rodríguez-Déniz, Suau-Sanchez, & Voltes-Dorta, 2013). The method requires the following inputs: an origin–destination matrix in a list format and a list of nodes with their respective coordinates in a projected coordinate system. If the links do not have precomputed lengths, this can be done as part of pre-processing to speed up the algorithm. Theoretically, our algorithm could take a dataset of unlimited size, but the limitations could be the processing power, computer memory, and length of calculation, with the complexity being $O\left(n\overline{K}^2\right)$, where n is the number of vertices in the network and $\overline{K}$ is the average degree of the network. One way to decrease the computational complexity when using a method such as ours on a big flow network would be to prune the network to reduce the number of links. This could be done by removing the edges where the distance between nodes exceeds a given distance. This type of pruning aligns with geographically local methods being used to analyse other types of large data.

GWCD is computationally heavy (Blondel et al., 2008; Raghavan, Albert, & Kumara, 2007) and the issue of computing time is typically more pronounced when using link-based CD as the number of links is much higher

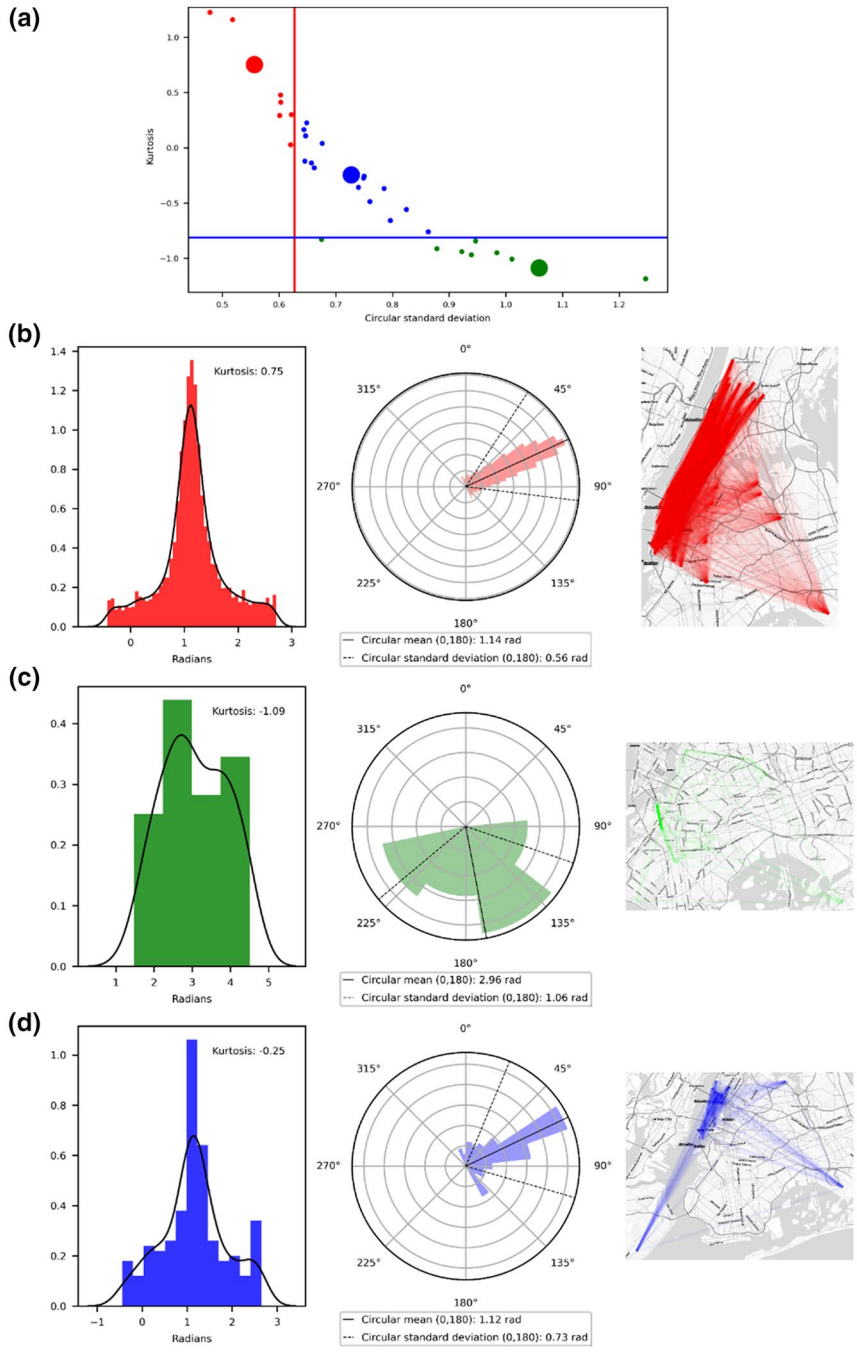

FIGURE 9 Categorization of communities in the taxi case study. The three large circles in panel a show three selected communities shown in panels b, c, and d in the respective colours. In comparison to Scottish commuting data, New York taxi data division into community types is not as highly polarized, but we can still detect different movement types. Each panel shows the different type of community and they show the distribution of direction of flows, the radial distribution of the direction, and the representation of the community on the map. The community shown in panel b has a narrow distribution and the flows are highly directed, which are characteristics of a movement-based community. Panel c shows the opposite, the distribution is wide and the directions are scattered, which then corresponds to a place-based community. Panel d shows a mixed community that cannot be classified into either of the two types

than the number of nodes (Blondel et al., 2008). Specifically, our analysis here on the Scottish commuting data (containing a network of ~1.6 million flows and ~50,000 nodes) was implemented in Python and takes about 12 hr on a dedicated workstation (2 × 12-core 2.10 GHz processor with 96 GB of RAM) to compute. For comparison, the modularity optimization method used by Blondel et al. (2008) can identify communities from a network of 118 million nodes in about 150 min. They do not discuss the processing power of their machine, but it is likely that they used either a cluster computer or a heavily parallelized workstation.

The current implementation of GWCD uses the HLC method (Ahn et al., 2010) and is limited to the application of undirected graphs. This is a major limitation for the study of spatial flows, because these methods cannot differentiate between flows going to or from a specific node. Given the importance of directionality in the study of movement (Jacoby & Freeman, 2016), a major direction of future research will be to extend current methods to capture directional flows in spatial networks.

Another major emphasis of future research would be to devise a mechanism to incorporate time and/or change over time into the similarity measure, and subsequently the community definition. Because patterns of movement change during the day, week, season, or across years, incorporating temporal dynamics will be useful for understanding spatial flows in modern flow datasets (e.g., bike-share data). Some of this is already on-going, for example, Sarzynska, Leicht, Chowell, and Porter (2016) explore temporal networks and time-dependent community structure and conclude that it is important to develop CD methods that take advantage of spatial and temporal information. To properly understand and model spatiotemporal processes such as movement, we need to know the relationship between what, where, when, and how all three components interact (Peuquet, 1994), and a spatio-temporal version of CD may be the way forward to investigate this.

6 | CONCLUSION

In this article we propose the use of the geographical information inherent in spatial flow networks to augment existing CD methods and allow for identification of regions with different movement types. We adapted a link-based CD method with geographical weighting to obtain the results that can be connected to real space. The approach we took allows us to use both pieces of information in the definition of the similarity matrix (i.e., by combining the Tanimoto coefficient with a distance-based weight). Our results highlight that communities detected using GWCD tend to be smaller, but more elongated, than those calculated using classical algorithms and have spatial attributes that enable us to explore in more detail what those communities represent. A major contribution of our work is that using geographically weighted, link-based clustering, we are able to identify different types of communities (movement, place, and mixed) that relate to underlying movement patterns and processes.

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council and TheScottish Graduate School of Social Science.

DISCLOSURE STATEMENT

No financial interest or benefit has arisen from this research.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available from the Census UK Data Service portal at https://census.ukdataservice.ac.uk/ and from the New York Taxi and Limousine Commission at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Example code used in this article can be found at https://github.com/lankabel/gwcd.

ORCID

Sebastijan Sekulić https://orcid.org/0000-0001-7208-7393
Jed Long https://orcid.org/0000-0003-3961-3085
Urška Demšar https://orcid.org/0000-0001-7791-2807

REFERENCES

Ahajjam, S., El Haddad, M., & Badir, H. (2018). A new scalable leader-community detection approach for community detection in social networks. *Social Networks*, *54*, 41–49. https://doi.org/10.1016/j.socnet.2017.11.004

Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–764. https://doi.org/10.1038/nature09182

Ball, R. M. (1980). The use and definition of travel-to-work areas in Great Britain: Some problems. *Regional Studies*, *14*(2), 125–139. https://doi.org/10.1080/09595238000185121

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, *499*(1–3), 1–101. https://doi.org/10.1016/j.physrep.2010.11.002

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Casado-Díaz, J. M. (2000). Local labour market areas in Spain: A case study. *Regional Studies*, *34*(9), 843–856. https://doi.org/10.1080/00343400020002976

Cerina, F., De Leo, V., Barthelemy, M., & Chessa, A. (2012). Spatial correlations in attribute communities. *PLoS One*, *7*(5), e37507. https://doi.org/10.1371/journal.pone.0037507

Chen, Y. (2015). The distance–decay function of geographical gravity model: Power law or exponential law? *Chaos, Solitons & Fractals*, *77*, 174–189. https://doi.org/10.1016/j.chaos.2015.05.022

Cheng, J., & Bertolini, L. (2013). Measuring urban job accessibility with distance decay, competition and diversity. *Journal of Transport Geography*, *30*, 100–109. https://doi.org/10.1016/j.jtrangeo.2013.03.005

Comber, A. J., Brunsdon, C. F., & Farmer, C. J. (2012). Community detection in spatial networks: Inferring land use from a planar graph of land cover objects. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 274–282. https://doi.org/10.1016/j.jag.2012.01.020

Coombes, M., & Bond, S. (2008). *Travel-to-work areas: The 2007 review*. London, UK: Office for National Statistics.

Coombes, M. G., & Openshaw, S. (1982). The use and definition of travel-to-work areas in Great Britain: Some comments. *Regional Studies*, *16*(2), 141–149. https://doi.org/10.1080/09595238200185161

Cormen, T. H., Stein, C., Rivest, R. L., & Leiserson, C. E. (2001). *Introduction to algorithms*. New York, NY: McGraw-Hill.

Coscia, M., Giannotti, F., & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, *4*(5), 512–546. https://doi.org/10.1002/sam.10133

De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2014). Mixing local and global information for community detection in large networks. *Journal of Computer and System Sciences*, *80*(1), 72–87. https://doi.org/10.1016/j.jcss.2013.03.012

De Montis, A., Caschili, S., & Chessa, A. (2013). Commuter networks and community detection: A method for planning sub regional areas. *The European Physical Journal Special Topics*, *215*(1), 75–91. https://doi.org/10.1140/epjst/e2013-01716-4

Demšar, U., Reades, J., Manley, E., & Batty, J. M. (2014). Edge-based communities for identification of functional regions in a taxi flow network. In K. Stewart, E. Pebesma, G. Navratil, P. Fogliaroni, & M. Duckham (Eds.), *Extended abstract proceedings of GIScience 2014* (pp. 55–60). Vienna, Austria: Department of Geodesy and Geoinformation, Vienna University of Technology.

Eldridge, J. D., & Jones, J. P., III. (1991). Warped space: A geography of distance decay. *The Professional Geographer*, *43*(4), 500–511. https://doi.org/10.1111/j.0033-0124.1991.00500.x

Expert, P., Evans, T. S., Blondel, V. D., & Lambiotte, R. (2011). Uncovering space independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7663–7668. https://doi.org/10.1073/pnas.1018962108

Farmer, C. J. Q., & Fotheringham, A. S. (2011). Network-based functional regions. *Environment and Planning A*, *43*(11), 2723–2741. https://doi.org/10.1068/a44136

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Fotheringham, A. S. (1981). Spatial structure and distance–decay parameters. *Annals of the Association of American Geographers*, *71*(3), 425–436. Retrieved from http://www.jstor.org/stable/2562901

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, UK: John Wiley & Sons.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Goddard, J. B. (1970). Functional regions within the city centre: A study by factor analysis of taxi flows in central London. *Transactions of the Institute of British Geographers*, *49*, 161–182. https://doi.org/10.2307/621647

Guimerà, R., & Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*(7028), 895–900. https://doi.org/10.1038/nature03288

Halás, M., Klapka, P., & Kladivo, P. (2014). Distance–decay functions for daily travel-to-work flows. *Journal of Transport Geography*, *35*, 107–119. https://doi.org/10.1016/j.jtrangeo.2014.02.001

Hannigan, J., Hernandez, G., Medina, R. M., Roos, P., & Shakarian, P. (2013). Mining for spatially-near communities in geo-located social networks. In *Proceedings of the 2013 AAAI Symposium on Social Networks and Social Contagion: Web Analytics and Computational Social Science*, Arlington, VA (pp. 16–23). Menlo Park, CA: AAAI.

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717–1736. https://doi.org/10.1080/13658816.2011.554838

He, K., Li, Y., Soundarajan, S., & Hopcroft, J. E. (2018). Hidden community detection in social networks. *Information Sciences*, *425*, 92–106. https://doi.org/10.1016/j.ins.2017.10.019

Illeris, S., & Pedersen, P. O. (1968). *Central places and functional regions in Denmark: Factor analysis of telephone traffic*. Lund, Sweden: Geografisk tidsskrift.

Jacoby, D. M. P., & Freeman, R. (2016). Emerging network-based tools in movement ecology. *Trends in Ecology & Evolution*, *31*(4), 301–314. https://doi.org/10.1016/j.tree.2016.01.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, *108*, 87–111. https://doi.org/10.1016/j.jnca.2018.02.011

Kaluza, P., Kölzsch, A., Gastner, M. T., & Blasius, B. (2010). The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, *7*(48), 1093–1103. https://doi.org/10.1098/rsif.2009.0495

Kim, P., & Kim, S. (2015). Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering. *Physica A: Statistical Mechanics and its Applications*, *417*, 46–56. https://doi.org/10.1016/j.physa.2014.09.035

Lancichinetti, A., & Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, *80*(1), 016118. https://doi.org/10.1103/PhysRevE.80.016118

Leicht, E. A., & Newman, M. E. J. (2008). Community structure in directed networks. *Physical Review Letters*, *100*(11), 118703. https://doi.org/10.1103/PhysRevLett.100.118703

McArthur, D. P., Kleppe, G., Thorsen, I., & Ubøe, J. (2011). The spatial transferability of parameters in a gravity model of commuting flows. *Journal of Transport Geography*, *19*(4), 596–605. https://doi.org/10.1016/j.jtrangeo.2010.06.014

McCulloch, K., Golding, N., McVernon, J., Goodwin, S., & Tomko, M. (2021). Ensemble model for estimating continental-scale patterns of human movement: A case study of Australia. *Scientific Reports*, *11*(1), 4806. https://doi.org/10.1038/s41598-021-84198-6

Médard De Chardon, C., & Caruso, G. (2015). Estimating bike-share trips using station level data. *Transportation Research Part B: Methodological*, *78*, 260–279. https://doi.org/10.1016/j.trb.2015.05.003

New York Taxi and Limousine Commission. (2019). *New York City Taxi Trip Data, 2009–2018* (ICPSR 37254). Retrieved from https://www.icpsr.umich.edu/web/ICPSR/studies/37254

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B - Condensed Matter*, *38*(2), 321–330. https://doi.org/10.1140/epjb/e2004-00124-y

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(Suppl. 1), 2566–2572. https://doi.org/10.1073/pnas.012582999

O'Sullivan, D., & Manson, S. M. (2015). Do physicists have geography envy? And what can geographers learn from it? *Annals of the Association of American Geographers*, *105*(4), 704–722. https://doi.org/10.1080/00045608.2015.1039105

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818. https://doi.org/10.1038/nature03607

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., … Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, *4*(1), 10. https://doi.org/10.1186/1756-0381-4-10

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, *84*(3), 441–461. https://doi.org/10.1111/j.1467-8306.1994.tb01869.x

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106. https://doi.org/10.1103/PhysRevE.76.036106

Rodríguez-Déniz, H., Suau-Sanchez, P., & Voltes-Dorta, A. (2013). Classifying airports according to their hub dimensions: An application to the US domestic network. *Journal of Transport Geography*, *33*, 188–195. https://doi.org/10.1016/j.jtrangeo.2013.10.011

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Roy, J. R., & Thill, J.-C. (2003). Spatial interaction modelling. *Papers in Regional Science*, *83*(1), 339–361. https://doi.org/10.1007/s10110-003-0189-4

Sarzynska, M., Leicht, E. A., Chowell, G., & Porter, M. A. (2016). Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, *4*(3), 363–406. https://doi.org/10.1093/comnet/cnv027

Siła-Nowicka, K., & Fotheringham, A. S. (2019). Calibrating spatial interaction models from GPS tracking data: An example of retail behaviour. *Computers, Environment and Urban Systems*, *74*, 136–150. https://doi.org/10.1016/j.compenvurbsys.2018.10.005

Song, C., Qu, Z., Blumm, N., & Barabasi, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021. https://doi.org/10.1126/science.1177170

Tanimoto, T. T. (1958). *An elementary mathematical theory of classification and prediction*. New York, NY: IBM.

Taylor, P. J. (1975). *Distance decay in spatial interactions*. Norwich, UK: Geo Abstracts.

Tranos, E., Gheasi, M., & Nijkamp, P. (2015). International migration: A global complex network. *Environment and Planning B*, *42*(1), 4–22. https://doi.org/10.1068/b39042

UK Data Service. (2015). *2011 Census: Flow data*. Retrieved from https://census.ukdataservice.ac.uk/get-data/flow-data.aspx

Van Der Laan, L., & Schalke, R. (2001). Reality versus policy: The delineation and testing of local labour market and spatial policy areas. *European Planning Studies*, *9*(2), 201–221. https://doi.org/10.1080/09654310123131

Wu, Y.-J., Huang, H., Hao, Z.-F., & Chen, F. (2012). Local community detection using link similarity. *Journal of Computer Science and Technology*, *27*(6), 1261–1268. https://doi.org/10.1007/s11390-012-1302-4

Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes. In *Proceedings of the 13th IEEE International Conference on Data Mining*, Dallas, TX (pp. 1–10). Piscataway, NJ: IEEE.

Yang, Y., Heppenstall, A., Turner, A., & Comber, A. (2019). A spatiotemporal and graph-based analysis of dockless bike sharing patterns to understand urban flows over the last mile. *Computers, Environment and Urban Systems*, *77*, 101361. https://doi.org/10.1016/j.compenvurbsys.2019.101361

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, *28*(11), 2178–2199. https://doi.org/10.1080/13658816.2014.914521

Zipf, G. K. (1946). The P1 P2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, *11*(6), 677–686. https://doi.org/10.2307/2087063

How to cite this article: Sekulić S, Long J, Demšar U. A spatially aware method for mapping movement-based and place-based regions from spatial flow networks. *Transactions in GIS*. 2021;00:1–21. https://doi.org/10.1111/tgis.12772