International Journal of Geographical Information Science
Vol. 00, No. 00, Month 200x, 1–2

RESEARCH ARTICLE

Interactional regions in cities: making sense of flows across networked systems

Kira Kempinska[a*], Paul Longley[b] and John Shawe-Taylor[c]

[a]*Department of Security and Crime Science, University College London, UK*; [b]*Department of Geography, University College London, UK*; [c]*Department of Computer Science, University College London, UK*

(*Received 00 Month 200x; final version received 00 Month 200x*)

*Corresponding author. Email: kira.kowalska.13@ucl.ac.uk

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

Do administrative boundaries correspond to the observable ways in which people interact in urban space? As cities grow in complexity, and people interact over long distances with greater ease, so partitioning of cities needs to depart from conventional gravity models. The current state-of-the-art for uncovering *interactional* regions, i.e. regions reflective of observable human mobility and interaction patterns, is to apply community detection to networks constructed from vast amounts of human interactions, such as phone calls or flights. This approach is well suited for origin-destination activities, but not for activities involving multiple locations, such as police patrols, and is blind to spatial anomalies. As a result of the latter, community detection generates geographically coherent regions, which may appear plausible but give no insights into forces other than gravity that shape our interaction patterns.

This paper proposes novel approaches to regional delineation that address the aforementioned shortcomings. Firstly, it introduces topic modelling as an alternative tool for extracting interactional regions from tracking data. Secondly, it presents refinements of the topic modelling and community detection approaches that can uncover interaction patterns driven by forces other than spatial proximity. When applied to police patrol data, our methodology partitions the street network into non-overlapping patrol zones and detects popular long-distance routes between police stations. These findings could be used in the design of effective police districts, especially in light of recent funding cuts that promise to impact upon the ways in which policing and specifically patrols are carried out.

Keywords: Interactional Regions; Topic Modelling; Community Detection; Network Analysis; Police Districts

1. Introduction

Cities are "not simply places in space but systems of networks and flows" (Batty 2013). As such, they represent highly structured and dynamic environments that provide the loci of human mobility and interaction. The structure of cities both shapes and is shaped by patterns of human interactions, and hence urban analytics should be founded upon areal units that reflect such patterning.

To this end we propose the concept of *interactional regions* which reflect the ways in which people are observed to move and interact. Interactional regions are spatial envelopes that commonly bound human activities and interactions, such as consumer transactions, taxi routes or police patrols. They respect the natural ways in which people interact across space and, as such, their definition is essential for effective business and service planning, including the assignment of administrative responsibilities in public resource allocation.

Administrative geographies are inevitably an uneasy compromise between existing and past patterns of spatial interaction, with the latter encapsulated in so-called 'place effects' (Fotheringham 1997). From this perspective, places can themselves be construed as the accretion of past interactions, making places unique, but nonetheless comparable with others that have interactional histories that may be similar in different ways. Boundaries may have been created many decades ago, when human interactions and

mobility were predominantly local and the conceptual separation of human populations into fixed and geographically coherent regions was plausible and useful. However, the accelerating scale and pace of societal evolution combined with observable changes in the frictions of distance result in new, multifaceted and increasingly complex patterns of human connectivity (Singleton and Longley 2009). It is these that nevertheless define contemporary interactions between established places (Thiemann *et al.* 2010). Spatial interaction patterns are no longer a simple manifestation of the distance attenuation functions of traditional gravity models (if they ever were) or of opportunity functions of radiation models (Simini *et al.* 2012), but of a far more complex range of interacting factors that can only be uncovered by analysing vast amounts of human-generated flow data, such as phone calls (Blondel *et al.* 2010, Ratti *et al.* 2010), monetary transactions (Brockmann 2010, Vanhove 1999) or vehicle flows (Karlsson and Olsson 2006, Manley 2014).

In practice, the current state-of-the-art for identifying interactional regions is to apply community detection techniques to the flow networks created by aggregating human flows between locations (Blondel *et al.* 2010, Manley 2014, Ratti *et al.* 2010). This approach enables analysis of interactions without the geographical presupposition inherent to gravitational models. However, it has two important limitations. Firstly, it is designed for datasets with clearly defined origins and destinations for each interaction. Examples of such datasets include phone calls (with caveats), taxi journeys or retail transaction data. Counterexamples include continuously generated data, such as tracking data from police vehicles or mail delivery vans, where journey origins and destinations are not functionally defined or known.

Secondly, and most importantly, community detection is shaped by the pre-existing spatial structure of settlements (Besussi *et al.* 2010, Expert *et al.* 2011). In most cases, it uncovers regions that are strongly determined by geographical proximity at the expense of other underlying forces shaping the interactions. For instance, traffic flows are typically dominated by low-cost short-ranged interactions. As a result, community detection is blind to spatial anomalies and only identifies regions which are compact in physical space. This leads us to the central question of our work: can we detect interaction patterns that build upon more than distance attenuation? In other words, if we control for gravity-like forces, what other forces shape our interaction and mobility patterns? And can we develop a standard network methodology to uncover them?

In this paper, we propose novel approaches to regional delineation that address the above limitations. Firstly, we propose a method of topic modelling for extracting interactional regions from new forms of data, i.e. tracking data with no origin and destination specified. Secondly, we extend community detection and topic modelling to uncover interaction regions driven by forces other than spatial proximity. We factor out the effect of space in order to reveal more clearly hidden interaction patterns between places.

We validate our methodology using GPS traces from police patrol vehicles in the London Borough of Camden. Our data are derived from a wider investigation into the local geography of criminal activity and proactive initiatives by police and citizens to reduce crime. The data are particularly relevant to introducing the concepts described in the paper because of the requirement to patrol the all street segments in the study area.

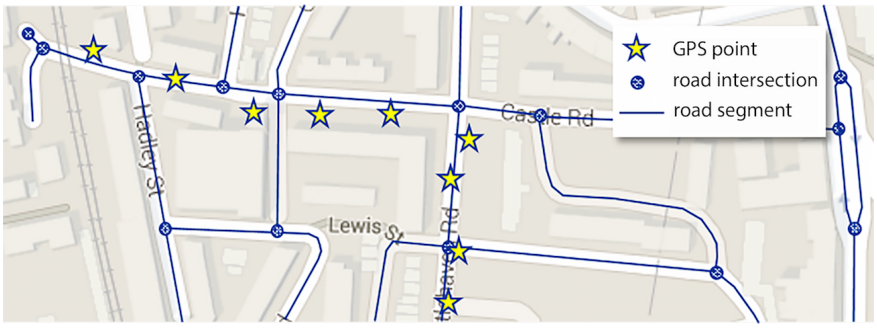


Figure 1.: Illustration of street segments and street intersections.

2. Methodology

Our methodology presents a comprehensive toolkit for extracting interactional regions from large volumes of mobility tracking data in networked environments, such as cities. The data are given as sequences of observations, each corresponding to an episode of mobility or a *journey*. The methodology is motivated by data generated by police patrol vehicles, but is equally applicable to any tracking data that pertain to separable episodes.

At the base of the methodology is a representation of tracking data as a flow network. Interactional regions are extracted as patterns on that network using two clustering approaches: community detection and topic modelling. Community detection assigns locations to regions given flows between *pairs of locations*, hence treating tracking data in an origin-destination fashion, whereas topic modelling mines flow patterns from *location sequences* corresponding to complete journeys. The two methods lead to different definitions of interactional regions, which will be clearly stated in the following subsection.

Finally, we propose novel extensions of community detection and topic modelling that enable us to uncover spatially anomalous interaction patterns. By accounting for spatial forms of cities, we answer the central question motivating this work: can we detect interaction patterns that are not due to space? The obtained *spatially-independent* interactional regions augment the more traditional view on interactional regions obtained from standard community detection and topic modelling techniques (Blondel *et al.* 2010, Manley 2014, Ratti *et al.* 2010) which do not disentangle spatial effects from other effects of interest.

2.1. *Flow Network*

We begin the flow network creation by mapping vehicle traces, in the form of sequences of GPS observations, to the underlying street network. We perform map-matching using the technique of ST-Matching proposed by Lou *et al.* (2009). The technique converts complete GPS traces of vehicle journeys into sequences of visited street segments (see an example GPS trace in Figure 1 and its map-matching output in Table 1). Each street segment is a piece of road, not necessarily straight, between two neighbouring road intersections and is represented by a unique identifier provided by Ordnance Survey (2017).

We construct the flow network by representing street intersections as network nodes (vertices) and vehicle visits to street segments as undirected network edges. This definition allows multiple edges between a pair of nodes, each corresponding to a single visit to the underlying street segment. We remove nodes that have no edges as they do not

No.	GPS ping (easting, northing)	Street segment (id)
1	(529830.826879,182824.079602)	osgb4000000030239207
2	(529901.218657,182821.928288)	osgb4000000031283337
3	(529982.352018,182809.024134)	osgb4000000030250373
4	(530037.990277,182811.410456)	osgb4000000031283336
5	(530101.259824,182812.320039)	osgb4000000031283336
6	(530183.132569,182801.38203)	osgb4000000031283326
7	(530176.143972,182778.158761)	osgb4000000031283326
8	(530181.99883,182739.597688)	osgb4000000031186773
9	(530178.955762,182707.345447)	osgb4000000031186773

Table 1.: Exemplary conversion from a GPS sequence to a sequence of visited street segments.

provide any information on mobility patterns.

2.2. *Interactional Regions as Communities*

Our first approach to interactional region extraction is community detection. In network science, community detection refers to the problem of finding the natural divisions of a network into groups of vertices, called *communities*, such that there are many edges within groups and few edges between groups (Newman 2010).

In our context, community detection mines interactional regions from the flow network as groups of highly interconnected street intersections (nodes). Thus, it leads to the following network-based definition of interactional regions:

Definition 2.1: (Interactional regions as communities) An interactional region is a collection of street segments that have high volumes of traffic flow between them.

2.2.1. Standard Approach

What is meant by "few edges between groups" and "many edges within groups" in community detection is debatable and different definitions have led to a variety of algorithms for community detection. The most common formulation of the problem, and the one adopted in this paper, is of modularity optimisation.

Modularity is a measure of the quality of a network partition, which has a high value when more edges in a network fall *within* rather than *between* communities. In practice, the current state-of-the-art for finding modules in spatial networks is to optimize the standard Newman-Girvan modularity (Guimerà *et al.* 2005, Onnela *et al.* 2011), which assigns vertices to the same community if there are more edges between them than one would expect were edges simply placed at random in the network. In the next section, we will argue that this approach overlooks the spatial nature of the system, or the city as it is in this case.

Modularity is formally defined as:

$$Q = (\text{fraction of edges within communities}) - (\text{expected fraction of such edges}) \quad (1)$$

It considers fractions of edges rather than absolute counts hence it is unaffected by the total number of edges in the network. In mathematical terms, modularity score reads:

$$Q = \frac{1}{2m} \sum_{C \in \mathcal{P}} \sum_{i,j \in C} [A_{ij} - P_{ij}] \tag{2}$$

where $i, j \in C$ is a summation over pairs of nodes i and j belonging to the same community C of a network partition into communities $\mathcal{P}$ and therefore counts edges within communities. A is the adjacency matrix storing the observed number of edges A_{ij} between nodes i and j and P is a matrix storing the expected number of edges between any two nodes. Our estimate of the expected number of edges depends on our null model. The most popular null model, proposed by Newman and Girvan (2004), is:

$$P_{ij} = k_i k_j / 2m \tag{3}$$

where $k_i = \sum_j A_{ij}$ is the degree of node i. Finally, modularity Q is normalized by the total number of edges in the network $m = \sum_{i,j} A_{ij}/2$.

Equation 3 defines that, under the Newman-Girvan (NG) null model, the expected number of edges between any two nodes is proportional to the product of the degrees of the nodes. That is, the more edges a node has, the more likely it is to connect to a different node in the network. Although this definition makes intuitive sense, it overlooks any underlying constraints that might impact on edge formation in spatially-embedded networks, such as spatial distance. We will address this limitation in the following section.

Modularity optimization is a computationally hard problem (Newman 2010). Algorithms that guarantee to find network partitioning with maximum modularity take exponentially long to run and hence are only useful for synthetically small networks (Brandes *et al.* 2007). Instead, therefore, we turn to a *heuristic* algorithm, an algorithm that *approximates* the optimal modularity in an efficient way. We use a popular heuristic algorithm known as the Louvain Method of community detection (Blondel *et al.* 2008).

Louvain Method scales well to large networks and is capable of clustering networks with weighted edges. It is advantageous in that users can easily modify the definition of modularity that it aims to maximise. This characteristic will be particularly useful when we introduce a spatial adaptation of the NG modularity in the next section.

Louvain Method approaches an optimal partition of a network into communities by first assigning each node to a different community and then iteratively merging communities into partitions that increase the overall modularity score Q. The algorithm converges when no further aggregation is found to increase the score.

2.2.2. Spatial Communities

The standard approach to community detection presented in Section 2.2.1 assumes that there are no underlying constraints that could impact on the formation of edges in our network. In other words, any clustering patterns that we observe are of interest to us. Unfortunately, this does not often hold true in practice, where we might want to exclude obvious patterns from those of interest in our analysis.

In our case, these obvious patterns are due to spatial proximity or, more precisely, the configuration of the underlying street network. We are bound to observe movement in our flow network between nodes that in reality are endpoints of the same street segment. On the other hand, we can be quite certain that there will be no direct traffic between nodes that have no connecting road segment. In this section, we propose a way of disentangling

these effects from our clustering results in order to discover interaction relations between places that arise not merely because of spatial adjacency.

We propose the following modification of the NG null model presented in (3):

$$P_{ij} \propto k_i k_j \cdot f(i,j) \quad \text{where} \quad f(i,j) = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are endpoints of the same street.} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The 0/1 function $f(i,j)$ incorporates our knowledge of the underlying street network. If two nodes are endpoints of the same street segment, then we retain the standard expectation proposed in (3) to reflect the fact that the more traffic passes through each node, the higher the chance that some of the traffic will occur between them. By contrast, if they are not directly connected, we reduce the expected flow between them to zero. Notice the proportionality sign in (4): once entries of P are calculated, P has to be renormalised to ensure that the total weight is conserved, i.e. $\sum_{ij} A_{ij} = \sum_{ij} P_{ij} = 2m$.

Intuitively, our proposal works by incorporating our knowledge of the street network into the calculation of the *expected* number of edges between nodes in the flow network. If nodes are endpoints of the same street segment, we expect some traffic between them and our expectation is uniform across all pairs of such nodes. The more traffic we observe, the more likely we are to put the nodes into the same community, i.e. a group of nodes with *higher than expected* traffic between them.

This approach to spatial community detection is inspired by Expert *et al.* (2011), who use a linear function akin to $f(i,j)$ to capture distances in Cartesian space between nodes on a flow network. Here we adapt this approach from Cartesian space to urban space, where interactions are influenced by the connectivity of the underlying street network.

Our spatial modification impacts on the expected number of edges calculated in (2), thus changing the value of modularity Q for any given network partitioning. Despite the change to Q, the same iterative approach as in Section 2.2.1, the Louvain Method, can be used to find a partitioning of the flow network into interactional regions that maximises the modified Q. We implement our spatial community detection with a generalized Louvain Method proposed by Jeub *et al.* (2016).

2.3. *Interactional Regions as Topics*

Our second approach to the extraction of interactional regions is topic modelling. Similar to community detection, this is an approach to finding clusters in data. However, instead of extracting them from pairwise similarities between items, it detects clusters as repetitive themes in unstructured collections of items.

Topic modelling was originally developed to discover main themes that pervade a large collection of documents (Blei 2012). Loosely speaking, it defines a topic as a collection of words concerning a common subject. It assumes that documents can exhibit multiple topics and mines these topics from large collections of documents by detecting groups of words that repeatedly occur together (*co-occur*) in documents.

Since its conception, topic modelling has been adopted to handle many kinds of data, including audio and music, computer code and social networks. Here, we adopt topic modelling to deal with vehicle journey data in order to uncover interactional regions. We assume that vehicle journeys can traverse one or more interactional regions (*topics*). We mine interactional regions from a large collection of vehicle journeys (*documents*), as groups of street segments (*words*) that often occur together in vehicle journeys. This

underpins a second, alternative, definition of interactional regions proposed in this paper:

Definition 2.2: (Interactional regions as topics) An interactional region is a collection of street segments that often co-occur in journeys.

Notice that the above definition of interactional regions differs slightly from Definition 2.1 in Section 2.2. It uses an extra level of information on interactions through the inclusion of complete vehicle journeys. This ensures that even spatially distant street segments in the same interactional region are related, since they often co-occur in vehicle journeys. This does not always hold true for community detection results, where interactional regions are formed based on pairwise relations between nodes. As a result, a *community* region might contain a pair of street segments because of high interactions between their neighbours, neighbours of their neighbours, etc., without any guarantee that a car has ever driven from one node to the other.

2.3.1. Latent Dirichlet Allocation

The most widely used topic model, and the one used here, is Latent Dirichlet Allocation (LDA) proposed by Blei *et al.* (2003). LDA is a *probabilistic model* that belongs to a family of generative probabilistic models. In generative probabilistic modelling, we treat our data as arising from a generative process that includes *unobserved (hidden) variables*. This generative process defines a *joint probability distribution* over the observed and the hidden random variables. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables. This conditional distribution is also called the *posterior distribution*.

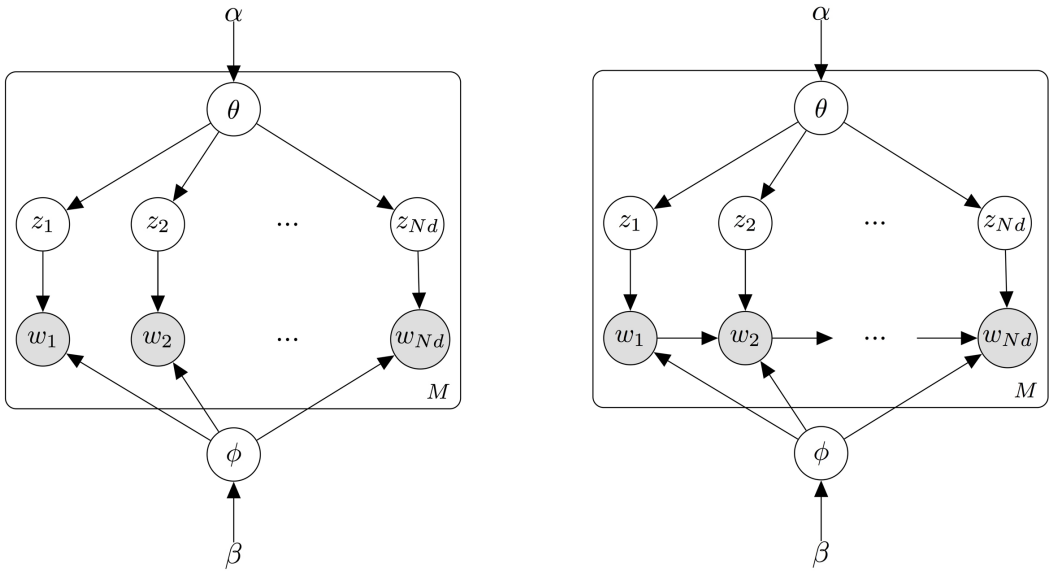
LDA attempts to capture the notion that documents exhibit multiple topics. It defines a generative process from which documents could have arisen. It states that the observed variables are the words of the documents; the hidden variables are the topics; and the generative process is as described here. The computational problem of inferring the hidden topic structure from the documents is addressed by the collapsed Gibbs sampler developed by Griffiths and Steyvers (2004). It is a sampling-based algorithm that approximates the posterior distribution by a finite number of samples from it.

LDA is described more formally with the following notation. There are K pre-defined topics, each taking a probabilistic distribution over a fixed vocabulary. When documents $D = \{d_1, d_2, ..., d_M\}$ are generated, a topic mixture θ for each document is sampled, with $\theta_{d,k}$ indicating the topic proportion of topic k in document d, from a Dirichlet distribution with prior α. Subsequently, topics $Z_D = \{z_1, z_2, ..., z_{Nd}\}$ for each word in the document are sampled from that mixture. Finally, based on the sampled topics, words $W_D = \{w_1, w_2, ..., w_{Nd}\}$ are chosen from the topics' distributions ϕ over the vocabulary, where ϕ_k is the distribution of topic k over the vocabulary, sampled from a Dirichlet distribution with prior β. The graphical model for LDA is illustrated in Figure 2(a).

Our core contribution is to adapt LDA to regional delineation problems by interpreting topics as interactional regions. The topics are inferred from large collections of vehicle traces (*documents*), where each trace is a sequence of visited street segments (*words*).

2.3.2. Spatial Topics

Conventional topic modelling is not well designed for spatial data. This is because the basic technique is not attuned to the nature of spatial data and hence cannot disentangle spatial patterns from other patterns that might be of more interest to the user. What is more, one of the assumptions of the core technique is that consecutive words within a document are independently sampled under the 'bag-of-words' assumption. These characteristics limit the usefulness of topic modelling for geographic problems in general and



(a) LDA (b) Token-Bigram LDA

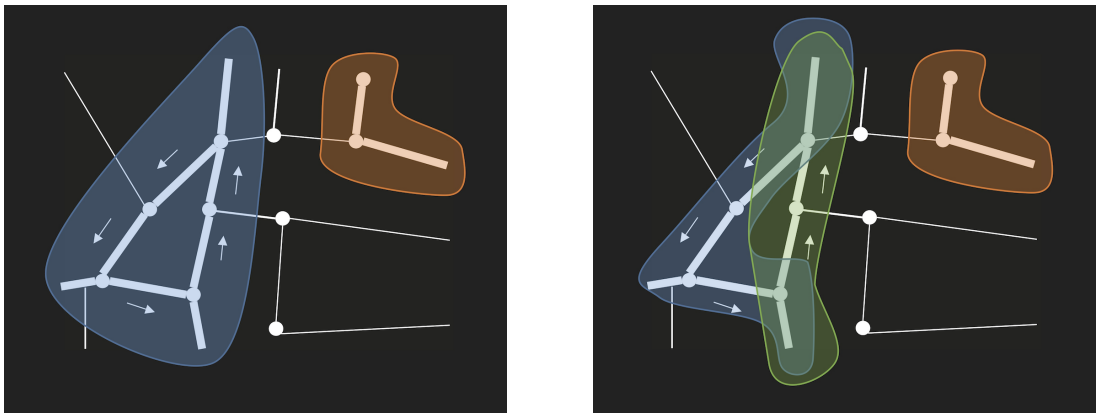
Figure 2.: Graphical illustration of topic models for interactional region extraction, assuming (a) independence and (b) dependence between consecutive street segments w_i and w_{i+1} in observed journeys. Each node is a random variable and each edge indicates statistical dependence between the variables. The rounded rectangles denote replication for all M vehicle traces.

interactional region extraction in particular. Our documents, vehicle traces, are inherently spatial and sequential. We thus need to incorporate these qualities in the model in order to identify any significant patterns in the data.

Our contribution is to accommodate the properties of spatial data by introducing the notion of dependence between consecutive words in the generative process captured by LDA (see Figure 2(b)). Such dependence has previously been proposed by Barbieri *et al.* (2013) as a token-bigram topic model, in which the dependence is interpreted as a transition probability $p(w_{i+1}|w_i)$, i.e. given an occurrence of word w_i, how likely is it that word w_{i+1} will occur next in the document sequence? In our case study, the transition probability is derived directly from the branching of the underlying street network. That is, given that a vehicle is on street segment w_i, how likely it is to move to street w_{i+1}? The probability is zero for non-adjacent street segments and inversely proportional to the number of street segments adjacent to w_i, k_{w_i}, otherwise:

$$p(w_{i+1}|w_i) = \begin{cases} 1/k_{w_i}, & \text{if } w_i \text{ and } w_{i+1} \text{ are adjacent.} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This modification requires a more universally-applicable inference algorithm than the collapsed Gibbs sampler used for standard LDA, as introduced in the previous section. We instead use the 'universal inference engine' implemented in the STAN probabilistic programming language (Stan Development Team 2016). The engine is based on a Gibbs sampler with a 'no-u-turn' extension (Homan and Gelman 2014) that uses adaptive parametrisation to eliminate the need of manual parameter tuning.



(a) Interactional regions as communities

(b) Interactional regions as topics

Figure 3.: Interactional regions in a synthetic case in which northward and southward vehicle journeys always follow distinct routes (arrowed). White lines represent street segments, their thickness is proportional to the number of visits to a segment, i.e. the number of edges in the flow network. Since community detection (a) does not consider journeys holistically, it groups northward and southward routes into a single interactional region. On the contrary, topic modelling (b) makes a distinction between the two routes and hence discovers an additional interactional region.

The proposed dependence term serves a similar purpose to the function $f(i, j)$ in the spatial community approach in (4). It enables spatial knowledge of the underlying street network to be accommodated while also removing the 'bag-of-words' assumption from the LDA model. The modified LDA can thus be used to detect interactional regions as collections of street segments that co-occur in vehicle journeys more often than expected based on their proximity in the underlying street network.

2.4. *Summary*

Our methodology employs two popular clustering approaches, community detection and topic modelling, to extract interactional regions from large amounts of digital vehicle traces. At its core, both methods are designed for episodic mobility and interaction data. However, they differ in their measure of interaction between locations (Definitions 2.1 and 2.2) which leads to differences in regional delineation. Definition 2.1 focuses on interactions between pairs of locations and hence is better suited for origin-destination data, whereas Definition 2.2 looks at interactions over sequences of locations and hence is better attuned to episodic mobility data with no functionally defined origins and destinations. When applied to mobility tracking data, the differences between the two methods are exemplified in Figure 3.

We attune both techniques to spatial data analysis by changing their *expected* measure of interaction between locations, i.e. before observing any vehicle traces. By default, in both methods the expected level of interaction is uniform for all pairs of street nodes. We customise it to reflect the structure of the underlying street network instead, e.g. non-adjacent street nodes are expected to have zero traffic directly between them (see Equations 4 and 5). The adaptations enable capturing interaction patterns that are not merely a result of the spatial arrangements of streets.

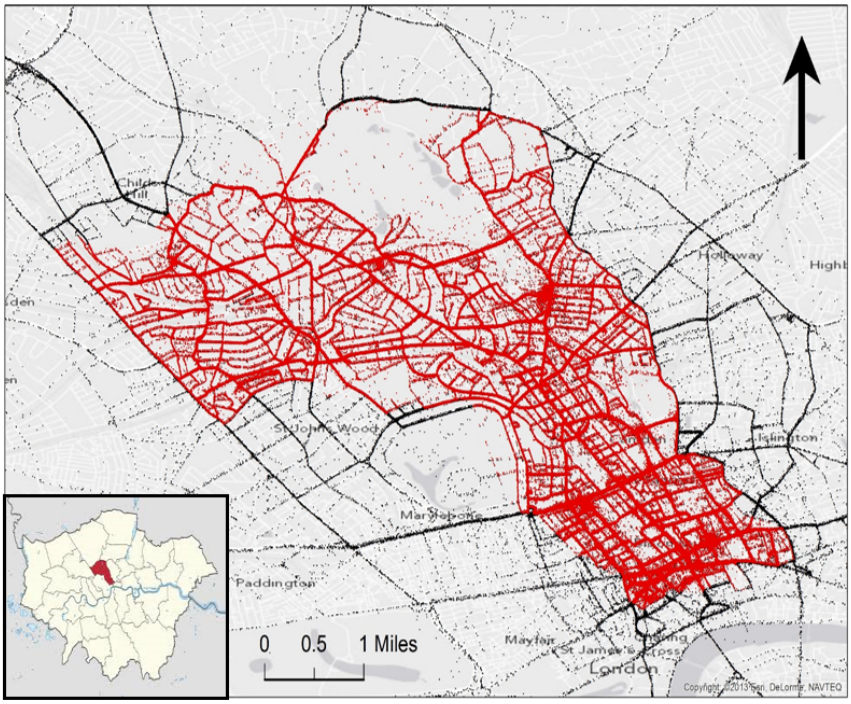


Figure 4.: GPS signals transmitted by police vehicles inside (red) and outside (black) the London Borough of Camden in March 2011.

3. Numerical Validation

We validate the proposed methodology on police patrol tracking data. Unlike other episodic activities, such as mail delivery or shopping, police patrolling is expected to take place in every part of the neighbourhood. The granularity of spatial coverage makes them particularly suitable for validation of a regional delineation methodology such as ours.

3.1. *Police Patrol Data*

The police patrol data are a complete set of GPS signals transmitted by police patrol vehicles during March 2011 in the London Borough of Camden, a borough in Central London with the total area of 21.8 km^2. The dataset comprises a total of 1,188,953 GPS signals from 5,513 journeys (see Figure 4). It was acquired for research purposes as part of the "Crime, Policing and Citizenship" project in collaboration with the Camden Metropolitan Police*. The flow network generated from the dataset is shown in Figure 5.

3.2. *Interactional Regions as Communities*

First, we extract interactional regions from police tracking data using community detection (see Definition 2.1). Standard community detection uncovers seventy-six interactional regions shown in Figure 6. The regions are small relative to the study area and are

*UCL Crime Policing and Citizenship: http://www.ucl.ac.uk/cpc/.

Figure 5.: Police flow network of Camden. Colour intensity of each street segment is proportional to the number of police vehicle journeys in March 2011.

strongly clustered in space, which indicates that police patrol activities might be dominated by short-distance journeys. When we account for spatial factors according to (4), the resulting interactional regions (see Figure 7) are no longer spherical but somewhat elongated and follow stretches of individual roads. They are also relatively small but we have no influence over region sizes when using community detection, which only outputs a single partitioning corresponding to maximal modularity in (2).

The differences between standard and spatial communities are intuitively plausible. Roads that are in close proximity to each other are likely to distribute high amounts of local, within region, traffic. These short-ranged interactions dominate interactional regions uncovered by standard community detection (as shown in Figure 6), indicating that spatial proximity plays a major role in their formation. When we incorporate spatial effects according to (4), we can focus on long-ranged interaction patterns instead. These rather follow long stretches of major roads (e.g. yellow community in Figure 7(b)) as high category roads attract more traffic than one would expect just based on spatial proximity.

The standard and spatial communities seem to reflect different modes of police patrolling. According to the wider literature (Chen *et al.* 2017) and our knowledge gathered through working closely with the Camden Police, police patrols can be roughly divided into *routine* and *emergency* patrols. The former is a form of preventive policing that require police to regularly visit crime hotspots, i.e. small geographical units with high

crime intensity, such as street segments or small groups of street blocks (Braga *et al.* 2014). This mode of behaviour is spatially clustered and hence well suited for standard community detection, as shown in Figure 6. The latter is a reactive policing effort that requires police vehicles to reach crime scenes as quickly as possible. Emergency patrolling relies heavily on major roads as they enable reaching distant crime scenes in a short amount of time. The most popular long-distance police routes are well depicted by spatial community detection in Figure 7. Note that these observations remain speculative, however, since there is no ground truth on what police officers actually did during their patrol journeys.

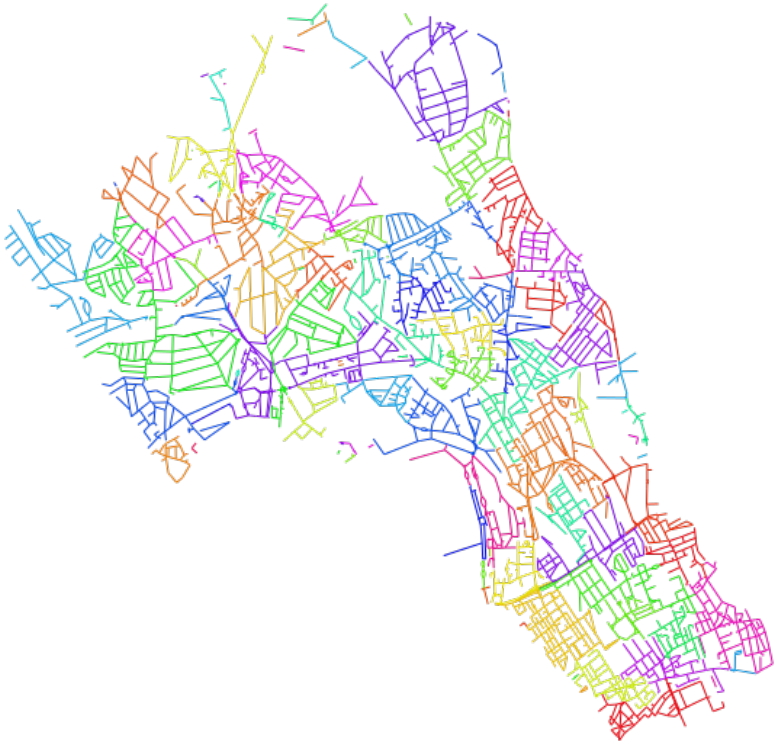
3.3. *Interactional Regions as Topics*

Second, we analyse interactional regions discovered using topic modelling. Similar to standard community detection, standard topic modelling produces interactional regions which are strongly determined by geographical factors. In contrast to community detection, however, topic modelling does not only show the *optimal* partitioning, but instead enables viewing interactional regions at multiple scales by varying the number of topics K that we fit to the data.

We show interactional regions at different scales in Figure 8. The larger they are, the more spatially constrained they become. This again reflects the fact that police journeys are predominantly local and thus short-scale interactions dominate any large-scale analysis. When very small interaction regions are chosen, topic modelling is capable of uncovering non-trivial interaction patterns such as long road stretches in Figure 8(b). The ability to uncover both gravity-like and other less-trivial interaction patterns by varying the parameter K puts topic modelling at a significant advantage to community detection. In contrast to community detection, topic modelling considers complete journeys when detecting functional relations. Since journeys tend to be longitudinal, so are shapes of the extracted interactional regions. The smaller the interactional regions, the subtler the routing choices they reflect.

Topic modelling sometimes leads to disconnected parts of the street network being identified as members of the same interactional region. This rather undesirable characteristic, visible as multiple subgraphs with the same colour in Figures 8 and 9, could be addressed by a similar probabilistic model for clustering, called a block model (Parkkinen *et al.* 2009). Block model replaces the 'bag-of-words' assumption of topic modelling with a 'bag-of-pairs-of-words' assumption that places more emphasis on clustering connected parts of the network together. This could be investigated in future work.

When we modify topic modelling to account for the underlying street network connectivity according to (5), we detect interactional regions with almost no spatial compactness. Even at very low resolution in Figure 9, they are rather stretches of roads than local neighbourhoods. The stretches often reappear in police journeys as they connect locations of mutual functional importance. In this case, they seem to be the roads connecting police stations (see Figure 9(b)). In contrast to spatial community detection, where most interactional regions contain as few as $\sim$10 street segments (see Figure 10(b)), spatial topic modelling can uncover interactional relations at much larger distances and of generally larger sizes (Figure 10(e)). These characteristics suggest superiority of topic modelling over community detection as a method for extracting interactional regions from episodic tracking data, such as police patrol data.



(a) all communities

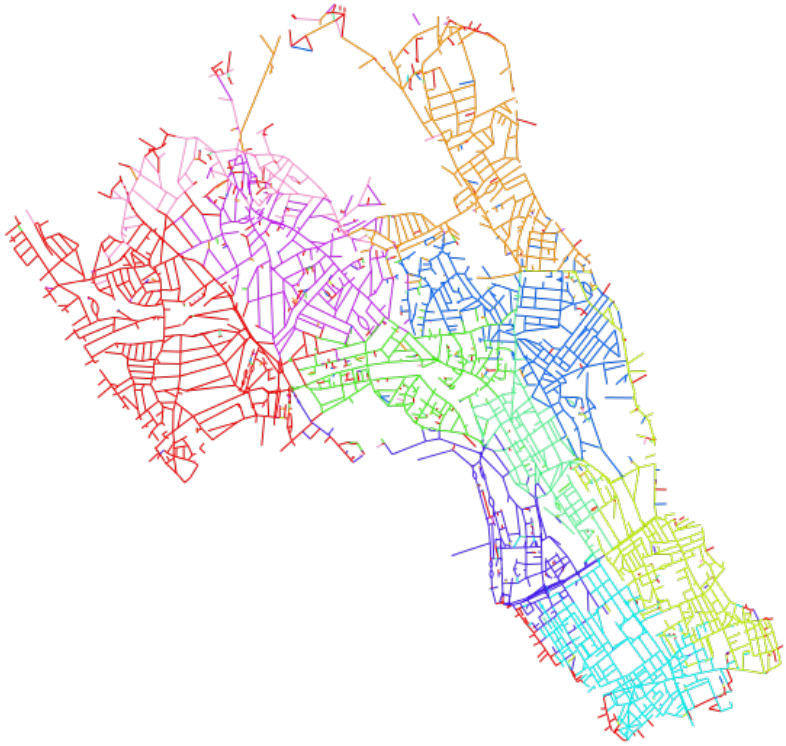
(b) largest communities

Figure 6.: Interactional regions as communities.

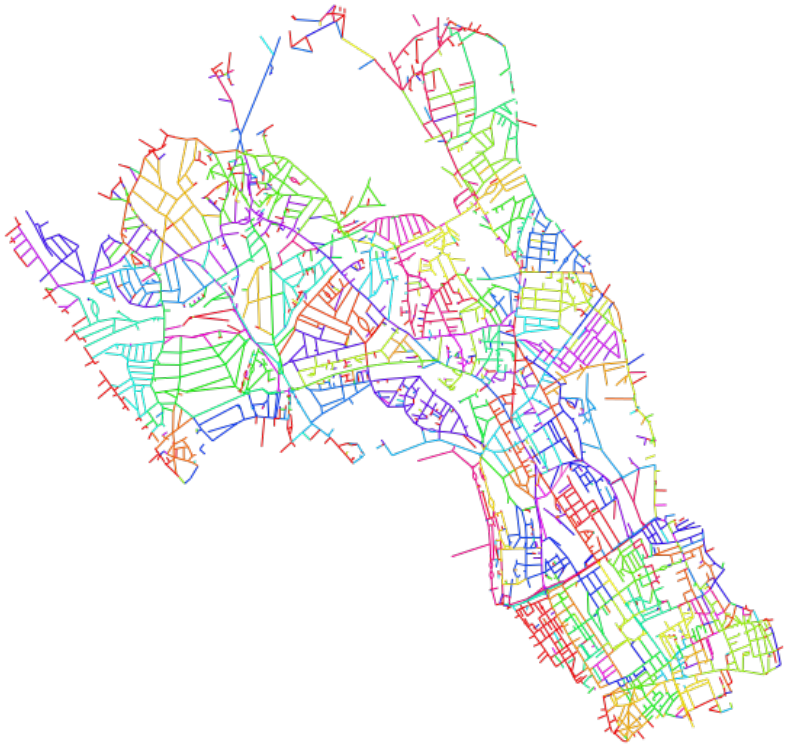
(a) all communities

(b) largest communities

Figure 7.: Interactional regions as spatial communities.

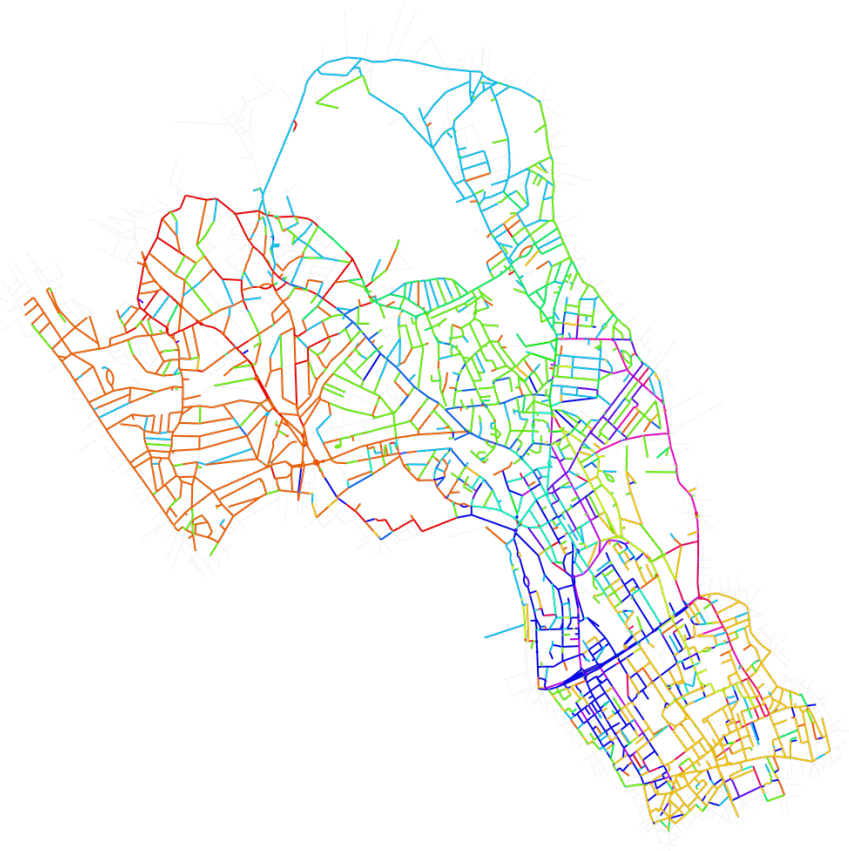


(a) 10 topics



(b) 150 topics

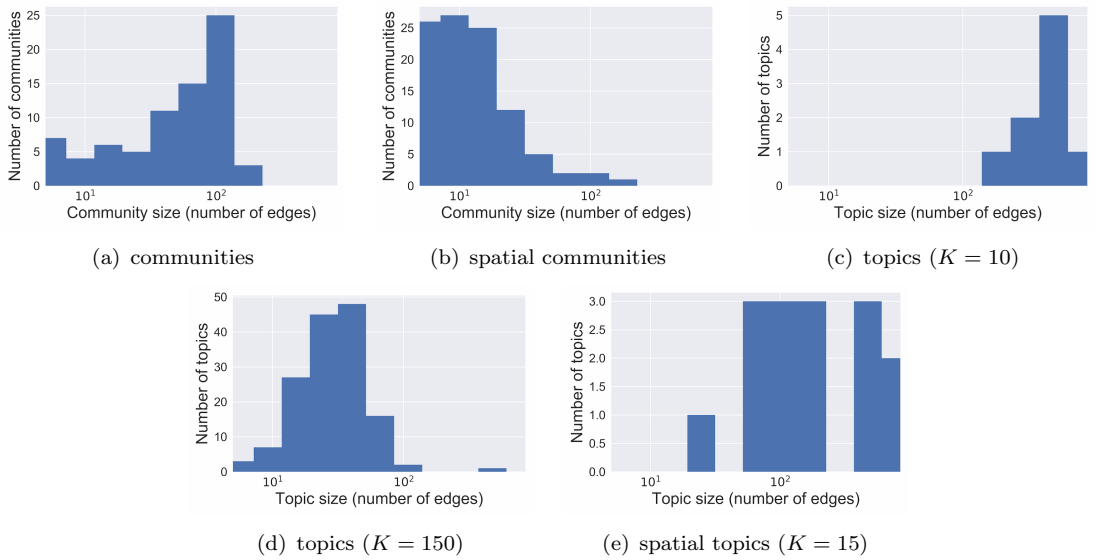
Figure 8.: Interactional regions as topics at different scales.



(a) all topics

(b) selected topics (with locations of police stations marked)

Figure 9.: Interactional regions as spatial topics.



(a) communities (b) spatial communities (c) topics ($K = 10$)

(d) topics ($K = 150$) (e) spatial topics ($K = 15$)

Figure 10.: Size distribution of international regions as (a) communities, (b) spatial communities, (c) 10 topics, (d) 150 topics, (e) spatial topics. Region size is equal to the number of street segments it contains. Note that region size is given in logarithmic scale on x-axis.

3.4. *Methods Comparison*

So far, our validation has focused on *qualitative* comparison of interactional regions uncovered with the proposed methods. Our focus now shifts to *quantitative* analysis to answer questions such as: how much different are the results from the different proposed methods? How can we measure their quality? We compare results from both the proposed methods, community detection and topic modelling, and their variants, standard and spatial.

We address the question of differentces between regional deineations by using adjusted mutual information score (MI) (Xuan Vinh *et al.* 2010), which is a measure of distance between different regional partitions. MI is equal to one only when two partitions are identical and is between 0 and 1 otherwise. Results are summarized in Figure 11 where we observe that interactional regions obtained from topic modelling and community detection are genuinely different (MI <= 0.35). Their largest difference is between spatial communities and spatial topics (0.06). The dissimilarity is not surprising since the two methods follow different definitions of what interactional regions are (Definition 2.1 and Definition 2.2). On the contrary, interactional regions coming from the same method show slightly higher similarity, such as topics (K=10) and topics (K=150) (0.37). Another interesting point is that spatial methods lead to very different regional delineations to standard methods (0.35 for spatial communities vs. communities; 0.23 for spatial topics vs. topics). This difference is already acknowledged in the previous two sections, where we notice that spatial methods produce regions that are much less spatially clustered than the corresponding standard methods. By design, the difference should arise from the fact that spatial methods remove the effects of spatial proximity when looking for regional delineation.

We confirm whether the differences between spatial and non-spatial methods in fact arise from their treatment of spatial adjacency by performing a randomisation test. The test randomly shuffles the geographical position of the nodes in the flow network while

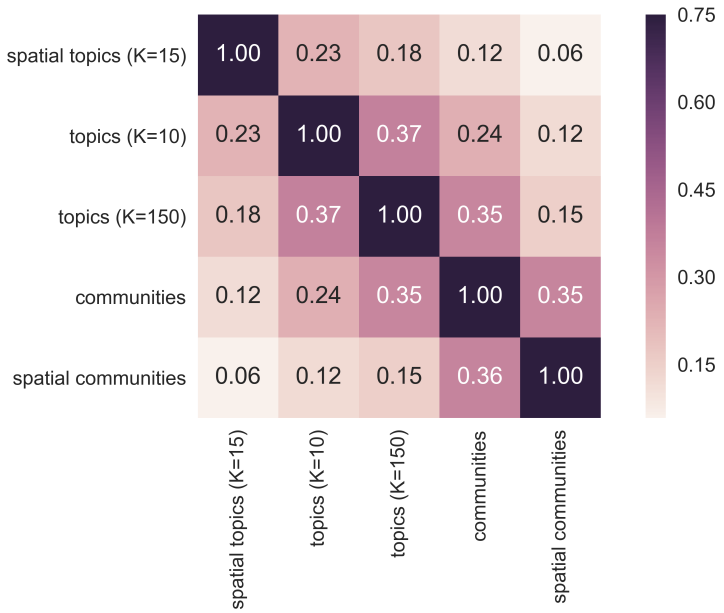


Figure 11.: Mutual information score between regional partitions obtained with the proposed methods.

	Original-Random
Spatial Topics ($K = 15$)	0.01613398 ± 0.02
Spatial Communities	0.35651846 ± 0.03

Table 2.: Average MI measured between the regional partition found on the original flow network and 100 randomized networks (Original-Random) for spatial topic modelling and spatial community detection.

keeping edges between them unchanged. As a result, in the randomised network, we observe the same volume of traffic between pairs of nodes but the traffic can now occur between nodes that are not connected by a street segment in reality (thus violating our spatial adjacency assumptions in (4) and (5)). The randomised network is no longer embedded in the underlying street network, but this has no effect on regional partitions uncovered with *standard* community detection and topic modelling, since these methods do not make use of the street arrangement information. On the contrary, *spatial* community detection and topic modelling, which assume the spatial embedding of the network, uncover partitions that are largely different from the ones they find in the real, spatially-embedded, flow network (see Table 2). Interestingly, the largest variation is shown by spatial topic modelling, where partitions found on the random and the real network show almost zero resemblance (0.016).

The randomisation test could be further extended to measure significance of the discovered partitions as in the work by Expert *et al.* (2011). This would bring us close to answering the second question posed: can we measure quality of the discovered partitions? The extension could constitute future work but it would, nonetheless, not fully answer the question of partition quality, which is conceptually difficult to answer in the lack of ground truth on the correct interactional region delineation. If ground truth was

available, we could assess the quality of different partitions by measuring their similarity to the ground truth using the mutual information score, as shown for different partitions in Figure 11. Since ground truth does not exist, we can only approximate partition quality through significance testing (as suggested above) or through usefulness of the discovered partition in a specific context, e.g. "do the discovered regions lead to increased accuracy in region-based route choice simulations?" as explored in the work by Kowalska *et al.* (2015), or "are the discovered partitions more stable across time?". Answers to these questions could create interesting extensions to this paper but are outside the scope of this work.

4. Conclusions and Directions for Future Research

In this paper, we have presented a comprehensive network-based methodology for extracting interactional regions from digitised vehicle traces in urban environments. The methodology used a large dataset of GPS vehicle traces to define a road traffic network and then uncovered interactional regions as densely connected areas within the network. It considered two approaches to the discovery of interactional regions: community detection and topic modelling. Community detection used aggregated traffic flows between pairs of street nodes when assigning them to regions. Topic modelling instead considered sequences of street nodes corresponding to complete vehicle journeys, hence potentially giving a more complete picture of how drivers perceive the urban space. The techniques were adapted to account for the effect of space upon the network topology, hence uncovering interaction patterns between places that arose not solely from spatial proximity.

Both community detection and topic modelling could detect short-ranged interaction patterns in police patrol data, suggesting that spatial proximity is a major force in spatial interactions. By adopting the methods to focus on spatially-anomalous interactions only, they could also uncover less-trivial long-ranged interaction patterns. Therefore, the proposed methodology, including both standard and spatial adaptations of the methods, provides a more detailed insight into interaction patterns than previously proposed using standard community detection only (Blondel *et al.* 2010, Manley 2014, Ratti *et al.* 2010). Our initial analysis suggested two advantages of topic modelling over community detection for episodic activity data. Firstly, it could detect either aggregate or granular activity patterns by varying a single parameter K. Secondly, it was capable of detecting longer activity trails, such as paths between police stations. On the negative side, however, topic modelling could lead to disconnected parts of the street network being classified as members of the same interactional region.

We validated our methodology using police patrol data due to their availability and high spatial granularity. We discovered interactional regions which seemed to correspond to two modes of police patrols: routine and emergency patrols. Spatially clustered interactional regions bounded routine patrol activities, restricted to neighbourhoods and minor roads, whereas elongated interactional regions corresponded to popular major road routes leading to police stations or emergency calls. We uncovered distinguishable patrolling preferences that could potentially be used in the design of effective police districts, especially in light of recent funding cuts and multiple London police districts being merged.

Our comprehensive methodology extends beyond police patrol data. It is applicable to vehicle flows in general, as well as other episodic flows, e.g. cyclist or pedestrian

journeys. As such, the methodology opens new avenues of quantitative analysis of urban dynamics. Depending on the dataset analysed, it can discover regional partitions that are case study-specific (e.g. regions extracted from *police* journeys) or more generalizable to the wider city population (e.g. regions extracted from *all* vehicle journeys).

Future research could extend the proposed methods to account for subtler spatial effects, such as differences in flows on major and minor roads. It could also validate the methods further using other city flow data or by using the discovered regions within a specific application, such as region-based route choice simulation frameworks (see Manley *et al.* (2015) as an example).

Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) are greatly appreciated.

We would also like to show our gratitude to Dr Ed Manley for very helpful discussions during the course of this research.

References

Barbieri, N., *et al.*, 2013. Probabilistic topic models for sequence data. *Machine Learning*, 93 (1), 5–29.

Batty, M., 2013. *The New Science of Cities*. MIT Press.

Besussi, E., *et al.*, 2010. The Structure and Form of Urban Settlements. *In*: T. Rashed and C. Jürgens, eds. *Remote Sensing of Urban and Suburban Areas*. Springer Netherlands, 13–31.

Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, 55 (4), 77.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Blondel, V., *et al.*, 2010. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42.

Blondel, V.D., *et al.*, 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10).

Braga, A.A., Papachristos, A.V., and Hureau, D.M., 2014. The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis. *Justice Quarterly*, 31 (4), 633–663.

Brandes, U., *et al.*, 2007. *In*: *On Finding Graph Clusterings with Maximum Modularity.*, 121–132 Berlin, Heidelberg: Springer Berlin Heidelberg.

Brockmann, D., 2010. Following the money. *Physics World*, 23 (02), 31–34.

Chen, H., Cheng, T., and Wise, S., 2017. Developing an online cooperative police patrol routing strategy. *Computers, Environment and Urban Systems*, 62, 19–29.

Expert, P., *et al.*, 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.

Fotheringham, A.S., 1997. Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21 (1), 88–96.

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (1), 5228–35.

Guimerà, R., *et al.*, 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles.. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (22), 7794–9.

Homan, M.D. and Gelman, A., 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15 (1), 1593–1623.

Jeub, L.G.S., *et al.*, 2016. A generalized Louvain method for community detection implemented in MATLAB. [online] [2017-07-22].

Karlsson, C. and Olsson, M., 2006. The identification of functional regions: theory, methods, and applications. *The Annals of Regional Science*, 40 (1), 1–18.

Kowalska, K., Shawe-Taylor, J., and Longley, P., 2015. Data-driven modelling of police route choice. *In*: *Proceedings of the 23rd GIS Research UK conference (GISRUK 2015)*, Leeds.

Lou, Y., *et al.*, 2009. Map-matching for low-sampling-rate GPS trajectories. *In*: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, nov. New York, New York, USA: ACM Press, p. 352.

Manley, E., 2014. Identifying functional urban regions within traffic flow. *Regional Studies, Regional Science*, 1 (1), 40–42.

Manley, E., Orr, S., and Cheng, T., 2015. A heuristic model of bounded route choice in urban areas. *Transportation Research Part C: Emerging Technologies*, 56, 195–209.

Newman, M. and Girvan, M., 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69 (2), 026113.

Newman, M.E.J.M.E.J., 2010. *Networks : an introduction*. Oxford University Press.

Onnela, J.P., *et al.*, 2011. Geographic Constraints on Social Network Groups. *PLoS ONE*, 6 (4), e16939.

Ordnance Survey, 2017. OS MasterMap Integrated Transport Network Layer. [online] [2017-07-22].

Parkkinen, J., Gyenge, A., and Sinkkonen, J., 2009. A block model suitable for sparse graphs. *In Proceedings of the 7th International Workshop on Mining and Learning with Graphs*.

Ratti, C., *et al.*, 2010. Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE*, 5 (12), e14248.

Simini, F., *et al.*, 2012. A universal model for mobility and migration patterns. *Nature*, 484 (7392), 96–100.

Singleton, A.D. and Longley, P.A., 2009. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29 (3), 289–298.

Stan Development Team, 2016. *Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0*. Technical report.

Thiemann, C., *et al.*, 2010. The Structure of Borders in a Small World. *PLoS ONE*, 5 (11), e15422.

Vanhove, N., 1999. *Regional policy: a European approach*. 3 Ashgate: Ashgate Publishing Limited.

Xuan Vinh, N., Julien Epps, U., and Bailey, J., 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11, 2837–2854.