Data-driven modelling of police route choice

Kira Kowalska[*1], John Shawe-Taylor[†2] and Paul Longley[‡3]

[1]Department of Security and Crime Science, University College London
[2]Department of Computer Science, University College London
[3]Department of Geography, University College London

November 4, 2014

Summary

In recent years, increasing digitisation of police patrol activities has enabled new insights into police patrol behaviour. This paper explores digitised traces of police patrol journeys in order to understand police routing preferences and to propose models that enable simulations of patrol behaviour. The models assume that police journeys are undertaken as series of "topics", which are inferred from vehicle GPS data using a widely used topic modelling technique called *latent Dirichlet allocation*. Initial experiments have shown that they are capable of reproducing police coverage patterns that would not be captured by alternative models assuming optimal behaviour.

KEYWORDS: route choice, topic modelling, GPS data, police vehicles

1. Introduction

In recent years, increasing digitisation of police patrol activities has brought new opportunities and challenges to policing. Not only has it enabled evaluation of current patrol strategies, but also a shift to more effective, data-driven approaches.

This research project analyses digitised traces of police patrol vehicles to propose data-driven models of police movements around the city. The models give explanations about police behaviour and enable simulations of police movement. Hence, they could aid the development of new patrol strategies and serve as an evaluation tool of police work on patrol.

The project is one of the first attempts to develop data-driven models of police patrol movement. This research gap arises mainly because of non-availability of data capturing detailed police movement around the city. The case study for this research will be the London Borough of Camden, for which GPS traces of police vehicles have been recently released for research purposes.

2. Methods

Modelling police vehicle movements is based on the assumption that police plan their journeys in stages, taking their preferred routes through neighbourhoods en route to destination. This assumption is in line with research into vehicle route choice undertaken by Manley (2013) who showed that drivers' behaviour is rather suboptimal and that their "route selection takes place in phases, linking locations and decision points on route to destination".

[*] kira.kowalska.13@ucl.ac.uk
[†] j.shawe-taylor@ucl.ac.uk
[‡] p.longley@ucl.ac.uk

The preferred routes are inferred as 'topics' from vehicle GPS data using *latent Dirichlet allocation*, a widely used topic modelling technique. Police journeys are then simulated as sequences of the inferred topics.

2.1. Topic modelling

Topic modelling techniques were originally developed to discover main themes that pervade a large collection of documents (Blei et al., 2003). At the moment, they are being adopted by other disciplines as well (e.g. population genetics (Pritchard, Stephens, & Donnelly, 2000), image classification (Fei-Fei & Perona, 2005)) with the purpose of finding underlying themes in unstructured collections of items.

In this project, we attempt to develop a novel application of topic modelling to route choice modelling. Although limited previous research used topic models to discover mobile behavioural patterns (Huynh, Fritz, & Schiele, 2008)(Farrahi & Gatica-Perez, 2012), no research has applied topic models to *vehicle GPS data* in particular. On top of that, no research has adopted the discovered patterns for movement modelling. Therefore, the work presented in this report, despite its premature state, could potentially contribute to the research community not only by applying topic modelling to a new type of data, but also through an innovative use of the discovered topics in vehicle route choice modelling.

The most common topic modelling algorithm and the one used in this project is *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). In the context of document analysis, it defines a topic as a distribution over a fixed vocabulary. It assumes that each document exhibits multiple topics and that words contained in the document are generated by firstly randomly picking a topic from the document's distribution over topics and then by randomly choosing a word from the topic's distribution over the vocabulary. Parameters defining the generative process are tuned to the data using a sampling technique called *Gibbs Sampling* (Steyvers & Griffiths, 2007).

Our research project requires novel interpretations of topics, documents and words. In our context of vehicle movement, words become street segments and documents become vehicle journeys (collections of visited street segments). Discovered topics are distributions over street segments. The topics reveal underlying clusters of street segments based on their co-appearance in vehicle journeys and hence enable a simplified representation of vehicle journeys as distributions over the topics.

2.2. Route choice modelling using topics

Topics inferred from the data are used to model police vehicle movement. The original street network is reduced to a *topic network*, in which each *topic node* is a collection of street segments that have the highest probability of appearing in that topic. Topic nodes are connected by an edge if their street segments are physically connected. Vehicles that want to travel from one street segment to another start at the topic node where the first segment is assigned and take the shortest path through the topic network from that node to the topic node containing their journey destination. This modelling framework reflects the intuition that journeys are undertaken in stages or as series of topics on route to the destination.

Two variants of the modelling framework are explored. In the first one, the topic network is *unweighted* and vehicles choose paths that minimise the number of topics traversed to their destination. In the second one, topic nodes are *weighted* by the total length (in metres) of street segments assigned to them. Vehicles subsequently choose paths that minimise journey length.

2.3. Model validation

The models are validated against the data by measuring the *Pearson correlation coefficient* (Agresti, 2008) between street coverage generated by the models and the actual street coverage, where coverage is defined as the number of vehicle visits at each street segment in Camden. The generated street coverage contains journeys between all possible origins and destinations in the Camden's street network, weighted by the probability of observing such an origin-destination pair in the actual data.

Since the models are probabilistic, the generated coverage at a street segment is represented by the sum of probabilities of all possible journeys that would visit that segment. An alternative approach would be to simulate agents (vehicles) based on the probabilistic rules and then use their journeys to calculate the generated coverage. The alternative approach would be prone to sampling bias though, especially if the number of agents was small.

3. Results and discussion

3.1. Police Vehicle Data

Police vehicle data motivating the project were released for research purposes in May 2012 as part of the "Crime, Policing and Citizenship" project[§]. The data include all GPS signals transmitted by police vehicles in the London Borough of Camden in the months of March 2010 and March 2011 (1,188,953 GPS signals in total). The frequency of GPS signal transmissions is roughly every 15 seconds.

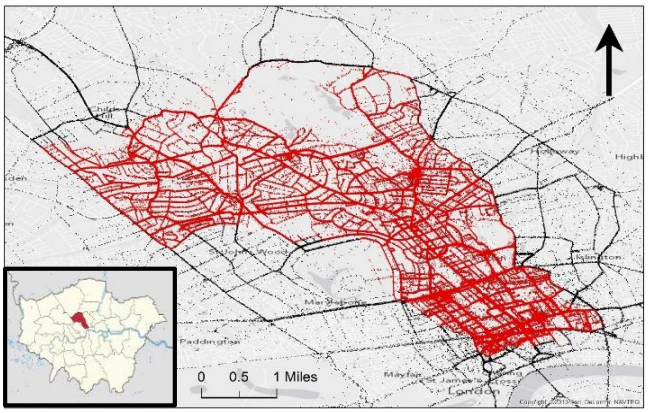

Figure 1. GPS signals transmitted by police vehicles inside (red) and outside (black) the London Borough of Camden in March 2010.

Police route choice preferences observed in the data are sub-optimal in terms of journey legnth as shown in Figure 2 and strongly biased towards the use of major roads (Figure 3). These patterns are in line with the findings by Manley (2014) that underpin our models.

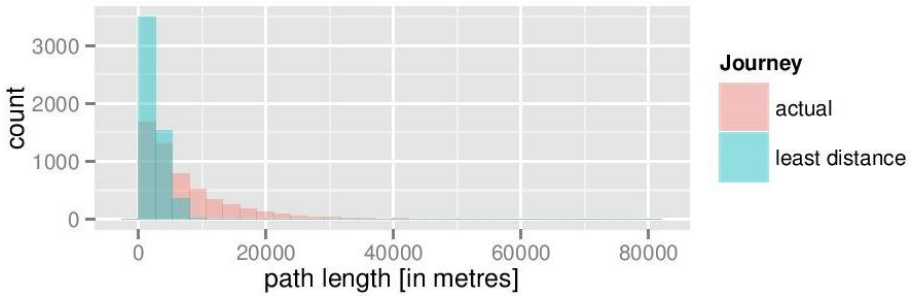

Figure 2. Distributions of lengths of journeys in March 2010 and their least distance alternatives.

[§] UCL Crime Policing and Citizenship: http://www.ucl.ac.uk/cpc/.

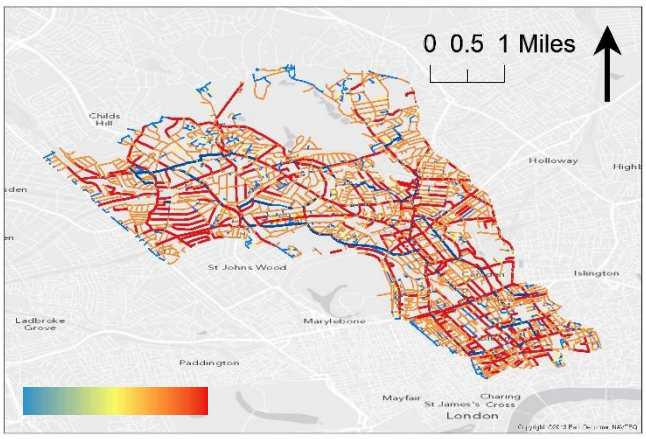

Figure 3. Difference between actual and least distance journeys (*actual* minus *least distance*) in March 2010; yellow corresponds to exact match.

3.2. Route choice modelling

Topics inferred from the data are shown in Figure 4. Latent Dirichlet allocation algorithm required specifying the number of topics *a priori* and this number was set to hundred following initial experiments into the influence of the number of topics on topic sizes. However, it is acknowledged that further research is required into optimising the number of topics for modelling purposes.

The assignments of segments to topics in Figure 4 (left) seem to reflect their network proximity, as well as a general road hierarchy. Segments of major roads tend to be clustered together forming 'stretched' topics, whereas segments of minor roads seem to be clustered within their neighbourhoods. These observations reflect the intuition that vehicles tend to travel longer distances along major roads but otherwise limit their journeys to nearby locations.

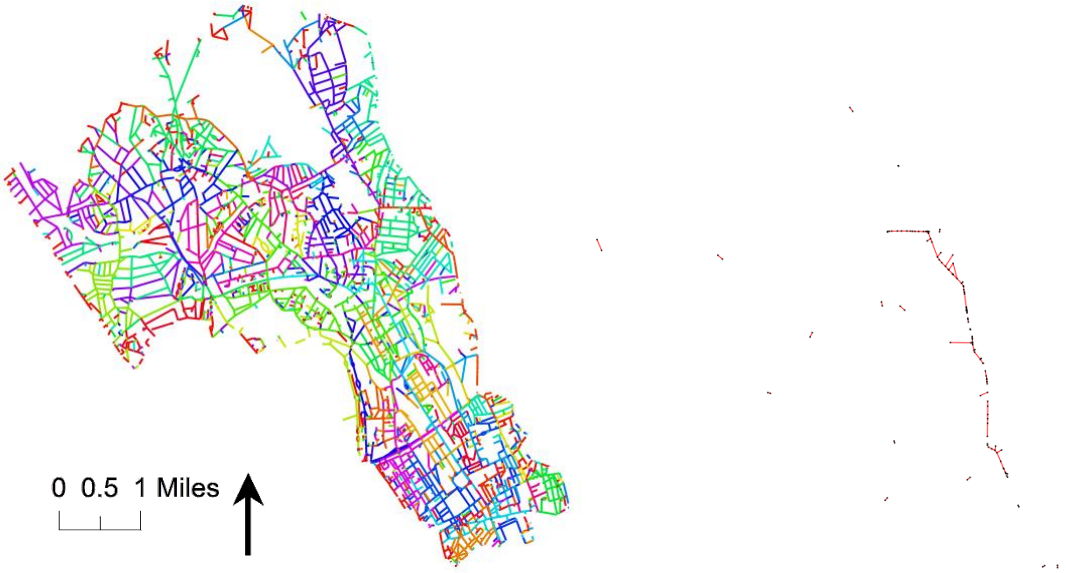

Figure 4. (left) Camden's street network coloured by topic assignments, (right) an exemplary topic, when hundred topics are inferred from police journeys in March 2010 in the London Borough of Camden.

The inferred topics are used to create a topic network. This proves to be problematic as a closer investigation of topics in Figure 4 reveals that topics are often disconnected (see example topic in Figure 4 (right)). The discontinuity might be due to our simplistic approach to assigning segments to topics,

in which a topic is treated as a defined collection of street segments rather than a distribution over all street segments. The issue requires further investigation though, which is outside the scope of this paper.

For modelling purposes, the discontinuity is tackled by creating a topic graph in which each topic is represented by multiple nodes, each representing one of its connected components. The resulting topic graph is shown in Figure 5 (right). An alternative approach could construct the graph based on the largest connected component of each topic only. This would, however, lead to a disconnected graph as shown in Figure 5 (left).

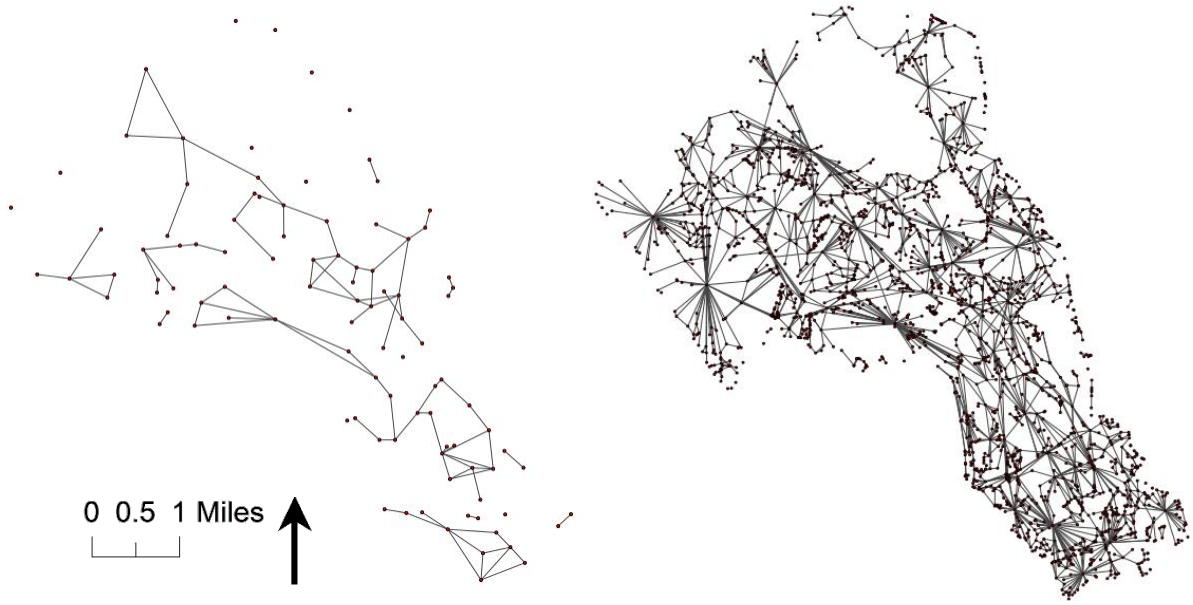

Figure 5. Topic graph created from (left) only the largest connected component of each topic, (right) all connected components of each topic, when hundred topics are inferred from police journeys in March 2010.

The topic graph is used to model police vehicle movement according to the procedures introduced in Section 2.2. The generated coverages for the *unweighted* and *weighted* model variants are shown in Figure 6. Their correlations with the actual police coverage (in Figure 7) are **0.204** and **0.349** respectively. Although these correlations are not substantially higher than a correlation of 0.315 between the data and a simplistic model assuming that vehicles always follow least distance paths (also in Figure 7), coverage patterns that they generate reflect subtle route choice preferences visible in the actual police coverage that cannot be captured under the unrealistic least distance assumption.

Figure 6. Police coverage generated by (left) unweighted and (right) weighted, topic graph from Figure 5 (right).

Figure 7. (left) Actual police coverage in March 2010 and (right) police coverage generated under the least distance assumption.

Further work is required to uncover the full potential of using topics in movement modelling. Possible extensions of the work presented in this scenario include:

- automated inference of the number of topics that would maximise modelling accuracy,
- higher order correlation metrics to measure model accuracy (e.g. correlation between counts on adjacent segments could better reflect the accuracy of a model in reconstructing movement patterns),
- addition of network connectivity information to the topic modelling algorithm in order to increase connectedness of discovered topics.

4. Biography

Kira Kowalska is a first year PhD student in the Jill Dando Institute of Crime and Security Sciences at University College London. Her main research interests lie in the area of machine learning and network analysis, particularly in application to crime and security issues.

Paul Longley is Professor of Geographic Information Science at University College London. His publications include 14 books and more than 125 refereed journal articles and book chapters. He is a former co-editor of the journal Environment and Planning B and a member of four other editorial boards. He has held ten externally-funded visiting appointments and given over 150 conference presentations and external seminars.

John Shawe-Taylor is a professor at University College London (UK) where he is the Head of the Department of Computer Science. His main research area is Statistical Learning Theory, but his contributions range from Neural Networks, to Machine Learning, to Graph Theory.

References

Agresti, A. (2008). *Statistical Methods for the Social Sciences* (4th ed.). Upper Saddle River, N.J: Pearson.

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

Farrahi, K., & Gatica-Perez, D. (2012). Extracting mobile behavioral patterns with the distant n-gram topic model. In *Proceedings of the 16th International Symposium on Wearable Computers* (pp. 1–8). IEEE.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 524–531).

Huynh, T., Fritz, M., & Schiele, B. (2008). Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*.

Manley, E. J. (2013). *Modelling Driver Behaviour to Predict Urban Road Traffic*. University College London.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, *427*(7), 424–440.