# Spatio-Temporal Clustering of Road Network Data

Tao Cheng[1] and Berk Anbaroglu[1,2]

[1] Dept. of Geomatic Engineering, University College London
Gower Street, London WC1E 6BT, UK
[2] Dept. of Geodesy and Photogrammetry Engineering, Hacettepe University
06800, Beytepe, Ankara, Turkey
{tao.cheng,b.anbaroglu}@ucl.ac.uk,
banbar@hacettepe.edu.tr

**Abstract.** This paper addresses spatio-temporal clustering of network data where the geometry and structure of the network is assumed to be static but heterogeneous due to the density of links varies cross the network. Road network, telecommunication network and internet are of these type networks. The thematic properties associated with the links of the network are dynamic, such as the flow, speed and journey time are varying in the peak and off-peak hours of a day. Analyzing the patterns of network data in space-time can help the understanding of the complexity of the networks Here a spatio-temporal clustering (STC) algorithm is developed to capture such dynamic patterns by fully exploiting the network characteristics in spatial, temporal and thematic domains. The proposed STC algorithm is tested on a part of London's traffic network to investigate how the clusters overlap on different days.

**Keywords:** spatio-temporal clustering, road network, spatio-temporal homogeneity and heterogeneity.

## 1 Introduction

As the amount of data which have spatial and temporal dimensions increases dramatically via the wide usage of sensors and crowd sourcing, spatio-temporal data mining are getting more popular. One of the tasks under spatio-temporal data mining is spatio-temporal clustering, which is the process of grouping data into clusters where the similarity between the observations (an 'observation' in this paper refers to a single measurement which has a value on thematic, spatial and temporal domains) of a cluster is high whereas the dissimilarity between observations at different clusters is high as much as possible. Clustering is used to find the patterns in the data, which will be very useful for pattern analysis.

Among others, finding the non-recurrent traffic congestion is the key activities of Transport for London in order to deliver the journey time reliability for London Olympic Games. However, the dynamics of the traffic reveals the complexity of the network performance which results in traffic congestions change with time, appearing in different size and at different area. The uneven density of road network makes the pattern heterogeneous cross the network. Furthermore, road networks in a week show

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 116–123, 2010.

different patterns due to the travel behaviour difference. Spatio-temporal data include information of three domains which can be used for clustering individually or jointly. Thematic domain defines the characteristics of the data. Spatial domain is used to describe the location of the data. And lastly temporal domain defines the timing of the observation. These domains are used to answer the questions '*what*', '*where*' and '*when*' respectively.

It is seen that initial research on clustering focused on spatial domain on point data where the density of points are taken into account [8, 10]. Then clustering has been conducted on the combined thematic and spatial domains. [2] combined the spatial and thematic distances into one distance measure by using a pre-defined value (i.e. *w*). This value determines the trade off between the two domains' distances. Choosing *w* is not trivial and it is chosen intuitively. [1] used spatial adjacency relation to cluster on road network data by also considering the traffic flow as the thematic domain. These two domains are combined with each other by considering both the inter-cluster similarity and intra-cluster dissimilarity. [9] mentioned the linkage between detecting spatial clusters and defining the spatial weight matrix. In their research the similarity criteria was based on Getis-Ord local statistic $G_i^*$. This local statistic is applied for all combinations of the contiguous neighbours and the combination which maximizes this statistic is chosen as the spatial neighbourhood.

Research has also been conducted on the combined temporal and thematic domains. Temporal clustering is the clustering of time series of observed values on the thematic domain. This approach can be used when a retailer wishes to know whether its customers are changing over time or a financial institution wants to determine if the credit card fraud transactions change over time. [3] did temporal clustering by using three indicators; number of clusters that each time stamp will have, minimum number of observations that each cluster should have and the minimum distance between two consecutive clusters; which are defined by the user. Their approach will lead to investigate how clusters at different time stamps are related with each other.

Efforts have been made to detect the clusters considering all three of the domains. For example, [4] divided the time line into fixed size intervals and calculated the similarity based on the thematic domain. A spatial distance threshold is defined to create a graph showing the similarity relations. However, choosing the spatial distance threshold is not a trivial issue. [5] used a probabilistic approach, space-time scan statistics, to detect emerging spatio-temporal clusters. This idea is well suited and applied on disease surveillance. However, the spatio-temporal process is assumed to follow a Poisson distribution which may not be the real case. [7] detected and tracked the clusters of moving point objects. The point observations intersecting in space at some time can be clustered by using the historical clustering information these observations. However, the main limitation is that the number of clusters should be determined a priori.

It is shown that efforts have been made in existing spatio-temporal clustering algorithms to detect clusters changing with time so that clusters capture the dynamicity of spatial phenomenon. Most research, however, either requires threshold values to evaluate the similarities in the spatial or temporal domains, or the number of clusters, which can be tedious for the user. Most importantly, those thresholds or numbers are actually should not be fixed since they are dynamic, relevant to the thematic attribute of the network. In terms of that, a fully dynamic approach should be adopted for spatio-temporal clustering which will be able to capture the dynamics, homogeneity and heterogeneity in the data.

Next section describes the proposed algorithm to cluster spatio-temporal network data. It is followed by a case study of transport network in London. Conclusion and future work is given in the final section.

## 2 Proposed Spatio-Temporal Clustering Algorithm

Developed algorithm will exploit the spatial, temporal and thematic domains of the network data. The algorithm is initially proposed at [11] and here it is further developed by incorporating the spatio-temporal search which will be examined under section 2.2. The algorithm is designed for network data and other data types which could be represented as a graph structure (G = (V, E) where V represents the set of spatial objects represented as vertices and E represents the adjacency between the spatial objects. A similar research was done by [6] where the edges are clustered based on their change from presence to absence or vice versa. This paper extends the research at [6] by relaxing the edge's thematic values from binary to real values. The similarity function is based on the values observed at the thematic domain and applied based on the spatial and temporal relations between the objects. These relations define the search direction which the similarity function will be applied.

### 2.1 Similarity Function

Similarity function is used as the basic component of clustering. It compares two time-series objects (e.g. $p_{1..t}$ and $q_{1..t}$ , where $p_k$ and $q_k$ defines the thematic attribute value of objects $p$ and $q$ respectively at $k^{\text{th}}$ time and $p_k$ and $q_k \in \mathbb{R}$ ) which are adjacent to each other (denoted as *adjacent(p,q)* ) at their network topology and each having $t$ observations in the temporal domain. Firstly, both of the time-series is divided into equal parts where each part will have only two consecutive observations ($p_k$, $p_{k+1}$ and $q_k$, $q_{k+1}$ where $k$ = 1, 2,.., $t$-$1$). This interval ($k$ to $k$+$1$) is called as the $k^{th}$ *basic temporal interval (BTI).* It consists of two consecutive observations in temporal domain so that it is possible to derive several different similarity metrics (slope of change, difference/mean of the two observations,..) to compare between an object and objects which are adjacent to it. Also, all of the possible similarities/dissimilarities between the two compared time series will be captured by this way (since it is not sound to have a basic temporal interval of size one). In other words when comparing the two time series each having $t$ observations, there will be $t$-$1$ similarity comparisons. When two edges are compared by the similarity function the positive similarities will denote the clusters. An intuitive rule for a positive similarity, which shall cover any phenomenon, is that the change (i.e. sign of the slope) between the two consecutive observations should be same at both of the time-series. This condition is formulated as: $\frac{p_k - p_{k+1}}{q_k - q_{k+1}} > 0$.

Apart from the change of direction, it is also believed that the thematic domain values should be close to each other. This closeness is characterized by the $\delta$ parameter at the following similarity function. The similarity function forms a linear buffer zone using the sum of the two consecutive observations and if the other time series' sum of the two consecutive observations falls in that buffer zone, then the time-series' exhibits a positive similarity. For simplicity, at further references to the similarity function, $\delta$ and *adjacent(p,q)* parts will be omitted.

$$simF(p_k, p_{k+1}, q_k, q_{k+1}, \delta \mid 0 < \delta < 1; p, q \in V \wedge adjacent(p,q)) = \begin{Bmatrix} 1, if \left| \frac{p_k + p_{k+1}}{q_k + q_{k+1}} - 1 \right| < \delta \\ 0, otherwise \end{Bmatrix}. \quad (1)$$

Yet, this is not the only similarity function that can be defined. Other binary similarity functions, as long as they take two consecutive observations of the two time series, can be defined by using background knowledge.

### 2.2 Search Directions

Once the similarity function is determined, the algorithm searches for positive similarities to form spatio-temporal clusters. Searching in spatio-temporal domain can be investigated under three categories: spatial-same time, temporal-same space and spatio-temporal.

Spatial-same time search compares two adjacent objects at same times. In other words, similarity function is evaluated for all of the adjacent objects (i.e. for all $p$ and $q$ pairs that are *adjacent(p,q)*): $simF(p_k, p_{k+1}, q_k, q_{k+1} \mid k = 1,2,..,t-1)$. Similarity search continues with the objects that are adjacent to $q$ at the positive similarities found at the comparison of $p$ and $q$. This recursive search continues until there are no more positive similarities found.

Second search direction is temporal-same space which analyses the temporal similarity in between the objects themselves at consecutive times: $simF(p_k, p_{k+1}, p_{k+1}, p_{k+2} \mid k = 1,2,..,t-2)$.This search direction is expected to give the highest number of clusters and there is no need for recursive search.

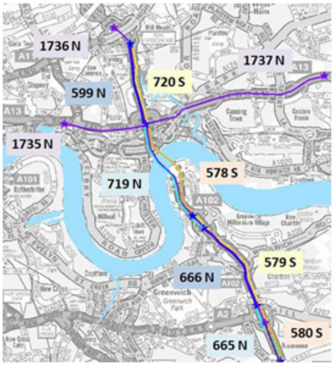
Third search direction, spatio-temporal search, is the search conducted in both space and time. The main idea in this search lies to model the flow in the network. Since in the road network vehicles flow, an event happening at one edge might show its effect at its adjacency after a time elapse. The main question under this search direction is to find the number of time steps which quantifies this time elapse. This number is referred as *stepSize*. The similarity function that will be used will be like: $simF(p_k, p_{k+1}, q_{k+stepSize}, q_{k+stepSize+1} \mid k = 1,2,..,t - stepSize - 1)$.

## 3 Case Study –Blackwall Tunnel at London

### 3.1 Data Description

Data is collected via the automatic number plate recognition (ANPR) cameras at 5 minute intervals between 28 December 2009 – 3 January 2010 for 11 links (a link is defined as the road segment(s) where two ANPR cameras –one at the beginning and one at the end of the link- operate and collect data according the number of cars passed) at a region which is renowned with its congestion; Blackwall Tunnel. This region has a road that passes underneath the river Thames and the region is very close to the banking centre of London; Canary Wharf. Spatial dimension is used via the adjacency matrix, temporal dimension can be thought in different scales, as minute, day or week based and the thematic domain is the average speed in km/h. Aim of the case study is to observe how spatio-temporal clusters occurring on different days of a week differ from each other and analyze how adjacent links behave.

There are 11 links where 3 of them (i.e. 1735N, 1736N, 1737N) overlap with the rest of the links. Because of this overlap, they are not adjacent to any of the links and discarded from the analysis. Letters shown near the link numbers denote the direction of flow, where N and S mean that the traffic on that link flows towards north and south respectively. The study area and the adjacency matrix is illustrated at figure 1.



| links | 578 | 579 | 580 | 599 | 665 | 666 | 719 | 720 |
|---|---|---|---|---|---|---|---|---|
| 578 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 579 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 580 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 599 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 665 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 666 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 719 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 720 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Fig. 1.** Blackwall Tunnel Region and the Adjacency Matrix

It can be seen that the traffic flow direction is incorporated in defining the adjacency matrix. If the directionality of the links has not been incorporated then for example 579 and 666 will be adjacent, however they are not adjacent since 579th link is flowing towards south whereas 666th links flows towards north.

Two parameters should be determined before searching for detecting the clusters. First is the δ parameter which will be used at the similarity function stated at equation 1. It is chosen as 0.1 after consulting to the colleagues at Transport for London. Second parameter is the stepSize, which is chosen after investigating the average journey time for the links on the first day of the analysis. The average journey time for the links is shown at figure 2. Solid black line shows the 5 minute level, and it is seen that most of the journeys are completed less than 5 minutes. This suggests that stepSize can not be not greater than 5 minutes, thus stepSize is chosen as one BTI. In other words, when the spatio-temporal search is conducted the similarity function will be in the form of $simF(p_k, p_{k+1}, q_{k+1}, q_{k+2} \mid k = 1,2,..,t-2)$.

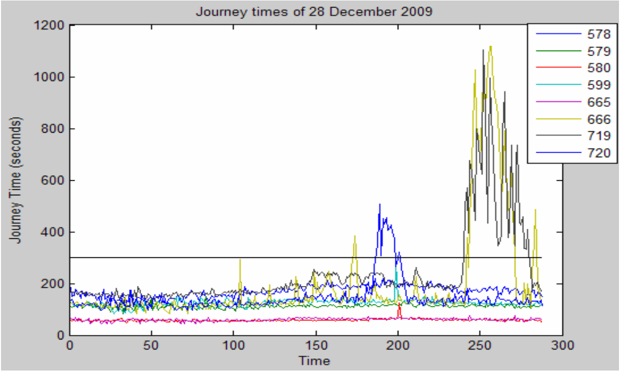


**Fig. 2.** Journey times of 28 December 2009 at Blackwall Tunnel Links

### 3.2 Results

Two of the search directions, spatial-same time and spatio-temporal, are tested since temporal-same space search is pure temporal analysis and does not reveal information about the network as a whole.

Each detected cluster has two components. First component states the links involved in that cluster and second denotes the BTIs which those links are clustered. This representation does not reveal much information, thus the representation is transformed into a binary time series where; if there is a cluster at the BTI, the function outputs one otherwise zero. These binary sequences of clusters were suitable for comparing the clusters that are detected on different days. To compare these binary sequences, Jaccard Coefficient (JC) is used, which is defined as:

$$JC = \frac{n_{11}}{n_{10}+n_{01}+n_{11}}. \tag{2}$$

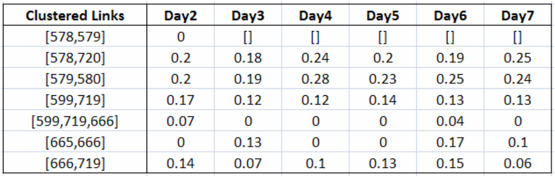
where $n_{11}$ denotes the number of times when both of the days has a cluster at that BTI (i.e. positive match), $n_{00}$ denotes the number of times when both of the days does not have a cluster at a BTI (i.e. negative match) $n_{10}$ and $n_{01}$ denotes the number of mismatches where one day has a cluster at a BTI whereas there is no cluster at that BTI in the other day.

For experimental purposes, clusters that occur on the first day of the analysis (28 December 2009) is compared with the clusters that occurred at the rest of the days (29 December – 3 January). It can be verified from the JCs that the days show distinct characteristics.

Spatial-same time and spatio-temporal search directions will be done in order to capture the spatio-temporal clusters.

Table 1 shows the JCs between the clusters shown at the left column and at comparison between the first day and the other days (day 2 –day 7) on the spatial-same time based search. '[ ]' means that either of the days does not have that cluster. It is seen that links 578 and 579 has a cluster only at the first and second days and at the other days; there is no 578-579 cluster which strongly suggests that these two links have different characteristics.

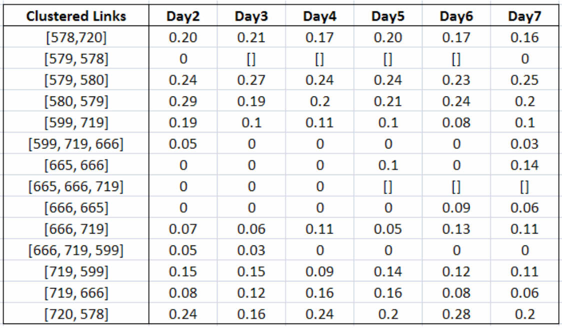
**Table 1.** Jaccard Coefficients for the spatial-same time search



| Clustered Links | Day2 | Day3 | Day4 | Day5 | Day6 | Day7 |
|---|---|---|---|---|---|---|
| [578,579] | 0 | [] | [] | [] | [] | [] |
| [578,720] | 0.2 | 0.18 | 0.24 | 0.2 | 0.19 | 0.25 |
| [579,580] | 0.2 | 0.19 | 0.28 | 0.23 | 0.25 | 0.24 |
| [599,719] | 0.17 | 0.12 | 0.12 | 0.14 | 0.13 | 0.13 |
| [599,719,666] | 0.07 | 0 | 0 | 0 | 0.04 | 0 |
| [665,666] | 0 | 0.13 | 0 | 0 | 0.17 | 0.1 |
| [666,719] | 0.14 | 0.07 | 0.1 | 0.13 | 0.15 | 0.06 |

Similar idea is applied when the spatio-temporal search is conducted and the Jaccard Coefficients are shown at table 2. The previous results are validated as there is no relation between the links 578 and 579 in spatio-temporal search as well. In addition, the links 579 and 580 showed the highest overlap between the days.

**Table 2.** Jaccard Coefficients for the spatio-temporal time search



| Clustered Links | Day2 | Day3 | Day4 | Day5 | Day6 | Day7 |
|---|---|---|---|---|---|---|
| [578,720] | 0.20 | 0.21 | 0.17 | 0.20 | 0.17 | 0.16 |
| [579, 578] | 0 | [] | [] | [] | [] | 0 |
| [579, 580] | 0.24 | 0.27 | 0.24 | 0.24 | 0.23 | 0.25 |
| [580, 579] | 0.29 | 0.19 | 0.2 | 0.21 | 0.24 | 0.2 |
| [599, 719] | 0.19 | 0.1 | 0.11 | 0.1 | 0.08 | 0.1 |
| [599, 719, 666] | 0.05 | 0 | 0 | 0 | 0 | 0.03 |
| [665, 666] | 0 | 0 | 0 | 0.1 | 0 | 0.14 |
| [665, 666, 719] | 0 | 0 | 0 | [] | [] | [] |
| [666, 665] | 0 | 0 | 0 | 0 | 0.09 | 0.06 |
| [666, 719] | 0.07 | 0.06 | 0.11 | 0.05 | 0.13 | 0.11 |
| [666, 719, 599] | 0.05 | 0.03 | 0 | 0 | 0 | 0 |
| [719, 599] | 0.15 | 0.15 | 0.09 | 0.14 | 0.12 | 0.11 |
| [719, 666] | 0.08 | 0.12 | 0.16 | 0.16 | 0.08 | 0.06 |
| [720, 578] | 0.24 | 0.16 | 0.24 | 0.2 | 0.28 | 0.2 |

From these results it is seen that daily patterns vary, and should be treated accordingly. In addition, number of positive similarities decrease as the spatial search is extended; which is indeed an expected outcome.

## 4 Conclusion

This research proposed an algorithm of spatio-temporal network clustering where spatial and temporal domains are exploited using the network topology (via adjacency matrix) and characteristics (via spatial-same time and spatio-temporal search directions). It is shown that, the proposed algorithm captures the dynamics of the network so that the spatio-temporal clusters appear and disappear with time. In addition, by this way both homogeneity and heterogeneity in the data is captured. Finally, user is involved only in determining the similarity on thematic domain which can be decided by domain expert.

Future work will be focused on different scales in temporal domain. In most the researches, as well as the case study mentioned in this paper, temporal dimension exhibits in different scales (time-of-day, day-of-week, etc.) and scaling the results at these different scales is a challenging problem. Secondly, detecting the appropriate stepSize is also a challenging problem, since this research assumed that stepSize does not change spatio-temporally. Finally, the linkage between the spatio-temporal clusters and the traffic congestion will be sought.

**Acknowledgments.** This research is jointly supported by UK EPSRC (EP/G023212/1), Chinese NSF (40830530) and 863 Programme (2009AA12Z206). Thanks to Transport for London for providing the data and domain knowledge. Second author thanks to Hacettepe University and Higher Education Council of Turkey for the PhD scholarship they provide.

## References

1. Wang, Y., Chen, Y., Qin, M., Zhu, Y.: SPANBRE: An Efficient Hierarchical Clustering Algorithm for Spatial Data with Neighborhood Relations. In: Fourth International Conference on Fuzzy Systems and Knowledge Discovery. FSKD 2007, pp. 665–669 (2007)
2. Lin, C., Liu, K., Chen, M.: Dual clustering: integrating data clustering over optimization and constraint domains. IEEE Transactions on Knowledge and Data Engineering 17, 628–637 (2005)
3. Adomavicius, G., Bockstedt, J.: C-TREND: Temporal Cluster Graphs for Identifying and Visualizing Trends in Multiattribute Transactional Data. IEEE Transactions on Knowledge and Data Engineering 20, 721–735 (2008)
4. Wei, L., Peng, W.: Clustering Data Streams in Optimization and Geography Domains. Advances in Knowledge Discovery and Data Mining, 997–1005 (2009)
5. Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K.: Detection of emerging space-time clusters. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 218–227. ACM, Chicago (2005)
6. Chan, J., Bailey, J., Leckie, C.: Discovering correlated spatio-temporal changes in evolving graphs. Knowledge and Information Systems 16, 53–96 (2008)
7. Rosswog, J., Ghose, K.: Detecting and Tracking Spatio-temporal Clusters with Adaptive History Filtering. In: IEEE International Conference on Data Mining Workshops. ICDMW 2008, pp. 448–457 (2008)
8. Yiu, M.L., Mamoulis, N.: Clustering objects on a spatial network. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 443–454. ACM, Paris (2004)
9. Aldstadt, J., Getis, A.: Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. Geographical Analysis 38, 327–343 (2006)
10. Harel, D., Koren, Y.: Clustering spatial data using random walks. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 281–286. ACM, San Francisco (2001)
11. Cheng, T., Anbaroglu, B.: Defining Spatio-Temporal Neighbourhood of Network Data. In: ISGIS, pp. 75–80 (2010)