DEFINING DYNAMIC SPATIO-TEMPORAL NEIGHBOURHOOD OF NETWORK DATA

Tao Cheng [1] 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, Ankara, Turkey
{tao.cheng, b.anbaroglu}@ ucl.ac.uk

Commission II, WG II/3

KEY WORDS: Spatio-temporal neighbourhood, spatio-temporal clustering, network complexity

ABSTRACT:

To improve the accuracy and efficiency of space-time analysis, spatio-temporal neighbourhoods (STNs) should be investigated and analysed in the classification, prediction and outlier detection of space-time data. So far most researches in space-time analysis use either spatial or temporal neighbourhoods, without considering both time and space at the same time. Moreover, the neighbourhoods are mostly defined intuitively without quantitative measurement. Furthermore, STNs of network data are less investigated compared with other types of data due to the complexity of network structure. This paper investigates the existing approaches of defining STNs and proposes a quantitative method to define STNs of network data in which the topology of the network does not change but the characteristics of the edges (i.e. thematic attribute values) change with time which requires dynamic STNs adapted to the properties of the network. The proposed method is tested by using London traffic network data.

1. INTRODUCTION

The amount of data which has spatial and temporal dimensions increases dramatically via the wide usage of fixed or mobile sensors. Extracting useful information from these data gains importance day by day which is referred as spatio-temporal data mining. Classification, prediction, outlier detection and clustering are among the major tasks of spatio-temporal data mining and analysis. These tasks involve analyzing spatio-temporal neighbourhoods (STNs) because STNs of an instance give important clues on the evolution of the instance itself. Better defining and identifying the neighbouring instances will lead to better modelling of the phenomenon under investigation.

So far most STNs root from spatial neighbourhoods (SNs) which then elongated in time. Different data forms as point, grid, polygon or network will exhibit different forms of STNs. Thus, different methods are used to define STNs on these data forms. For point data, a distance threshold is usually used to define spatial neighbourhood and temporal neighbourhood is usually defined intuitively (Celik et al., 2006). Lu et al. (2003) and Chen et al. (2008) used k-Nearest Neighbour (k-NN) based on Mahalanobis distance to find the SNs of an instance. Then STN is treated as a static entity where spatial neighbourhood is simply elongated in time dimension (GeoPKDD, 2006; Zhang et al., 2008). Grid data possesses a regular structure where several intuitive definitions for spatial neighbourhood exist. These neighbourhood strategies can be considered as metaphor of chess pieces and named as rook, queen and bishop neighbourhood. Yin and Collins (2007) used rook neighbourhood as the spatial neighbourhood and considered one time step backward as temporal neighbourhood to detect moving objects on videos. For polygon data, if two polygons share a common edge then they are thought to be spatial neighbours. Billard et al. (2007) predicted an epidemic in time across 12 states of US based on spatial neighbourhoods considering one former time stamp. Network data is the least investigated among all the data types and networks are treated as graph structures where the spatial neighbourhoods are defined by graph connectivity. Shekhar et al. (2001) defined STNs as spatial neighbourhoods that are adjacent at the graph with temporal neighbourhood consisting of previous time stamps of the spatial neighbourhoods. Although space and time are integrated in STNs, the neighbourhoods are fixed for an individual like in space and time. However, due to the complexity of networks, the neighbourhoods are actually dynamic.

This paper is motivated from aforementioned facts: 1) space and time are not treated in an integrated manner; 2) there is not a quantitative method for defining dynamic STNs for network data. To achieve this quest, literature related with spatial and spatio-temporal clustering is given under the second section. Third section discusses the proposed algorithm to cluster on spatio-temporal network data in which the topology of the network does not change but the thematic attribute associated with the edges of the network changes with time. The case study is discussed and results are shown in the fourth section. Conclusion and future work is given in the fifth section.

2. RELATED WORK ON SPATIAL AND SPATIO-TEMPORAL CLUSTERING

We believe that the originating point to attack the problem to define STNs should be spatio-temporal clustering, because the underlying ideas of both “neighbourhood” and “clustering” is same: to group the observations so that similar observations will fall into the same grouping (i.e. neighbourhood or cluster respectively). Considering these, the research problem can be restated as: dynamic spatio-temporal clustering on spatially embedded network data. There are three domains under spatio-temporal clustering: thematic, spatial and temporal attributes. Thematic attribute gives the information about the phenomenon observed, spatial and temporal attributes give the location and timing of the observation respectively. This section describes the literature conducted on spatial and spatio-temporal clustering.

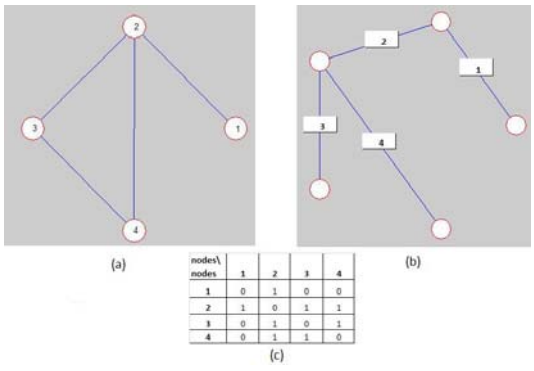
Spatial clustering problem is also referred as ‘Dual Clustering’ by Lin (2005). They proposed two distance functions: one in spatial and another in thematic domain. These functions are

combined with a pre-defined weight value to get one distance function. Choosing the pre-defined weight value is not trivial and it is chosen intuitively in their research. Wang (2007) used only spatial neighbourhood relations to cluster network data without considering the temporal domain. As a result, the dynamics in the network are not captured.

Spatial clustering is not sufficient to understand 'events' since to describe an event, one needs to answer the questions of *what*, *when* and *where*. In other words, thematic, temporal and spatial domains should be combined in a consistent way to have a better understanding of spatial phenomenon. Wei (2009) divided the time line into fixed size intervals and calculated the similarity based on the thematic domain. Spatial domain is used by means of defining a spatial distance threshold. However, how to choose the spatial distance threshold was not discussed. In addition, clustering results depend on the size of the chosen temporal interval. Neill (2005) emphasized on the significance of temporal domain. They used a probabilistic approach to detect emerging spatio-temporal clusters. However, the spatio-temporal process is assumed to follow a Poisson distribution which may not be the real case or time-consuming tests should be done to verify this assumption. Chan et al. (2008) captured the temporal dynamics of a graph by inspecting on the presence or absence of an edge. Their main task is to detect the regions where the change (absence/presence of an edge) is spatio-temporally correlated.

3. SPATIO-TEMPORAL CLUSTERING ON SPATIALLY EMBEDDED NETWORKS

Theoretically, one can represent the spatio-temporal objects as either vertices or edges in an undirected graph. An example of this is shown at figure 1. Figure 1 (a) represents the objects at vertices and Figure 1 (b) represents the same objects at edges. Figure 1(c) is the adjacency matrix for both of the graphs shown at Figure 1(a-b). To be consistent with the case study, from now on the representation shown at Figure 1 (b) will be used. Thus, spatio-temporal objects are the edges of the graph and vertices connect the objects coincident to them. In either case, the idea behind the representation is to obtain the adjacency matrix.



nodes\ nodes	1	2	3	4
1	0	1	0	0
2	1	0	1	1
3	0	1	0	1
4	0	1	1	0

Figure 1: Different graph representations of same data

Although the algorithm is designed for network data, this algorithm could be used for spatio-temporal clustering whenever the spatio-temporal phenomenon (which can exhibit in point, line or polygon) could be represented as a graph structure (G = (V, E) where V represents the set of vertices and E represents the set of edges).

Once the graph structure of the spatio-temporal phenomenon is acquired, then a matrix showing the connectivity between vertices (or edges); adjacency matrix; is created for the graph structure. While creating the adjacency matrix (if exists), the direction of the edges could be incorporated.

Up to now, the spatial domain is used to acquire the adjacency matrix of the spatio-temporal phenomenon. Temporal and thematic domains are exploited at this stage. Temporal domain is divided into equal parts where each part will have only two consecutive observations in the thematic domain. This is called as the *basic temporal interval*. For example basic temporal interval k of the object p consists of the two thematic attribute observations of p^{th} object at consecutive times of k-1 and k. At each comparison step, basic temporal interval is shifted one time step. Thus, if the time-series has a length of t, there will be $t - 1$ similarity results for the two adjacent objects' similarity comparison. Since it consists of two consecutive (in temporal domain) observations, it is possible to derive several different similarity metrics (slope of change, difference/mean of the two observations,..) to compare between an object and the 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). This is the first novelty of this research, since there is no need to specify a window size at temporal domain and it is designed to be the simplest possible, having two consecutive observations. In addition, this will allow capturing all of the possible similarities between two time series.

The similarity function is defined at basic temporal interval of k for two adjacent objects p and q with *at least* four inputs (i.e. p_{k-1}, p_k, q_{k-1}, q_k) where the thematic attribute value of p and q at times k-1 and k are denoted as p_{k-1} and p_k and q_{k-1} and q_k respectively. Similarity function takes at least these four inputs, because some other parameters (which should be defined using background knowledge) may be needed to define the flexibility of similarity comparison.

For the objects to be labelled as positively similar at *basic temporal interval k* two requirements should be fulfilled: Firstly, the direction of change in thematic attribute values (i.e. slope) should be same and secondly, the thematic values of both objects should be similar which is quantified by the parameter δ. This requirement needs to be symmetric (e.g. if spatial object p is found to be positively similar at basic temporal interval k with the spatial object q, then q should also be positively similar with p at k^{th} basic temporal interval) , thus has two parts separated by a logical *or* operator. These two requirements for a positive similarity are illustrated at equations 1 and 2 respectively. If either of these conditions hasn't met, then the similarity function will return a negative similarity result.

$$\frac{p_k - p_{k-1}}{q_k - q_{k-1}} > 0 \tag{1}$$

$$\begin{aligned} &(1-\delta)(q_k + q_{k-1}) < p_k + p_{k-1} < (1+\delta)(q_k + q_{k-1}) \\ &\qquad\qquad \vee \\ &(1-\delta)(p_k + p_{k-1}) < q_k + q_{k-1} < (1+\delta)(p_k + p_{k-1}) \end{aligned} \tag{2}$$

These similarity criteria are one of the many possibilities, however we tried to make it as generic as possible.

This time-series similarity comparison is done for all the adjacent objects which will constitute the dynamic spatial neighbourhood (i.e. spatial neighbourhood changes with time) of an object rather than STN. This is because; the comparison between adjacent objects is the done at the same basic temporal intervals. To capture the STN, temporal comparison should also be done. This will be achieved by comparing the object p at basic temporal interval k with itself at basic temporal interval $k+1$. The requirements for this comparison are same as spatial comparison stated above with the only difference that q_{k-1} and q_k is replaced with p_k and p_{k+1} respectively.

After the similarity comparisons are done for all adjacent objects, spatio-temporal clustering search is extended through the adjacencies of the adjacent objects. This time, adjacent objects are compared with their adjacencies at the intervals where a positive similarity is determined in the previous stage, so that spatio-temporal clusters will be spatially linked within themselves. Similar with the spatial search, temporal search will also extend to temporal adjacencies as long as there is positive similarity. This combination of spatial and temporal searches based on the similarity metric will constitute the spatio-temporal clusters and indeed the STNs. This search continues, until there is no more similarity is found. Therefore, the second novelty of this proposed algorithm is that there is no need to define spatial or distance threshold. By this way, the spatio-temporal clustering is conducted on all three domains: thematic, spatial and temporal domains.

4. CASE STUDY

4.1 Data Description

Algorithm presented in the third section is applied on the spatio-temporal traffic data of London road network. The data consists of average journey time of 11 objects obtained at 5 minutes intervals on 28 December 2009. These 11 links (i.e. 'objects' of this case study) are near the Blackwall Tunnel (figure 2) which is known as its unexpected congestions due to traffic accidents. Each link has an id (number) and a direction indicator (N or S tells traffic flow is towards north or south). Thematic attribute is converted from average journey time into average excess time per kilometre.

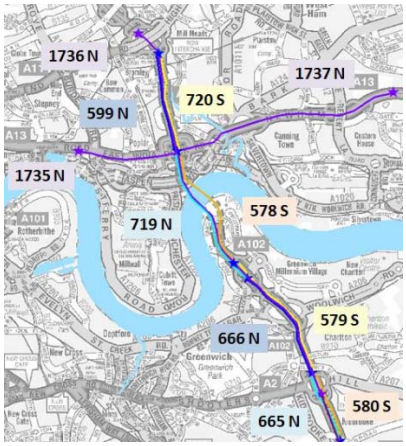


Figure 2: Map of the links in case study

To calculate the excess time per kilometre one need to define the average free flow journey time which we defined as the average of journey times occurred between 02:00 – 06:00 which is a common time interval used to observe the free flow characteristics of the link. Equation 3 shows the calculation of excess time per kilometre, where $jt_{d_{i,t}}$ denotes the average journey time of the link i at time t on day d. Similarly $jt_{d_{i,avg(02:00-06:00)}}$ denotes the free flow journey time at the link i and day d. Excess journey time is calculated as the difference between the observed journey time and free flow journey time. Then, this difference is divided by the length of the link to get the thematic attribute to be used to do spatio-temporal clustering. This metric is used to for spatio-temporal clustering, since it is a good metric to measure congestion.

$$excess_{per_{km(d,i,t)}} = \frac{jt_{d_{i,t}} - jt_{d_{i,avg(02:00-06:00)}}}{length(i)} \quad (3)$$

Another thing to define is the rules to define the adjacency of the links. An intuitive rule; flow of traffic should be in same direction and links should coincide at a vertex, is used to create the adjacency matrix. For instance, links 665N and 580S coincides at the same vertices; however they are not considered as adjacent since the flow of traffic is in opposite directions. With this definition of adjacency, adjacency matrix for the objects of figure 2 is created as shown at figure 3 where the links (1735-1737)N are not included because they overlap with other objects.

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

Figure 3: Adjacency Matrix of the objects of Figure 2

Rules to define the adjacency matrix can even be more specified by considering a traffic rule which is; an event happening at downstream will effect the upstream but not the vice versa. With this definition, 599N will be adjacent with 719N, but 719N will not be adjacent with 599N because what ever happened at 719N will not affect 599N because the traffic is flowing from 719N to 599N. However, this idea is not incorporated when defining the adjacency matrix in this research.

Final thing to define is the δ parameter in the similarity function. It is decided intuitively as 0.1 leading to the interpretation that two adjacent objects will be similar as long as their sum of the thematic attribute (i.e. excess time per km) values at a basic temporal interval lies in 10 percent zone of other adjacent object's sum of the thematic attribute (i.e. excess time per km) values at that basic temporal interval.

4.2 Results

Implementations were conducted on Matlab 2008a, however due to time constraints only spatial search part was implemented without doing the temporal search part. Once the similarities of adjacent objects are determined and the similarity search is extended based on the adjacency matrix and positive

similarities. Thus, the results stated at figure 4 shows the dynamic spatial neighbourhood discovery rather than the STN.

On the left column of figure 4 the clustered links and on the right column the basic temporal intervals in which they are clustered are shown. Basic temporal intervals of a day start with 1 which indicates the time between 00:00-00:05 and ends with 288 which indicates the time between 23:55-00:00.

Links	Basic Temporal Intervals
[578,720]	[152,167,204,237,262]
[579,580]	[153,155,198,224,226,253,257,283]
[580,579]	[153,155,198,224,226,253,257,283]
[599,719]	[113,117,126,136,137,139,144,206,220,221]
[599,719,666]	139
[665,666]	[38,128,141,194,203,229]
[666,665]	[38,128,141,194,203,229]
[666,719]	[79,94,123,139,148,150,166,200,264]
[666,719,599]	139
[719,599]	[113,117,126,136,137,139,144,206,220,221]
[719,666]	[79,94,123,139,148,150,166,200,264]
[720,578]	[152,167,204,237,262]

Figure 4: Clustering Results of 28 December 2009

Since the main program searches for all of the individual objects (this can be seen from the first object under the different clusters) and creates the clusters based on that search, the adjacent links are clustered at the same basic temporal intervals which implies the symmetric nature of the clusters. For example when searching for the object 579, similarities were found at the basic temporal intervals of 113, 117, 126, 136, 139, 144, 206, 220 and 221 with object 719. When the main program runs for the object 719, it clusters with object 599 at the same basic temporal intervals. This search is redundant and similarity of the matrix can be exploited to eliminate this redundant search. However, this depends on the adjacency matrix and if the adjacency matrix is not symmetric then, all the objects should be searched.

Results indicate that, as the similarity search extends towards the adjacencies of adjacent links, detected similarities decreases. This is an expected result, since spatio-temporal correlation will decrease with the increase of the spatial distance between the objects. There is only one basic temporal interval (i.e. 139) and where a spatio-temporal cluster is formed among the links 599, 719 and 666.

These results show that the dynamic nature of the spatio-temporal clusters can be captured by using the proposed algorithm: spatio-temporal clusters grow and shrink with time (since the property of the network change with time).

It is clear that, these results utterly depend on the chosen similarity function and its parameters as well as the rules to define the adjacency among the objects. As aforementioned, the similarity function needs to be defined by considering background knowledge about the phenomenon.

5. CONCLUSION

This paper addressed the issue of importance of defining STNs. It is seen that most of the spatio-temporal data mining tasks use the STN concept. Commencing from the analogy between neighbourhood and cluster, this research proposed an algorithm on spatio-temporal network clustering for the task to determine the STN. It is shown that, the proposed algorithm captures the dynamics of the network. Both spatial and temporal information are used effectively to reduce the computational cost of detecting spatio-temporal clusters. Other than choosing the similarity function and related parameters of the similarity function, the user is not involved in choosing any spatial or temporal parameter.

There are several drawbacks of the proposed algorithm as well. Firstly, it cannot handle the cases where there is a loop at an edge (i.e. edge from a vertex to itself). Secondly, similarity is treated as a binary characteristic. And finally, the algorithm can only handle spatially embedded (network topology does not change with time) network objects.

The case study did not capture the STN due to time limitations but only captured the dynamically changing spatial neighbourhood. Implementing and testing the temporal search part is left as a future work. 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.). Future research will also focus on the extending the spatio-temporal clustering to different time scales. In addition understanding the dynamic nature of the spatio-temporal clusters (e.g. finding the patterns of growth and shrink) is another challenging task to be sought.

ACKNOWLEDGEMENTS

This research is jointly supported by UK EPSRC (EP/G023212/1), Chinese NSF (40830530) and 863 Programme (2009AA12Z206). Second author thanks to Hacettepe University and Higher Education Council of Turkey for the PhD scholarship they provide. Data is kindly provided from TfL. We also would like to thank to the anonymous reviewers for their comments.

REFERENCES

Billard, L., Lee, S.D., Kim, D.K., Lee, K.M., Lee, C.H., Kim, S.S., 2007. Modeling Spatial-Temporal Epidemics Using STBL Model. In: *International Conference on Machine Learning and Applications,* 629-633.

Celik, M., Shekhar, S., Rogers, J.P., Shine J.A., 2006. Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining: A Summary of Results. In *18th IEEE International Conference on Tools with Artificial Intelligence,* 106-115.

Chan, J., Bailey, J., Leckie, C., 2008. Discovering correlated spatio-temporal changes in evolving graphs. *Knowledge and Information Systems*, 16, pp. 53-96.

Chen, D., Lu, C.T., Kou, Y., Chen Y., 2008. On Detecting Spatial Outliers. *GeoInformatica*, 12(4): 455-475.

GeoPKDD. 2006. Last visited: 11 December 2009 http://www.geopkdd.eu/files/dev_public/D2.2M.pdf

Lin, C.R., Liu, K.H., Chen, M.S., 2005. Dual clustering: integrating data clustering over optimization and constraint domains. *IEEE Transactions on Knowledge and Data Engineering,* , 17, pp. 628-637.

Lu, C. T., Chen, D., Kou, Y., 2003. Detecting spatial outliers with multiple attributes. *15th IEEE International Conference on Tools with Artificial Intelligence,* 122-128.

Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K., 2005. Detection of emerging space-time clusters. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, Chicago, Illinois, USA, pp. 218-227.

Shekhar, S., Lu, C., Zhang, P., 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, California, USA

Wang, Y., Chen, Y., Qin, M., Zhu, Y., 2007. SPANBRE: An Efficient Hierarchical Clustering Algorithm for Spatial Data with Neighborhood Relations. In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery,* pp. 665-669.

Wei, L., Peng, W., 2009. Clustering Data Streams in Optimization and Geography Domains. In: *Advances in Knowledge Discovery and Data Mining*, pp. 997-1005.

Yin, Z., Collins. R., 2007. Belief Propagation in a 3D Spatio-temporal MRF for Moving Object Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition.*

Zhang., S., Yao, H., Liu, S., 2008. Dynamic background modeling and subtraction using spatio-temporal local binary patterns. In: *ICIP 2008,* 1556-1559.