Spatio-Temporal Data Mining
– Coping with the Increasing Availability of Motion Data in Geography

***Patrick Laube*[1]**

[1]School of Geography and Environmental Science
University of Auckland. Auckland, New Zealand
Phone: +64 9 373-7599 # 88202 Fax: +64 9 373-7434
Email: p.laube@auckland.ac.nz

Presented at SIRC 2005 – The 17th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 24th-25th 2005

ABSTRACT

This paper discusses two data mining tasks to cope with the challenge of an increasing availability of high-resolution, individual-based motion data in Geography. First, techniques for motion patterns detection are presented adopting a 'retrieval by content' approach. Second, lifeline similarity measures are proposed as a means for 'descriptive modelling'.

***Keywords and phrases*:** spatio-temporal data mining, geospatial lifelines, trajectories, motion patterns, similarity measures

1.0 INTRODUCTION

GIScience is challenged by a previously unseen increase in the availability of high-resolution individual-based motion data, so-called *geospatial lifelines* (Mark 1998). The analysis of lifelines of moving point objects (MPOs) has seen a growing interest in various application fields, such as human movement analysis, traffic planning, animal behaviour science, and sports scene analysis. At the same time, Geographic Information Systems still suffer from the legacy of static cartography; hence the analytical power of GIS to cope with large space-time datasets is still limited (Peuquet, 2001).

Geography has only recently recognised the potential of data mining (Miller and Han, 2001). Data mining is the application of specific algorithms for extracting patterns from data. The classical data mining tasks are *exploratory data analysis (EDA)*, *descriptive modelling*, *predictive modelling*, *discovering patterns and rules*, and *retrieval by content* (Hand et al. 2001). In the last decade GIScience has seen several data mining approaches for motion analysis, mainly based upon EDA techniques. Exploratory (spatial) data analysis (Anselin, 1998) has been identifies as a good way to explore not only spatial but also spatio-temporal data. Several authors have, for example, adopted Hägerstrand's classical time geography framework for exploring motion processes (Forer, 1998; Forer 2002, Kraak and Koussoulakou, 2004). However, it has also been acknowledged that visual inspection reaches its limits if numbers of MPOs and lengths of lifelines increase (Kwan, 2000).

The other data mining tasks recieve much less attention. This paper discusses the potential of *retrieval by content* and *descriptive modelling* for motion analysis. For the first task a user has a motion pattern of interest in mind and wishes to find similar patterns in the data. The second task aims at finding a descriptive model that describes all of the lifeline data, e.g. by deriving lifeline clusters.

2.0 RELATED WORK

This section presents a selection of references that shall help to identify geography's potential in applying data mining techniques for the analysis of motion data. A geographic example and a side-step into two neighbouring disciplines provide some useful suggestions for mining motion data.

Sinha and Mark (2005, p. 115) propose a lifeline distance (dissimilarity) to detect clusters of people with similar environmental exposure history, 'providing a basis for revealing possible regions in space-time where environmental hazards might have existed in the past'. Thus, the authors base their similarity measure on spatial and temporal proximity of lifelines in a Hägerstrandian space-time aquarium (Hägerstrand, 1970).

A time series can be considered as the 1D case of trajectories. Hence, the computer science related field of time series analysis is a rich source for analytical concepts for motion analysis. Typical time series are trading prices of stock markets or climatic variables such as temperature variation. The analysis of time series involves searching for patterns such as 'going down' or 'going up'. Such queries follow the retrieval by content data mining approaches, trying to match a defined sub-trajectory to a given trajectory (Qu, Wang, and Wang, 1998). An alternative approach involves clustering similar (sub-)trajectories, thus following predictive modelling (Rafiei and Mendelzon, 2000).

Finally, in the field of video content retrieval, 2D trajectories emerge from tracking MPOs from surveillance video footage. Again, similarity queries aim at efficiently finding similar trajectories from large video footage data bases (Ng, 2001; Shim and Chang, 2003).

As these examples show, many motion analysis approaches mine exclusively the geometric properties of the lifelines, referring to some variant of a Euclidean distance between consecutive fixes building the backbone of trajectories. However, this neglects the trajectory derivatives such as speed or turning angle, which may be of greater importance for the moving object, for example the speed of an airplane or the swivelling motion azimuth of a grazing herbivore. Here lie significant opportunities for geography to contribute its expertise about topological relations between spatial entities. The remainder of this paper discusses two data mining approaches explicitly focusing on derived motion descriptors, one adopting the *retrieval by content* task (Section 3.0), and the other adopting the *descriptive modelling* (Section 4.0) task.

3.0 RETRIEVAL BY CONTENT – MOTION PATTERNS

This section presents research undertaken aimed at extending GIScience's conceptual and methodological toolbox for the analysis of tracks of MPOs. The adopted data mining task is *retrieval by content*.

The research addresses the following questions: Can we identify and formalise a set of generic *motion patterns* that can be found in the trajectories of moving point objects? How can we automatically detect such predefined motion patterns? And, finally, how can we evaluate the relevance and meaningfulness of motion patterns?

This approach is based on a family of so-called *relative motion (REMO)* patterns, with the term relative motion denoting the interrelation of motion attributes of different moving point objects over space and over time (Laube and Imfeld, 2002). Hence, motion patterns are referred to as predefined formalised search templates of motion attributes such as speed, change of speed, motion azimuth, or sinuosity. Such a pattern could refer to a specific time when a defined set of *n* MPOs move concurrently in the same direction (*concurrence* pattern). Such a concurrence pattern could, for instance, when a caribou herd starts its seasonable migration towards the calving areas. Laube et al. (2004) extend the motion pattern family proposing intrinsically spatio-temporal, hence dynamic, patterns that do not exist in either space or time alone, such as *flocking*, *converging*, and *diverging*.

The proposed data mining process includes the conceptualisation, formal description and detection of motion patterns in the tracks of moving point objects. This required the development of a pattern description formalism as well as novel data structures and adapted pattern detection algorithms (Gudmundsson et al. 2004; Laube et al. 2005). Finally, methods were suggested to evaluate the relevance of the detected patterns: synthetic motion data, parameterised with real observation data, are used for Monte Carlo experiments in order to assess the interestingness of found motion patterns (Laube and Purves, in press).

4.0 DESCRIPTIVE MODELLING – LIFELINE SIMILARITY

A second data mining task suited for spatio-temporal analysis is *descriptive modelling* (Hand et al. 2001). With respect to analysing lifelines, a descriptive model shall be found which would describe all of the motion data, e.g. by deriving clusters of similar lifelines. Whereas section 3.0 presented published research, this section outlines preliminary work.

Let's consider the following example. For a study in animal behaviour science, biologist tracked the motion of a set of migrating caribou (Figure 1). One possible task would then be to quantitatively assess the similarity of the trajectories in order to identify animals that moved in a similar way. Hence, given a set of *n* equal length geospatial lifelines of GPS tracked animals, derive a hierarchy that groups those lifelines that show similar motion properties throughout their motion process.

Assuming that bearing might be an interesting motion property for a migrating animal, one could now develop a lifeline similarity measure for the motion azimuth. One way of doing this is to compare the MPO's motion azimuth over time. For every time-step (B) and for every pair MPO the respective motion vectors can be assessed with respect to motion azimuth similarity (C): same direction = 1, perpendicular direction = 0.5, and opposite direction = 0. This results in an average motion azimuth similarity for every pair in the group, collected in a similarity matrix (D). Finally, clustering algorithms identify similarly moving animals and outliers (E).

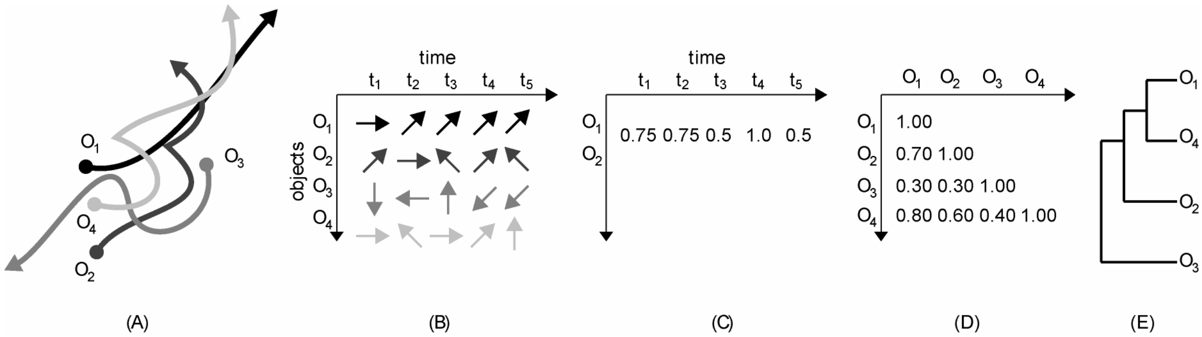

Figure 1: Clustering lifelines. Given the lifelines of four MPOs (A), a motion azimuth vector is computed for every time-step $t_1, \ldots, t_5$ (B). The comparison of the motion azimuth for every MPO with every other MPO (C) produces a motion azimuth similarity matrix (D) that is thereafter used for cluster analysis. Since O_3 moves most of the time in the opposite direction then all the other MPOs, it is isolated in the dendrogram (E).

This procedure is, of course, not limited to motion azimuth, but could also be applied for speed, acceleration, or sinuosity.

5.0 DISCUSSION

The inclusion of derived motion properties such as speed, acceleration, motion azimuth, and sinuosity provides more insights in the motion process than only looking at the positions of the fixes. The identification of the spatio-temporal event of many objects performing the same manoeuvre at the same time (i.e. a concurrence pattern), is not possible with only looking at the fixes of the involved trajectories alone. Or regarding the descriptive clustering example, considering a similarity measure basing on bearings is arguably more suitable for the understanding of the navigation of, for example, migrating caribou than just looking at the spatio-temporal clustering of the involved lifelines.

Furthermore, investigating the derived motion properties in a dynamic way, constantly proceeding throughout the lifeline, allows a deeper understanding of spatially and temporally implicit information captured in a trajectory than only looking at totalling descriptive measures such as a mean Euclidean distance. If the investigated lifelines represent, for instance, annual migration paths of animals, it may be very relevant when and where the MPOs 'were close', 'moved in a similar direction' or 'converged'. Of course, such detailed patterns rely on the availability of high-resolution tracking data. However, given the expected increase of such data, analysts should seize this opportunity and come up with new analytical concepts.

Both presented approaches offer (semi-)automated data mining processes that produce quantifiable and reproducible results, and thus offer a more objective analytical process than visual inspection, basing on the skills and knowledge of an analyst. However, as shown in Laube and Purves (in press), pattern detection must also be undertaken with care as it is only a starting point for subsequent research in investigating the interrelations and processes from the discovered patterns.

A further opportunity not addressed in this paper, is the integrative power of GISystems allowing the investigation not only of geometric trajectories but also including the underlying geography that fosters the motion processes. A predominant motion direction of a moving animal may easily be explained by the orientation of the valley the animal moves in. Thus, motion patterns or similarity measures may be defined that include not only the geometry of observation points, but also attribute information referring to the underlying geography. Since both, *retrieval by content* and *descriptive modelling,* are not restricted to the human's visual perception of maximally three dimensions, such approaches are very much suited to extend the analysis of motion beyond solely its spatial properties.

6.0 CONCLUSIONS

I argue that *retrieval by content* and *descriptive modelling* should be established alongside EDA as alternative approaches for the analysis of dense spatio-temporal data. This is especially true for distributed databases covering peoples' everyday movement, which probably extend beyond what human analysts can keep track of by pure visual inspection.

A deeper understanding of motion requires not only the investigation of the coordinates of trajectories but also the inclusion of derived motion descriptors such as speed, acceleration, motion azimuth, and sinuosity. Furthermore, spatially and temporally implicit data mining approaches, which constantly integrate the 'who?', 'where?', 'when?', and 'what?' moving through the lifeline, are preferable over approaches that collapse lifelines in a single statistical measure.

ACKNOWLEDGEMENTS

Patrick Laube's current work is funded by the Swiss National Science Foundation, grant no. PBZH2-110315. The author would furthermore like to acknowledge the invaluable input from Pip Forer, David O'Sullivan and Todd Dennis and three unnamed reviewers, all of University of Auckland.

REFERENCES

Anselin, L. (1998) Exploratory spatial data analysis in a geocomputational environment. *Geocomputation: a primer*, New York: John Wiley & Sons, pp. 77-94.

Forer, P. (1998) Geometric approaches to the nexus of time, space, and microprocess: Implementing a practical model for mundane socio-spatial systems. *Spatial and Temporal Reasoning in Geographic Information Systems*, Oxford: Oxford University Press, pp. 171-190.

Forer, P. (2002) Timelines, environments, and issues of risk in health: the practical algebra of x, y, t, a. *GIS for Emergency Preparedness and Health Risk Reduction*, Berlin Heidelberg: Springer, pp. 35-60.

Gudmundsson, J, van Kreveld, M. and Speckmann, B. (2004) Efficient Detection of Motion Patterns in Spatio Temporal Data Sets. *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, Washington DC, USA 2004. pp. 250-257.

Hägerstrand, T. (1970) What about people in regional science. *Papers of the Regional Science Association*, 24, pp. 7-21.

Hand, D., Manilla, H. and Smyth, P. (2001) *Principles of data mining*. Cambridge: MIT Press.

Kraak, M.-J. and Koussoulakou, A. (2004) A visualization environment for the space-time-cube. *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling*, Berlin Heidelberg: Springer, pp. 189-200.

Kwan, M.-P. (2000) Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C*, 8:1-6, pp. 185-203.

Laube, P. and Imfeld, S. (2002) Analyzing relative motion within groups of trackable moving point objects. *Geographic Information Science, Second International Conference, GIScience 2002*, Boulder CO, USA 2002, Lecture Notes in Computer Science, 2478, Berlin Heidelberg: Springer, pp. 132-144..

Laube, P., and Purves, R. S. (in press) An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems.*

Laube, P., Imfeld, S. and Weibel, R. (2005) Discovering Relative Motion Patterns in Groups of Moving Point Objects. *International Journal of Geographical Information Science*, 19:6, pp. 639-668.

Laube, P., Van Kreveld, M. and Imfeld, S. (2004) Finding REMO - detecting relative motion patterns in geospatial lifelines. *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling,* Leicester, UK 2004, Berlin Heidelberg: Springer, pp. 201-214.

Mark, D. M. (1998) Geospatial Lifelines. *Integrating Spatial and Temporal Databases, Dagstuhl Seminars, no. 98471*.

Miller, H.J. and Han, J. (2001) Geographic data mining and knowledge discovery: An overview. *Geographic data mining and knowledge discovery*, London: Taylor and Francis, pp. 3-32.

Ng, R.T. (2001) Detecting outliers from large datasets. *Geographic data mining and knowledge discovery*, London: Taylor and Francis, pp. 218–235.

Peuquet, D.J. (2001) Making Space for Time: Issues in Space-Time Data Representation. *GeoInformatica*, 5:1, pp. 11-32.

Qu, Y., Wang, C., and Wang, X. S. (1998) Supporting fast search in time series for movement patterns in multiple scales. *Proceedings of the 7th International Conference on information and Knowledge Management*, Bethesda, ML, USA, 02-07 November, 1998, ACM Press, New York, NY, 251-258.

Rafiei, D. and Mendelzon, A. (2000) Querying Time Series Data Based on Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 12:5, pp. 675–693.

Shim, C.-B. and Chang, J.W. (2003) Efficient similar trajectory-based retrieval for moving objects in Video Databases. *CIVR 2003*, Lecture Notes in Computer Science, 2728, Berlin, Heidelberg: Springer, pp. 163-173.

Sinha, G. and Mark, D.M. (2005) Measuring similarity between geospatial lifelines in studies of environmental health. *Journal of Geographical Systems*, 7, pp. 115-136.