VISUAL DATA MINING: BUILDING 'SPACE-TIME-CUBE' SOCIO-ECONOMIC CONDUITS FOR PROVINCIAL LEVEL MODELLING OF THE REGIONAL DEVELOPMENT PROCESS IN SPAIN 1955-1977

ROBERT J. ABRAHART AND ROY P. BRADSHAW
School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD.
Tel: +44 (0)115 84 66145; Fax: +44 (0)115 951 5249; Email: bob.abrahart@nottingham.ac.uk

BIOGRAPHY

Lecturer in Geographical Information Science (University of Nottingham: UK). Elected Chair of RGS-IBG Geographical Information Science Research Group; Associate Editor for Water Resources Research; Co-Editor of "GeoComputation" (Taylor & Francis, 2000) and "Neural Networks for Hydrological Modelling" (A.A. Balkema Publishers, 2004); Principal research interests include GIS and Neural Network Modelling.

INTRODUCTION

Leading economists such as Paul Krugman and Jeffrey Sachs continue to emphasise the substantial influence and importance of geographical factors on the process(es) of economic development. However, most existing models of economic development, including those that incorporate some element of spatial differentiation are still broad simplistic approximations. The "Stages of Growth" Model (Rostow, 1960) is one of the earliest and much criticised examples that still persists in a tacit and implied or accepted manner. This model states that since successful societies are recorded as having passed through a particular sequence of development stages then other societies that wished to develop would also need to pass through the exact same sequence of events. The "Inverted "U" Hypothesis" (Kuznets, 1955; Williamson, 1965) is a similar concept from a similar period which states that inequalities in the national [regional] distribution of income will worsen during the initial stages of economic development and that the differences in inequalities will improve as the nation [region] reaches higher stages of development. The fundamental assertion that all countries must pass along the same route to achieve development has however been subjected to strong and continued disapproval. Krugman (1996) has instead proposed a recent alternative rationalization based on countervailing economic forces and attempts to explain the overall pattern of development in terms of a "developed core region" and a "disadvantaged peripherial region". His model embraces the earlier traditions of Myrdal (1957) and Hirschman (1958) in which centripetal forces, such as the pre-existence of a skilled labour force, are factors that would tend to promote geographical concentration, whereas centrifugal forces, such as high land rents, are factors that would tend to promote geographical dispersion. The socio-economic spatial outcome will thus depend on the movement of items (such as goods and workers or resources) and the differing opportunities that exist (such as transport costs and wage levels) across the land; such that and the development process is based on "self-organizing economies" in which economic activities become concentrated in the richer spatial unit(s). This model benefits from the incorporation of spatial elements but is nonetheless limited with respect to the assumption of a two tier spatial setup [one rich and one poor]. The spatial and temporal philosophical position is also based on two-dimensional geometries and mind-sets and does not accept that differential economic development could occur in progression sequences or destination points at individual sites in time or space.

VISUAL DATA MINING

The ongoing need to digest and understand enormous volumes of digital information has challenged our power to gain useful spatial insights and to acquire important geographical knowledge that would otherwise be lost to the scientific world. The concurrent development of spatiotemporal datasets and location-aware computing present important opportunities for fresh scientific discoveries to appear based on the methodologies of *Data Mining* [DM] and *Knowledge Discovery in Databases* [KDD]. Numerous technical definitions exist for the different fields of research that are involved but one common thread is recognised: DM is the core non-trivial process in KDD that will [or at least has the potential to] map or convert measurement records into valid, novel, useful and understandable patterns or knowledge that is related to the matter in hand. DM is used to extract meaningful information out of massive datasets in an automated or semi-automated manner and it is axiomatic that the usefulness of different sorts of extracted information will be related to the specific needs and requirements of individual end-users. DM tools and strategies are in most cases used for the identification and

extraction of structures e.g. trends, clusters, periodicities, associations, correlations, quantizations and granularities. The resultant structures are often suitable candidates for the beneficial application of information visualization [InfoViz] techniques in which 2-D or 3-D colour graphics and/or animation are used to depict the structure of our information in a non-traditional manner. End users will also be permitted to navigate through the visualizations and be empowered to perform certain modifications based on graphical interaction - such that the phenomenal capabilities of human observation and perception can be used to explore and to interpret what is often a set of complex abstractions in a direct manner. This paper seeks to combine the two approaches and is about the development and testing of a dedicated and integrated suite of traditional DM and InfoViz technologies to meet to the challenges of spatiotemporal KDD. This tight coupled integration presents a novel method that has high potential and is hereinafter referred to as *visual data mining* [VDM].

SPATIOTEMPORAL DATA MINING

The expression 'spatiotemporal data mining' is in this paper used to refer to the identification and extraction of implicit knowledge, spatial and temporal relationships, or other relevant patterns of interest that have no explicit existence in the spatiotemporal database i.e. are not stored. Most recent data mining efforts in the field of geographical information science have been applied to static representations of space; but real geographic patterns and relationships and phenomena will evolve in both spatial and temporal senses. Thus space-time investigation is central to our understanding of geographic process and events such that if what might otherwise be obscured but useful geographical information could be extracted out of spatiotemporal datasets it is manifest that such material could be used to bring about improved possibilities with respect to the forecasting or prediction of spatial processes or events. It was a popular practice in earlier investigations to treat *spatial* data mining and *temporal* data mining as a pair of separate and independent challenges such that limited attention has been paid to the development of efficient strategies for the exploration of spatiotemporal datasets, in part due to the larger size of spatiotemporal datasets, but also due to the complexities of the different datasets that might perhaps be involved e.g. in terms of representations, structures and types. For a recent discussion of relevant algorithms see Ester et al. (2001) or Roddick & Spiliopoulou (2002). Indeed, most current methods that are applied to geographic datasets, use simple representations of geographic objects and spatial relationships e.g., point objects, polygons, and distances in 'Euclidean Space' [ES] (Buttenfield et al. 2001). For space-time data models it also is standard practice for time and space to be preserved as the principal dimension(s) in which spatiotemporal relationships and spatiotemporal processes are permitted to exist; hence other attributes in the spatiotemporal dataset are considered to be subordinate to the integrated spatial and temporal dimensions.

SPACE-TIME-CUBE VISUALIZATION

The expression 'space-time-cube visualization' is in this paper used to refer to the modern utilisation of earlier structures and concepts that are experiencing a popular resurgence due to better tools and increased demand for better scheduling of transport services and lifestyle activities in time and space. Hägerstrand (1970) proposed a theoretical structure that could be used to investigate the constraints that affect an individual's presence in space and time and to depict the activities of individuals in a space-time context. The original concept was adopted and developed to investigate individual human activities and behaviours (Carlstein et al., 1978; Carlstein, 1982; Golledge & Stimson, 1997). Earlier studies used the time-geographic concepts in a theoretical or semi-quantitative manner (Parkes and Thrift 1980). More recent studies involve the development of appropriate techniques and data structures for the digital description and computational visualization of numerous individual space-time-paths e.g. Kwan (2000a; 2000b; 2003); Shaw & Wang (2000); Wang & Cheng (2001); Frihida *et al.* (2002); and Miller (2004). The space-time-cube [STC] concept considers time and space to be of equal standing and to be inseparable factors in the investigation of working activities and recreational pastimes. STC adopts a 3-dimensional orthogonal viewpoint. The horizontal axes are used to record the position and locational changes of objects. The vertical axis is used to provide an ordered and synchronized sequence of events. There are three fundamental analytical concepts (Figs. 1 & 2). The *space-time path* is the trajectory of an individual's movements in physical space over time. This provides detailed information about the spatial and temporal characteristics of a particular individual, including starting/ending points and times, location of particular activities, sequential order of events, as well as relative position of specific events that occurred throughout the lifespan of each individual. The *space-time prism* comprises that set of possible locations that a person could travel to in continuous orthogonal space-time coordinates (Lenntorp, 1976). The *potential path area* is the space-time prism projected onto a two-dimensional platform, thus given that a space-time path can be used to represent the historical movements of each individual person, the space-time prism and potential path areas depict possible accessible space and region under a certain set of constraints. Figure 3 illustrates various spatiotemporal relationships that can be investigated between several different individual time-space-paths in the STC modelling environment.

BUILDING 'SPACE-TIME-CUBE' SOCIO-ECONOMIC CONDUITS

This paper will highlight the potential strengths and proven benefits of VDM. The paper will also demonstrate improved 'Data-to-Knowledge' [D2K] acquisition possibilities with respect to spatiotemporal trajectories. The reported investigation combines 'spatiotemporal data mining' and 'space-time-cube visualization' in a real world context. It also builds on previous studies and extends the space-time-cube concept from individual people in ES to aggregated statistical units in socio-economic NES. Two unconnected technologies are integrated to develop a practical solution to the challenges of modelling and understanding the national or regional development process(es). Miller & Han (2001) suggested that existing data structures, queries, indexes and algorithms should be expanded to cope with more complicated geographical objects (e.g., objects that move or evolve over time) and relationships (e.g., 'Non-Euclidean Space' [NES] distances, directions, and connectivities). It is also important to stress that conventional data mining methods in artificial intelligence do not recognize the uniqueness of spatial and temporal dimensions such that there are several possible routes to improvement that could be adopted. It is axiomatic that the most difficult option would be to develop a set of innovative or modified or specialist spatiotemporal mechanisms that could perform spatial data mining operations i.e. tools that that could be used, to the largest extent possible, to discover the rich set of spatial and temporal patterns or relationships and that are embedded in unsophisticated spatiotemporal datasets [Openshaw (1998) provides a list of potential requirements]. The simpler option that is instead demonstrated in this paper comprises a sequence of different operational procedures that are applied to investigate the regional development process, working at the provincial level, for Spain 1955-1977.

Step 1: DM methods are used to convert a complex multi-dimensional spatiotemporal socio-economic dataset from a set of statistics collected in ES to a stack of classified point pattern mappings organized in 'Socio-Economic' NES [see Agarwal & Abrahart (2003) for further details on non-linear multi-dimensional reduction using the self-organizing feature map (SOFM) algorithm (Kohonen, 1997)]. The modeling inputs comprised nine indicators for the 47 mainland provinces of Spain [Table 1] provided in the manner of eleven annual returns that spanned the period 1955 - 1977. Fig. 4 depicts a set of selected provincial trajectories viewed in socio-economic 2D-NES.

Step 2: Traditional statistical descriptors are computed on each point pattern dataset at each time step to provide a second set of point pattern datasets e.g. spatial median.

Step 3: InfoViz methods are applied to construct a spatiotemporal socioeconomic virtual world model of the development process in 3D-NES. The model contains space-time-path trajectories and statistical conduits that can be switched on or off. It offers a useful tool that can be 'experienced' or 'interrogated' and thus used to challenge traditional models and mindsets with respect to the 'transition process' in a meaningful and/ or applied context.

The main advantage of our approach is that it permits a number of individual regions, or smaller spatial units, to be included within the same model. Each region can thus be at a different stage of modernisation such that it is possible to probe and evaluate the changing socio-economic relationships between spatial units that follow variable convergent or divergent trajectories over time and space as units advance through different stages of economic development. Twelve provincial case studies will be considered [Table 2].

REFERENCES

Agarwal, P. and Abrahart, R.J. 2003. "Towards an integrated data mining environment for urban analysis using Self-Organising Maps and Geographical Information Systems". *GISRUK 2003: Proceedings GIS Research UK, 11th Annual Conference, City University, London, 9-11 April 2003.* pp. 239-243.

Buttenfield, B., Gahegan, M., Miller, H., Yuan, M. 2001. *Geospatial data mining and Knowledge Discovery.* UCGIS white paper on Emergent Research Themes.

Carlstein, T., Parkes, D. and Thrift, N. (eds.) 1978. *Timing Space and Spacing Time (Vol. 2): Human Activity and Time Geography.* New York: John Wiley & Sons.

Carlstein, T. 1982. *Time Resources, Society and Ecology.* London: George Allen and Unwin.

Ester, M., Kriegel, H-P and Sander, K. 2001. "Algorithms and applications for spatial data mining". Chapter 7 in: Miller, H.J. and Han, J. (eds.) 2001. *Geographic Data Mining and Knowledge Discovery.* London: Taylor and Francis. pp. 160-187.

Frihida, A., Marceau, D. and Thériault, M. 2002. "Spatio-temporal object-oriented data model for disaggregate travel behavior". *Transactions in GIS.* 6(3): 277-294.

Golledge, R. and Stimson, R. 1997. *Spatial Behavior: A Geographic Perspective.* New York: The Guilford Press.

Hägerstrand, T. 1970. "What about people in regional science?" *Papers of the Regional Science Association*. 24: 7-21.

Hirschman, A. 1958. *The Strategy of Economic Development*. New Haven [CT]: Yale University Press.

Kohonen, T. 1997. *Self-Organising Maps*. 2nd Ed. Berlin: Springer

Kwan, M.-P. 2000a. "Human Extensibility and Individual Hybrid-Accessibility in Space-Time: A Multi-Scale Representation Using GIS". In D. Janelle and Hodge, D. (eds.) *Information,Place, and Cyberspace: Issues in Accessibility*. Berlin: Springer-Verlag. pp. 241-256.

Kwan, M-P. 2000b. "Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: A methodological exploration with a large data set". *Transportation Research C*. 8:185-203.

Kwan, M-P. 2003. "Geovisualisation of Activity-Travel Patterns Using 3D Geographical Information Systems". *Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, 10-14 August 2003*.

Krugman, P.R. 1996. *The Self-Organizing Economy*. Cambridge [MA]: Blackwell Publishers.

Kuznets, S. 1955. "Economic Growth and Income Inequality". *American Economic Review*. 445: 1-28.

Lenntorp, B. 1976. *Paths in Time-Space Environments: A Time Geographic Study of Movement Possibilities of Individuals*. Lund: Gleerup.

Miller, H. 2004. *A Measurement Theory for Time Geography*. [DRAFT] http://www.geog.utah.edu/~hmiller/papers/time_geog_measure.pdf

Miller, H.J. and Han, J. (eds.) 2001. *Geographic Data Mining and Knowledge Discovery*. London: Taylor and Francis.

Myrdal, G. 1957. *Economic Theory and Under-developed Regions*. London: Duckworth.

Openshaw, S. 1998. "Geographical data mining: key design issues". *GeoComputation'99: Proceedings Fourth International Conference on GeoComputation, Mary Washington College, Fredericksburg, Virginia, USA, 25-28 July 1999*. [CD-ROM]
http://www.geocomputation.org/1999/051/gc_051.htm

Parkes, D.N. and Thrift, N.J. 1980. *Times, Space and Places: A Chroneographic Perspective*. UK: Wiley & Sons.

Roddick, J.F. and Spiliopoulou, M. 2002. "A Survey of Temporal Knowledge Discovery Paradigms and Methods". *IEEE Transactions on Knowledge and Data Engineering*. 14(4): 750-767.

Rostow, W.W. 1960. *The Stages of Economic Growth: A Non-Communist Manifesto*. Cambridge: Cambridge University Press.

Shaw, S-L and Wang, D. 2000. "Handling Disaggregate Spatiotemporal Travel Data in GIS". *GeoInformatica*. 4(2): 161-178.

Wang, D. and Cheng, T. 2001. "A spatio-temporal data model for activity-based transport demand modeling". *International Journal of Geographic Information Science*. 15: 561-585.

Williamson, J.G. 1965. "Regional Inequality and the Process of National Development: A Description of the Patterns". *Economic Development and Cultural Change*. 13: 3-54.

Yu, H. 2004. "Spatio-temporal GIS Design for Exploring Interactions of Human Activities" *GIScience 2004: Proceedings Third International Conference on Geographic Information Science, Adelphi, Maryland, USA, 20-23 October 2004*. http://www.ucgis.org/UCGISFall2004/studentpapers/files/yu.pdf

1. Construction and public works
2. Food, drink and tobacco industries
3. Chemicals and chemical product industries
4. Metallurgical industries
5. Agriculture
6. Transport and communications
7. Commerce
8. Financial services and banking
9. Property incomes, rentals

Notes:

- The above variables constitute the main sectors of the economy, particularly at the time of the study which is 1955-1977.
- Change in any one of these variables (sectors) has a significant impact on the provincial economy as a whole.
- No major economic sectors are omitted from our list of indicators.
- Minor economic sectors and those sectors not common across the country are omitted e.g. fishing only found in the coastal provinces.

Table 1: List of socio-economic indicators

1. Pontevedra (poor, rural, peripheral)
2. Madrid (rich, industrial/commercial, central)
3. Cordoba (agricultural, southern, Andalucian)
4. Guipuzcoa (rich, industrial, Basque)
5. Gerona (rich, industrial, Catalan)
6. Albacete (average wealth, agriculture and some industry, central)
7. Caceres (stagnating agricultural province, central)
8. Oviedo (declining industrial area, peripheral)
9. Malaga (poor but growing tourist area, peripheral)
10. Navarra (fast growing industrial and commercial area, peripheral)
11. Zaragoza (traditional agriculture and growing industry, regional capital)
12. Logrono (fast growing agricultural province)

Notes:

- These provinces were considered to be good "diagnostic" examples with regard to our investigation into the trajectories of development:

Table 2: List of illustrative cases studies to be considered

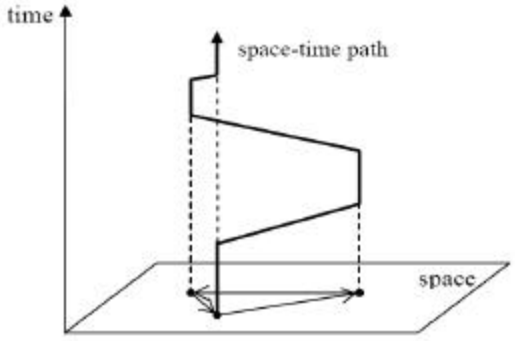


Fig.1: Space Time Path
[Source: Yu (2004)]

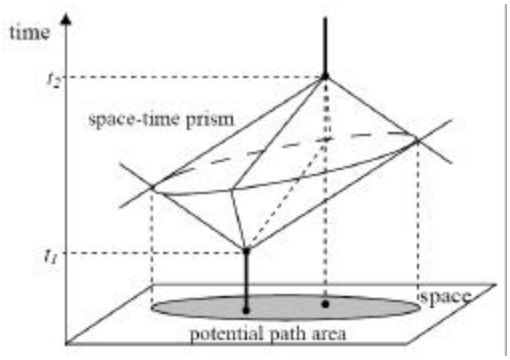
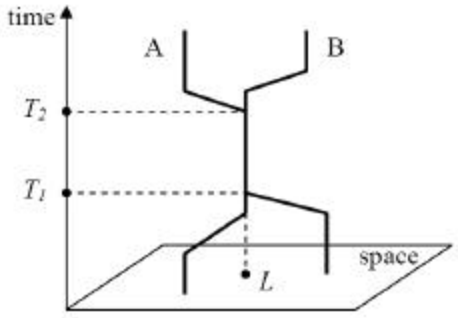
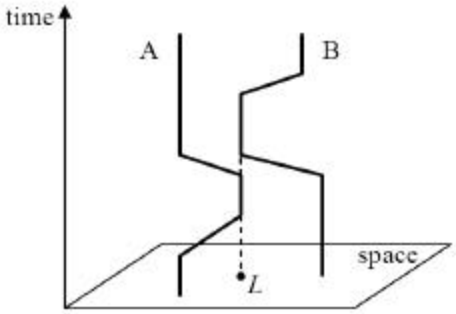


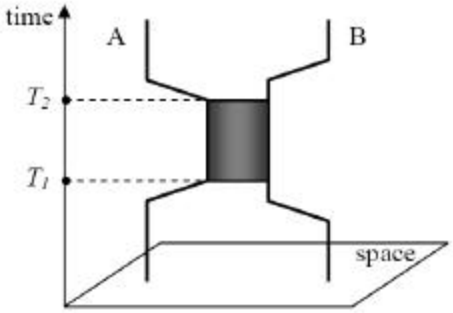
Fig.2: Space Time Prism/ Potential Path Area
[Source: Yu (2004)]



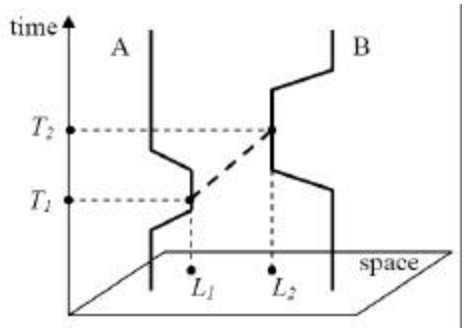
(a) Co-location in space and time



(b) Co-location in space



(c) Co-location in time



(d) No co-location in space or time

Fig.3: Spatio-temporal relationships
[Source: Yu (2004)]

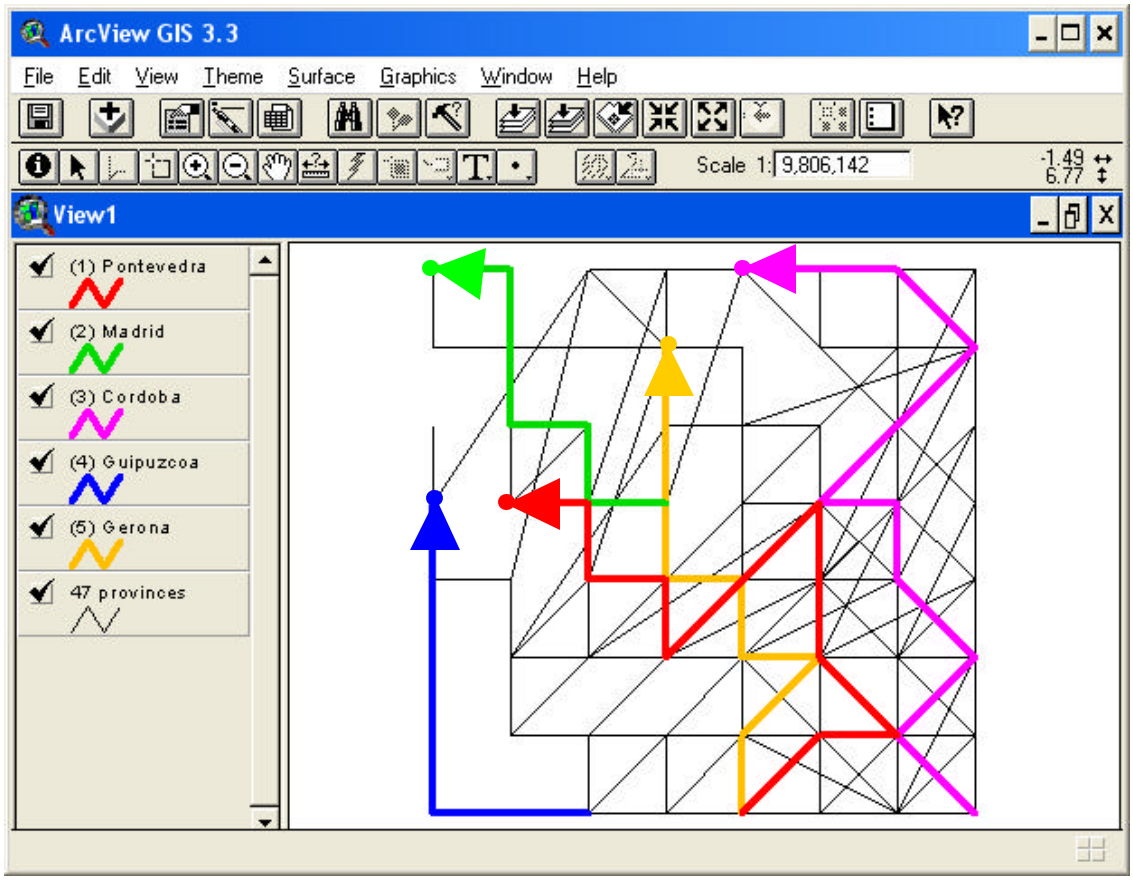


Fig. 4: Socio-economic trajectories for some selected mainland provinces (1955-1977)