Papers of the Applied Geography Conferences (2007) 30: 10-19

VISUAL EXPLORATION OF US DOMESTIC AIRLINE MARKET STRUCTURES

Jun Yan
Department of Geography and Geology
Western Kentucky University
Bowling Green, KY 42101

Jean-Claude Thill
Department of Geography and Earth Sciences
University of North Carolina at Charlotte
Charlotte, NC 28223-5973

1. INTRODUCTION

With the wide-spread adoption of information technology and the global diffusion of GIS, a large volume of digital spatial data continues to accumulate. In the field of domestic air travel, aggregate data at a fine geographic resolution have become available. For instance, ever since the enactment of the Airline Deregulation Act of 1978, all large certified domestic air carriers conducting scheduled passenger operations (except helicopter carriers) in the US are required to participate in a Passenger Origin and Destination Survey (GAO, 2003). The reports submitted by each air carrier are then compiled by the Office of Airline Information (OAI) of US Bureau of Transportation Statistics (BTS) to form a database, called the Airline Origin and Destination Survey (DB1B). This database contains a 10 percent sample of all airline tickets from all the reporting carriers by quarter each year. DB1B data tables from 1993 to 2005 are currently accessible to the public at the BTS website (http://www.bts.gov). DB1B data consist of three files, DB1BCoupon, DB1BMarket, and DB1BTicket, of which DB1BMarket is the most useful for studying US domestic airline markets. DB1BMarket contains information related to passenger travel by air from an origin airport to a destination airport. An origin-destination pair is often characterized as a directional market in that it is defined by the first departure airport on a ticket and the ultimate arrival airport. In this sense, the study of airline directional markets fits in the framework of spatial interaction research. In DB1BMarket files, other main data items include number of passengers, fare, and distance for each directional market. Due to its fine geographic details (at the airport level), this database can reasonably be expected to contain a wealth of information that could provide unparalleled insights into the formation of airline market structures over time and across space. DB1BMarket is the significant data source for studying US domestic air traffic patterns and airline market structures. In fact, a more widely used dataset in airline research, known as the Consumer Air Fare Report (http://ostpxweb.dot.gov/aviation/), is a derivation from this database.

The ability to explore the patterns in high-dimensional spatial interaction data of various descriptions from large digital databases, such as DB1B, ought to be emphasized if one wishes to gain from the data any useful information relevant to the underlying spatial interaction processes. Many of these newly-available data have not yet been as thoroughly examined as they should be, largely due to the lack of the research in exploration of spatial interaction data, even though exploratory techniques to representing spatial interaction data date back to Ullman's (1957, 1959) seminal analysis of US commodity flows and Chicago Area Transportation Study. In Ullman's method, movement is represented by "desire lines" or aggregated to rectangular flow bands with width proportional to flow magnitudes. However, movement and flow mapping quickly becomes quite problematic as soon as the size of the spatial system becomes greater than trivially small. The recognition of spatial structures is hence gravely hindered by purely visual approaches to the exploration of spatial interaction data to the point that some form of data compression is advocated to reduce the apparent complexity in large flow matrices.

To cope with increased data complexity and volumes in geography, as in many other disciplines, attention has recently been paid to approaches conceived in scientific computing to analyze large databases (Longley *et al.*, 1998; Openshaw and Abrahart, 2000). With fast-growing computational power at the disposal of researchers, and a huge collection of digital geographic data, this makes a lot sense. Methods able to cope with fewer assumptions are well suited in geography, considering the complex nature of most geographic research problems. As Peter Gould (1981) put it, "letting the data speak for themselves" is becoming a guiding principle for modern spatial analysis. In the digital era, we are often faced with situations that are fundamentally different from the conditions under which many existing analytical methods are usually applied. Data might contain important trends and relationships of which we are unaware, and which may violate some assumptions required by existing methods. Exploratory spatial data analysis (ESDA) methods play increasingly important roles in analyzing new data, especially at the beginning stage of analysis. Via exploration, data not yet fully understood can be classified, summarized and formed into high-level structures so that they can subsequently be developed into useful concepts for later in-depth spatial analysis and modeling use. Recent advances in geographic data mining (GDM), geographic knowledge discovery (GKD) and geographic visualization (GVis) represent the collective efforts in spatial science to extend the functionality and applicability of ESDA by offering new mechanisms to filter through large geographic databases (Buckley, *et al.* 2000; Buttenfield, *et al.* 2000; MacEachren, 2001; Miller and Han, 2001).

This paper adopts an approach for visually and computationally exploring geographic patterns embedded in large spatial interaction databases, such as DB1BMarket. A special type of artificial neural network, Self-Organizing Maps (SOM), is used for the purpose of handling data complexity in both theme dimensionality and data volume. Following this introduction, section 2 of this paper gives a review of research of data exploration in spatial interaction. Section 3 then briefly explains the SOM method; Section 4 presents some results from use of DB1BMarket data. The final section gives the conclusions.

2. DATA EXPLORATION IN SPATIAL INTERACTION SYSTEMS

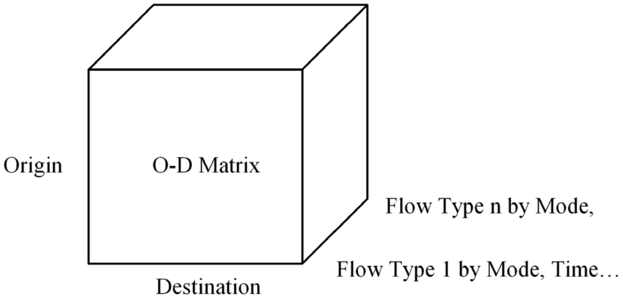
In the past, research in spatial interaction data exploration mainly went in two directions. The first approach focused on the improvement of visual techniques. Waldo Tobler is one of a few scholars who continuously contributed to this field. In the 1970s and 1980s, he published a series of papers that ally mathematical modeling and cartographic mapping methods (Tobler, 1976, 1979, 1987, 1988). The approach serves to visualize the surface of net flows between origins and destinations. The surface is displayed as a field of vectors, which approximate the gradient of a scalar potential computed from the relative net exchanges of flows. Tobler refers to this innovative data exploration method as "winds of influence", by using the earth science analogy of 'pressure field' that gives rise to winds. In addition to Tobler's early work, Marble and colleagues (Marble *et al.*, 1995, 1997) made a significant contribution to the exploratory analysis of spatio-temporal interregional flows with a new approach that makes extensive use of scientific visualization. Their approach implements some dynamic graphics-based tools, which allows analysts to map various interregional flows, to examine both the total set and subsets of the flows, and to compare flow volumes with selected characteristics of origins and destinations. However, cartographic mapping becomes unwieldy when the number of pairwise interactions is very large and multidimensional flow matrices are under examination. To facilitate the visualization and analysis of geographic structures in multiple interaction flows, Marble and his colleagues adopted a data projection method, projection pursuit (PP) that reduces the dimensionality of flow matrices. The second direction involves some form of data compression to reduce the complexity found in large flow matrices. In order to identify structures in large geographic databases with multidimensionality, a crucial task is to compress the number of attributes (*data projection*) and the number of data vectors (*data quantization*) without losing too much useful information. Uninteresting data or attributes need to be filtered out to retain essential structures and to group similar data. Conventional multivariate statistical methods, such as factor analysis (FA), principle component analysis (PCA), multidimensional scaling (MDS), and PP, basically serve the needs of data projection, while k-means deals with data quantization.

As early as the 1960s, some spatial scientists explored the effectiveness of reducing the complexity in spatial interaction data to uncover essential relationships within transportation flow matrices. Pioneering work was mainly done by Berry (1962, 1966), who developed three FA approaches to identify the major commodity flow patterns of India from 63 36x36 commodity flow matrices. The first approach consisted of an R-mode analysis, whereby flow destinations are factored to identify clusters of destinations with similar profile of incoming flows. Q-mode analysis accomplishes the same for flow origins. The third approach extracts structures among thematic dimensions (for instance, commodity types) on each origin-destination pair (*dyad*). Black (1973) later used the term “Dyadic Factor Analysis” to describe the latter modality of application of FA. However, one restriction of methods like PP, FA and MDS is that they are only methods of data projection by identifying a smaller number of latent components that represent the fundamental structures. To achieve both data projection and data quantization, a common solution is to carry out them sequentially: either data projection first followed by data quantization, or the other way around. The problem with this approach is that the conclusions inferred from the second step are conditional upon the outcome of the first step. Methods capable of both tasks simultaneously do not exhibit this flaw and should be given preference since many spatial interaction data, such as DB1BMarket, are often constituted in large databases with high dimensionality, and thus both data quantization and data projection are essential. As shown in Figure 1, in addition to the origins and destinations, many spatial interaction systems are complicated by a third dimension, such as flow type, transport mode, transport time, or any other quality or quantity of spatial interaction.

3. METHOD

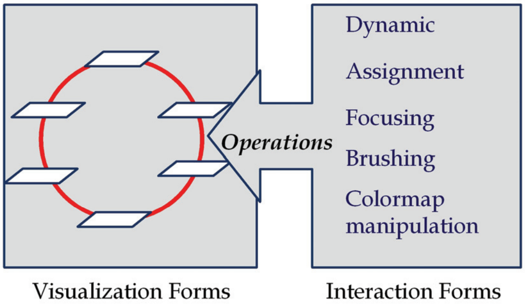
In this paper, an integrated ‘*computational and visual*’ approach known as the VIsual Spatial Interaction DAta MINer (VISIDAMIN) is used to facilitate data exploration and knowledge discovery in large spatial interaction databases (Yan and Thill, 2007a; 2007b). The computational method, Self-Organizing Maps (SOM), is a special kind of competitive neural network. The principle behind SOM is rather simple. Neurons in the output layer compete with each other based on a similarity measure, such as Euclidian distance. The winner earns the right to represent the input data vector on the basis of a certain dissimilarity measure in the attribute space (Kohonen, 2001). SOM allows the winner neuron, as well as the neurons in its neighborhood, to learn the new input, match and adapt so that each neuron gradually specializes to represent similar input data. Simply put, it works like a flexible net that is capable of folding onto the ‘*cloud’* formed by input data. Consequently, input data vectors with more similarity are assigned close to each other on the predefined map grid. It thus preserves the natural order in the attribute space of the input data.

FIGURE 1
DIMENSION TUBE OF SPATIAL INTERACTION DATA



With SOM as the core DM engine, a prototype interactive visual data mining (VDM) environment is developed, following the framework of MacEachren *et al.* (1999). In this environment various visualization forms are linked through a range of user interaction techniques (Figure 2). Each visualization form can be seen as a different view of the data under study. For instance, '*cartographic map*' displays the geographic distributions of interaction flows among sets of origins and destinations, while '*SOM component plane*' is for suggesting the properties of clustered structures in high dimensional attribute space. By linking them together, it is possible to examine how the clustered structures detected in attribute space are geographically defined. In addition to linking, some other interaction techniques are also implemented, including assignment, colormap manipulation, focusing, and brushing. Besides SOM feature maps and cartographic maps, this environment also implements some additional exploratory techniques to facilitate the understanding of original data as well as the evaluation of the results. Examples include scatter plot, star coordinate plot and parallel coordinate plot.

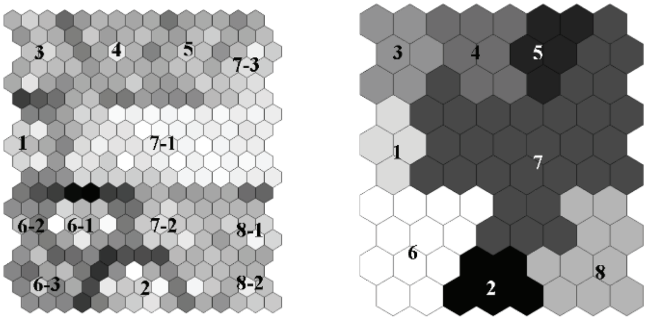
FIGURE 2
THE VDM FRAMEWORK



4. RESULTS AND DISCUSSIONS

Case studies are carried out on 2002 DB1BMarket data file. The training input consists of a dyadic matrix wherein each column containing either the share of the market held by a particular airline (as a percentage) or the average market airfare charged by this airline (in current US dollars). Airports located in the same metropolitan area are pooled together. Only markets between the (2002) 278 metropolitan areas in the contiguous US are considered, and so potentially there are 278 by 277, a total of 77006, directional markets. In 2002, a total of 34 certified air carriers operated inside the US. As a result, there are also 68 data fields in the final input data table. The primary aim of data complexity reduction is to find some prototype data that can represent the original data reasonably well. On this premise, the structures identified in the representative data can be assumed true in the original data as well. In SOM, a '*distance matrix*' technique can be used to visualize the patterns in SOM prototype data. In principle, colors (or other visual variables) can be assigned to each neuron on the basis of a certain statistic of inter-neuron distances, *e.g.* minimum, median, or maximum of the distances to its neighbors. In Figure 3a (*U-matrix*), the darkness of the map units (hexagons) denotes the actual distance of SOM neurons to their respective neighboring neurons: the darker the hexagon, the larger the distance. A cluster is viewed as an area on the map with low distances separated by borders consisting of high distances (thus large distances between neurons). Several clusters can be identified as marked in Figure 3a. A similar pattern is observed in Figure 3b, the result of a k-means clustering solution based on the prototype data. Of all the possible numbers of clusters, the clustering with k = 8 is reported here since it results in the best Davies-Bouldin index. The Davies-Bouldin index is a validity index for evaluating clustering results. It is a function of the ratio of the sum of within-cluster difference to between-cluster separation. The lower the index value the better the clustering (Davies and Bouldin, 1979).

FIGURE 3
CLUSTERS BASED ON ONLY MARKET SHARE INFORMATION



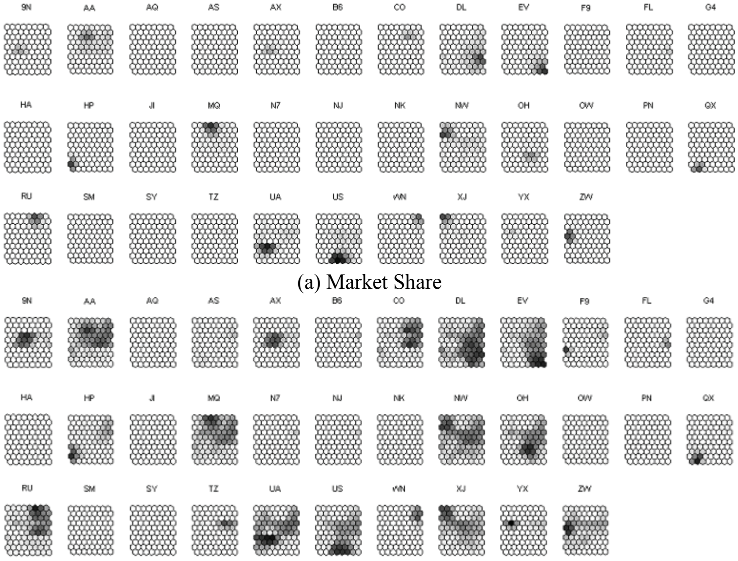
(a) U-Matrix (b) Clusters Identified by K-means

The SOM thus extracts high-level structures marked in the clustering of similar prototype data. The clustered structures located by SOM confirm that US domestic airlines, especially major airlines, tend to serve different markets in order to reduce direct competition with each other, which is widely reported in Taaffe *et al*. (1996). However a more vital question is what factors are attributed to each cluster. This can be answered by visualizing the spread of the values of each component via SOM '*component planes*' (Figure 4). Prototype data have components of the same number to that of input data, each corresponding to an original attribute. By comparing distance-matrices to component planes, we can conclude what component or group of components (airlines) mean most to a certain cluster. For instance, Cluster 2 (Figure 3 and Figure 4a) has high values of US Airways (US) market share. Thus the neurons in Cluster 2 can be interpreted as the markets controlled by US Airways. Following this example, we can then mark each cluster in Figure 3a by certain properties, that is, the components (airlines) that contribute to it more visibly (Table 1).

Component planes can also be used to examine the associations among components (airlines in terms of market share and airfare). At first glance, the overall patterns in Figure 4a (market share) and Figure 4b (airfare) may seem quite different. However, when attention is on the most meaningful information (high values), close correspondence exists. Invariably, neurons with high to moderately high market share also exhibit rather high fares. This means airlines tend to set high airfare in the markets where they dominate, as indicated by Figure 5. This is consistent with many previous studies and shows that the considerable impact that the level of competition has on the practice of setting fare in the US domestic airline industry (Goetz and

Sutton, 1999; Vowles, 2000a; Goetz, 2002). Also indicated in Figure 4, some airlines can still charge high airfare in certain markets even if they account for relatively low market share. For instance, American (AA) has two areas of high airfare as seen in the component planes: one where it has full control; the other where it only has large but not dominant market share. However, in this area Delta (DL) has high market share. Collectively they have the full control of these markets and thus either can charge high airfares (Figure 6). Also notice that Atlantic Southwest (EV) and Air Wisconsin (ZW, a regional subsidiary of Northwest) charge their customers highest airfare since they receive relatively very low competition from other airlines in their respective markets. On the other hand, the markets within Cluster 7-1 have the lowest airfare due to the lack of dominance of any particular airlines. This further proves that a high level of competition usually leads to low pricing.

FIGURE 4
SOM COMPONENT PLANES (2002 DB1BMARKET)



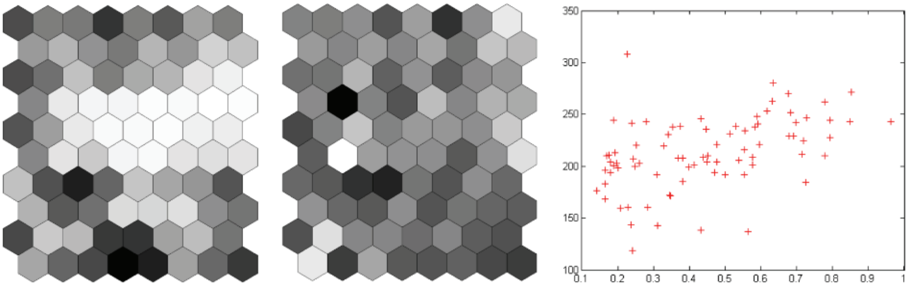
(a) Market Share

(b) Average Airfare

TABLE 2
CLUSTERS BASED ON 2002 MARKET SHARE DATA

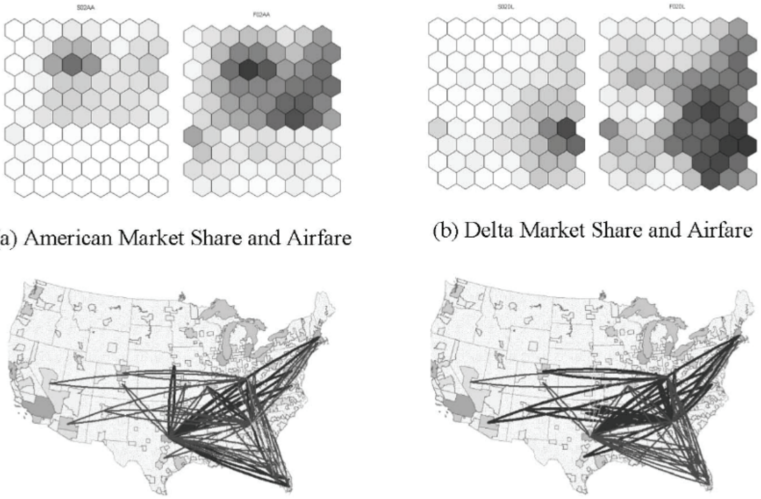
Cluster #	Cluster Property (Airline)
1	Air Wisconsin (ZW)
2	US Airways (US)
3	Northwest (NW), Mesaba (XJ)
4	American (AA), American Eagle (MQ)
5	Continental (CO), Continental Express (RU)
6-1	United (UA)
6-2	American West (HP)
6-3	Horizon (QX)
7-1	No dominant airlines
7-2	Comair (OH)
7-3	Southwest (WN)
8-1	Delta (DL)
8-2	Delta (DL), Atlantic Southeast (EV)

FIGURE 5
MAXIMUM MARKET SHARE AND MAXIMUM AIRFARE



(a) Maximum market share (b) Maximum airfare (c) Maximum airfare against maximum market share

FIGURE 6
MARKETS CONTROLLED BY AMERICAN AND DELTA COLLECTIVELY



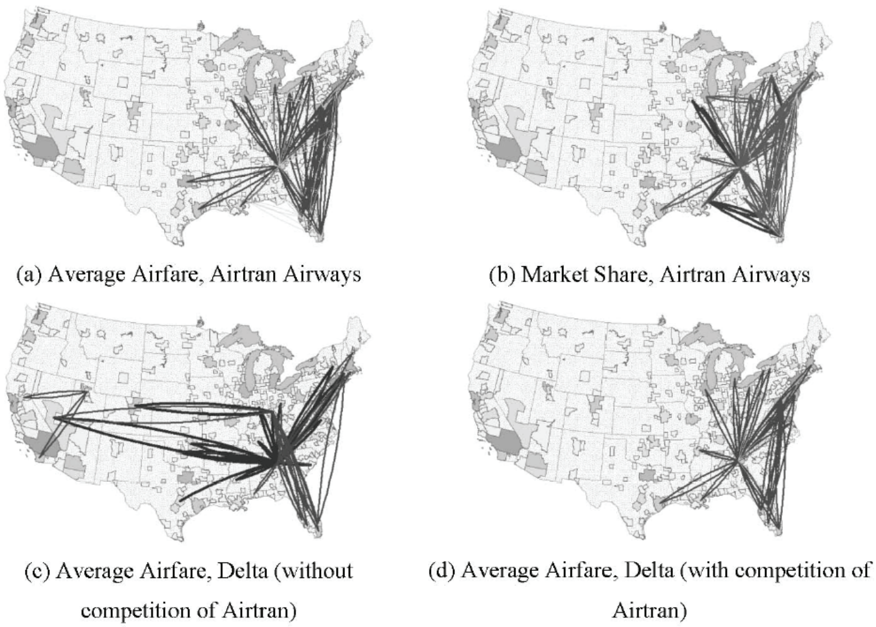
(a) American Market Share and Airfare (b) Delta Market Share and Airfare

(c) American Market (d) Delta Market

Note: (1) Average airfare ($): ― < 100 ― 100 - 200 ― > 200 ; (2) Passengers >= 30,000 and share (AA+DL) >= 80%

Travelers obviously benefit in the markets with the existence of low-fare airlines, *e.g.,* Southwest, JetBlue, and Airtran, even if these low-fare airlines may have relatively low market share. Thanks to the competition of low-fare airlines, even full-service airlines are forced to set lower fare in these markets. The markets projected at the upper right-hand corner of component planes (Cluster 7-3, Figure 3a) are typical examples of this so-called "*Southwest Effect*" (Vowles, 2000b, 2001). This suggests that in some cases the quality of competition matters the most if there are low-fare carriers involved, even though they may only account for a small market share. As shown in Figure 7, among the markets of Delta, those with competition from Airtran are associated with much lower average airfares than those without.

FIGURE 7
GEOGRAPHIC EFFECTS OF LOW-FARE AIRLINES (AIRTRAN AIRWAYS)



(a) Average Airfare, Airtran Airways

(b) Market Share, Airtran Airways

(c) Average Airfare, Delta (without competition of Airtran)

(d) Average Airfare, Delta (with competition of Airtran)

Note: (1) Market share: — < 10% — 10% - 60% — > 60% ; (2) Average airfare ($): — < 100 — 100 - 200 — > 200 ;
(3) Airtran Airways: passengers > 20,000; Delta: passengers > 20,000

5. CONCLUSIONS

Large volumes of spatial interaction data continue to accumulate. When dealing with them, it is often a prerequisite to reduce the data complexity beforehand. The SOM method has the advantage of collapsing origin-destination information and interaction attribute information simultaneously to retrieve essential and other well-defined relationships embedded in the data. The consistent geometric properties of SOM output provide a powerful platform for visual data exploration, validation and evaluation. As illustrated in this paper, native SOM visualization forms can advantageously be integrated in an interactive visualization environment.

Findings in US domestic air travel data suggest that SOM is capable of identifying clustered structures in large spatial interaction data. Via SOM, we are able to obtain an overview of a quite complex data set. The results indicate that airline carriers tend to serve markets in different geographic regions, reducing direct competition. In addition, the results confirm the importance of the competition level in determining the price of air services, that is, in the markets where the level of competition is high airfare tends to stay low. The SOM method is also instrumental in uncovering other interesting but less obvious relationships. For instance, our findings indicate that in a market served by two major airlines, the airfare usually remains at a relatively high level, while if the competition is from a low-fare carrier, the price is often driven down significantly. In sum, all of these suggest that SOM is capable of locating rather localized, focused, or partial structures as well as essential relationships in the entire dataset.

Visualization is commonly realized by using visual variables to describe the different kinds of information in the data. Usually only a limited number of visual variables can be applied to a single visualization; otherwise it will become too complex to comprehend. The idea is, in addition to reducing data complexity, to adopt multiple visualization forms so that the number of visual variables can be multiplied and information in display can be greatly increased. In the VDM environment, a variety of visualizations are implemented and linked together. As a result, various aspects of spatial interaction can be cross-examined intensively in order to fully understand the data. The improved interactivity is useful since the purpose of data exploration is, after all, to suggest hypothesis. For instance, linking SOM feature maps with flow maps offers a way to take a look at how the structures revealed by SOM are geographically defined.

6. REFERENCES

Berry, B.J.L. 1962. *Structural Components of Changing Transportation Flow Networks*. Fort Eustis, VA US Army Transportation Research Command.

Berry, B.J.L. 1966. *Essays on Commodity Flows and the Spatial Structure of the Indian Economy*. Department of Geography, Research Paper No. 111. University of Chicago Press. Chicago.

Black, W. R. 1973. Toward a Factorial Ecology of Flows. *Economic Geography* 49:59-67.

Buckley, A., M. Gahegan, and K. Clarke. 2000. *Geographic Visualization: A UCGIS White Paper on Emergent Research Themes Submitted to UCGIS Research Committee*. http://www.ucgis.org/emerging/Geographicvisualization-edit.pdf. Last accessed on 1 August 2002.

Buttenfield, B., M. Gahegan, H. Miller, and M. Yuan. 2000. *Geospatial Data Mining and Knowledge Discovery: A UCGIS White Paper on Emergent Research Themes Submitted to UCGIS Research Committee*. http://www.ucgis.org/emerging/gkd.pdf. Last accessed on 1 August 2002.

Chicago Area Transportation Study. 1959. *Final Report, Vol. I, Survey Findings*. Chicago, Il, pp. 96-99.

Gould, P. 1981. Letting the Data Speak for Themselves. *Annals of the Association of American Geographers* 71:166-176.

Goetz, A.R. 2002. Deregulation, Competition, and Antitrust Implications in the US Airline Industry. *Journal of Transport Geography* 10:1–19.

Goetz, A.R., and C.J.Sutton. 1997. The Geography of Deregulation in the US Airline Industry. *Annals of the Association of American Geographers* 87:238–263.

General Accounting Office (GAO). 2003. *Code of Federal Regulations. Title 14– Aeronautics and Space, Part 241*. Washington, DC.: Government Printing Office.

Kohonen, T. 2001. *Self-Organizing Maps*. 3rd edition. Berlin, Heideberg: Springer.

Longley, P.A., S. M. Brooks, R. McDonnell, and B. MacMillan. 1998. *Geocomputation: A Primer. Chichester*: John Wiley and Sons.

Marble, D. F., Z. Gou, and L. Liu. 1995. Visualization and Exploratory Data Analysis of Interregional Flows. In: *Proceedings of the 1995 Conference on Geographic*

Information Systems in Transportation (GIS-T), 128-136. Washington D.C.: American Association of State Highway and Transportation Officials (AASHTO).

Marble, D. F., Z. Gou, L. Liu and J. Sauders. 1997. Recent Advances in the Exploratory Analysis of Interregional Flows in Space and Time. In: *Innovations in GIS 4*, eds. Kemp, Z., 75-88. London: Taylor & Francis.

MacEachren, A. 2001. An Evolving Cognitive-Semiotic Approach to Geographic Visualization and Knowledge Construction. *Information Design Journal* 10(10):26-36.

MacEachren, A., M. Wachowicz, D. Haug, R. Edsall, and R. Masters. 1999. Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographic Visualization with Knowledge Discovery in Database Methods. *International Journal of Geographic Information Science* 13(4):311-334.

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

Openshaw, S., and R. Abrahart. 2000. *GeoComputation*. New York: Taylor & Francis.

Taaffe, E.J., H.L. Gauthier, and M.E. O'Kelly. 1996. *Geography of Transportation*. 1st Edition. Upper Saddle River. New Jersey: Prentice Hall.

Tobler, W. 1976. Spatial Interaction Patterns. *Journal of Environmental Systems* 6:271-301.

Tobler, W. 1979. Cellular Geography. In: *Philosophy in Geography*, ed. Gale and Olsson, 379-386. Reidel: Dordrecht.

Tobler, W. 1987. Experiments in Migration Mapping by Computer. *The American Cartographer* 14(2):155-163.

Tobler, W. 1988. The Quadratic Transportation Problem as a Model of Spatial Interaction Patterns. In: *Geographical Systems and Systems of Geography: Essays in Honor of William Warntz*, ed. W. Coffey, 75-88. London: University of Western Ontario.

Ullman, E.L. 1957. *American Commodity Flows: A Geographical Interpretation of Rail and Water Traffic Based on Principles of Spatial Interchange*. Seattle, WA: University of Washington Press.

Vowles, T.M. 2000a. The Effect of Low Fare Air Carriers on Airfares in the US. *Journal of Transport Geography* 8(2):121–128.

Vowles, T.M. 2000b. The Geographic Effects of US Airline Alliances. *Journal of Transport Geography* 8(4):277–284.

Vowles, T.M. 2001. The 'Southwest Effect' in Multi-airport Regions. *Journal of Air Transportation Management* 7(4):251–258.

Yan, J. and J. Thill. 2007a (in press). Visual Exploration of Spatial Interaction Data with Self-Organizing Maps. In: *Self-Organizing Maps: Applications in Geographic Information Science*, ed. P. Agarwal and A. Skupin. London: John Wiley & Sons.

Yan, J. and J. Thill. 2007b. Visual Data Mining in Spatial Interaction Analysis with Self-Organizing Maps. (unpublished ms. In review.)