*International Journal of Geographical Information Science*
Vol. 00, No. 00, March 2007, 1–23

# Highlighting Space-Time Patterns: Effective Visual Encodings for Interactive Decision Making

Mike Sips, Jörn Schneidewind, Daniel A. Keim
*(March 2007)*

The research reported in this paper focuses on integrating analytical and visual methods in order to explore complex patterns in geo-related multivariate data sets and to understand changes of these patterns over time. The goal is to provide techniques that are able to analyze real-world Data Warehouses, a typical architecture to manage such geo-related multidimensional data sets, in order to support analyst's decision making process. Challenges arise because real word applications usually have to deal with millions of records, with dozens of dimensions, and spatio-temporal context. Therefore, a tightly integration of automated analysis and interactive visualizations are integrated is needed (as proposed in the context of Visual Analytics). Our approach uses the well studied capabilities provided by Data Warehouses supporting knowledge discovery and decision making to analyze spatio-temporal behavior of pattern in high dimensional-spaces. The topic of the paper is to show possible interplays between automated analysis and geo-spatial visualization.

## 1 Introduction

Research and development as well as numerous application scenarios produce rapidly growing data volumes with increasing complexity and dynamics. This is possible because of rapid advances in the technology of processors, storage and networks. The data is usually not only characterized by a number of multivariate dimensions, but also by geo-related and temporal attributes. Data Warehouses are a common way to organize and manage such data sets, since they not only provide appropriate ways to model such relational data along the space, time and attribute dimensions, but also provide meaningful levels of abstraction. The analysis of the data contained in Data Warehouse environments sets is an important task, since for decision makers and analysts it is essential to rapidly extract relevant information from this flood of data in order to turn raw data into valuable knowledge (see Andrienko et al. (2003) and Dykes and Mountain (2003)). A key challenge is to gather the potential valuable information from complex data spaces, analyze the data, and

Stanford University, USA, e-mail: msips@graphics.stanford.edu
University of Konstanz, Germany, e-mail: schneide@inf.uni-konstanz.de
University of Konstanz, Germany, e-mail: keim@inf.uni-konstanz.de

present the results in a meaningful and intuitive way that allows the user to quickly identify the important information, and to react on critical process states or alarming incidents with proper actions; within minutes when possible (e.g. network security, fraud detection scenarios and emergency response). Interactive analytical reasoning in complex, high-dimensional, and huge data volumes has become one of the most important research issues in Visual Analytics (see Thomas and Cook (2005) for more details).

Interactive visualizations are of great value in dealing with vast volumes of complex data in order to support analytical reasoning and decision making since presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem solving and gaining deeper domain knowledge. The combination with automated data analysis methods holds enormous potential to provide valuable and previously unknown information, it encourages the data analyst to interact directly with the data and the involved algorithms, solving problems by applying domain expertise and general background knowledge to form and validate new hypotheses, and it allows the data analyst to shift and adjust exploration goals in order to discover alternative options (see Keim et al. (2004) and de Oliveira and Levkowitz (2003) for further details on visual data ming techniques).

The aim of our paper is to investigate an integrated approach combining automated analysis with a suite to easy to understand visual encodings highlighting the spatio-temporal behavior of dynamically changing data stored in Data Warehouses. The goal is to support analytical reasoning and visual assistance to facilitate proper actions. A good example is credit card fraud protection where the geographic information of credit card transactions at certain points in time can help to prevent fraud. Credit card companies may verify customer authorizations for those transactions which show a great difference in their dynamics (in a distance measure) in a very short time span (same credit card number was processed within one hour, from locations that are 1000 miles away), or where the history of the spatio-temporal behavior corresponds to high risk patterns (e.g. many online high amount purchases from a single location, e.g. high risk countries).

The integrated analysis along all dimensions (space, time, multivariate profiles) holds great potential to provide valuable and previously unknown information that can identify complex phenomena, especially multivariate space-time patterns. To support fraud detection in the above scenarios, methods for analyzing space time pattern must be tightly coupled with visual representations, that clearly show potentially fraud patterns diverging form regular transaction patterns. The research challenge is to provide comprehensive visual encodings for multi-dimensional data spaces allowing to identify multivariate geo-patterns, reveal their relationship, and follow their changes in time as well as under-

stand the reason for these changes. Effective visual reasoning is based on the visual understanding of patterns in an environment with multiple dimensions and the projection of their future status. Our approach supports analytical reasoning and decision support by

- showing multivariate pattern at certain time steps and their spatio-temporal dynamics by using clustering and data abstraction techniques
- allowing the data analyst to adjust results interactively by selecting central themes and dimensions
- visualizing the strength with which data items are associated to the clusters in order to allow the data analyst to understand the meaning of the structures
- highlighting temporal behavior in multiple perspectives and levels using coordinated views

To meet these tasks, we focus on the combination of automated data analysis methods and smart visual encodings that follow the Visual Analytics Mantra Keim (2005).

## 2 Highlighting Space-Time Patterns – An interactive approach

Our approach provides a suite of visual encodings combined with Analytical techniques to effectively explore and highlighting space-time patterns in Data Warehouse environments. Data Warehouses are today the typical way to organize and integrate large data volumes from multiple data sources; therefore our visual analysis methods are adapted to this architecture and it's interfaces. Data Warehouses are able to model relational data along space, time and attribute dimensions as well as provide meaningful levels of data abstraction. This is realized by organizing the data based on a multi-dimensional data model, called Data Cubes. Before introducing our technique, we therefore give a brief overview on Data Warehouse concepts that are related to our research (details can be found in Gray et al. (1997)).

### 2.1 *Multi-Dimensional Data Cubes*

A Data Cube is typically defined by dimensions and measures, which are similar to independent and dependent variables in traditional analysis ( Stolte et al. (2002)). For example, in a credit card transaction relation, the customer would be a dimension, while the transaction amount would be a measure. Data Cubes typically contain locations (e.g. country, state, city), time (e.g. year, quarter, month), and a number of attribute dimensions (e.g. products).

Data Cubes can be navigated and queried via Online Analytical Processing (OLAP), which provides an effective way to access the available data by providing drill-down / roll-up functionality. OLAP tools are typically used by analysts to produce standard reports (e.g. to analyze a companies product sales per year or sales per country), but have limited capabilities in detecting more complex space-time patterns (e.g. How did change the product sales per country over the years?).
Standard techniques take typically the time or space dimension into account, when analyzing attribute dimensions (it corresponds to the analysis of a single slice in the Data Cube). Furthermore standard analysis usually focuses on aggregations of attributes (e.g. How was the sales amount for a particular product for a certain regions?) rather than on their multivariate analysis (e.g. Are there clusters of regions that indicate similar product sales patterns?). To reveal such complex space-time patterns, sophisticated analysis methods are needed that take the space, time and attribute dimensions of the Data Cube into account.

### 2.2 *Application Data*

We demonstrate the benefit of our technique by using two data sets in the field of business intelligence. The first one is the InfoVis Contest 2005 data set (Grinstein et al. (2005)), which reflects technological trends in the USA. The data set contains characteristics (e.g. location, product sales, employment count) of technology companies, organized by year from 1989 till 2003. The data cube is modelled along a time (year, month), a space (location of companies: state, city) and attribute dimensions( sales, employees, etc.). The aim is to characterize trends, patterns, or structure in the data across all dimensions. The second data set contains (anonymous) real world product sales data in Italy. Here the Data Cube is defined by time (year, quarter, month), a geographic location (e.g. Sicily) and multivariate attributes reflecting a multidimensional profile including the product type, the amount of sales etc. The aim here is to identify any major sales pattern and its changes, by taking time and space dimensions into account.

### 2.3 *Interactive Exploration of the Data Warehouse*

In our research we focus on the visual analysis of the dynamics of spatio-temporal behavior of patterns extracted from Data Warehouse Environments. Our system supports the visual analysis of space-time pattern by organizing the underlying Data Cube around the temporal and geo-spatial dimensions, that means, the multivariate profile expressing the attribute properties are measured in terms of individual time steps and geo-spatial locations.

**figure 1 should be placed here (top of page)**

For example, for each industry sector the sales, number of employees etc. are measured for every combination of time step and geo-spatial location. These measures allow us to analyze relationships among the multivariate attributes and their spatio-temporal properties. A related data operator from the GIS Community is Map Cubes proposed by Shekhar et al. (1999). Map Cubes combine standard data cubes with galleries of maps (cartographic visualizations of dimensions). Our approach focuses on the analysis of multivariate patterns in the Data Cube directly and then visualizes them with respect to their time and space properties.

In their work Bedard et al. (2003) successfully demonstrate the unique capabilities to explore multi-dimensional data in an interactive and intuitive way offered by the SOLAP architecture in order to support the data analyst in the geographical knowledge discovery (the combination of cartographic and statistical views is called SOLAP).

We use similar interactive drill-down and roll-up capabilities (see figure 1), but in order to provide an effective analysis we organize our highlighting step around a central theme (e.g. in figure 3 the sales dimension has been chosen as the central theme) that controls the analysis and highlighting (e.g. in figure 3 the data analyst is interested in analyzing spatio-temporal behavior of the sales periods and the impact to the number of employees) as well as the visual output (e.g. in figure 3 the clusters are sorted from left to right according the cluster centroids along the sales dimensions while the color show the number of employees).

A related work to our visual output are linked micromap visualizations proposed by Carr and Pierson (1996) . Micromap plots link small generalized maps (exact positions and boundaries are not important) with statistical panels. The primary goal here is to show geo-spatial patterns in statistical summaries. In our work we follow the visual simplification of the maps to show an overview about the spatial distribution of statistical variables. In contrast to linked micromaps we organize the visual output of the second chosen attribute (e.g. the number of employees) around the central theme in our analysis (e.g. the sales amount). The goal of our design is to link geo-spatial patterns to features such as clusters of the high-dimensional data space (e.g. figure 3 shows clearly a linear correlation between the amounts of sales and the number of employees in the resulting clustering). In contrast to linked micromaps, our visual encodings, are tightly integrated into the knowledge discovery supported by the Data Warehouse.

Our system provides several data as well as visual ordering methods, such as sorting and reordering around the central theme. For example, the data analyst can sort the computed clusters, as a result of an multivariate analysis step, according to their compactness.

**figure 2 should be placed here (top of page)**
These reordering supports the understanding of dynamic behavior of companies in their sales dimension. Figure 2 shows the result of such a highlighting step. It is easy to detect some stable sales patterns in Italy over a single year, like a cluster in the south of Italy (orange), that identifies similar sales patterns these regions from February till April.
After an initial representation of these patterns at the highest summarization level (States or Provinces), the user is able to refine the results or to get more details on demand. Our system further supports the interactive exploration of the data cube by allowing the data analyst to select some subsets of the multivariate profile. The interface allows the data analyst to select a certain level of detail in the underlying Data Cube structure via drill down / roll up functionality. For example, when analyzing sales data over certain years, the selection (State, Year, and Sales) would analyze the sales of a product per country over the year. The user could then drill down to (County, Year, Sales), to analyze the sales for all counties of a selected State over the year.

## 3 Highlighting Space-Time Pattern

One important observation is that the analysis of spatio-temporal behavior depends largely on the questions to be answered or the hypothesis to be checked. Since the data analyst selects a subset of dimensions, and a central theme around that the visual analysis should be arranged, one idea to presenting spatio-temporal dynamics of patterns would be to compute and visualize attribute differentials explicitly, using a change map Monmonier (1989). Change maps are a well-known approach that creates multiple sequenced maps, using the same underlying map, to allow an understanding of spatio-temporal behavior by the data analyst. We follow the idea of change maps using sequenced maps since change maps are a an intuitive way to show spatio-temporal behavior. Since change maps are helpful to show individual attribute differentials, they do provide only little insight to complex dynamics of spatio-temporal behavior in Data Warehouses. For that reason we combine this idea with the model (e.g. grouping in clusters) created by automated data analysis to explicitly show coordinated views to stability, uncertainty and changes of patterns in multi-dimensional data spaces.
An approach proposed by Skupin and Hagelman (2003) uses SOM embeddings of multivariate profiles to explicitly visualize changes on the 2-D surface of the SOM. Although SOM-based visualization is very popular in GIS applications they are often hard to interpret without additional analysis functionality, like linked views using Parallel Coordinates, therefore we propose a more explicit analysis of patterns in multi-dimensional data space in order to support inter-

active decision making.

### 3.1 *Multivariate Analysis and Abstraction*

In our approach we use clustering methods to process every multivariate profile at every single time step to group similar profiles into clusters in order to find spatio-temporal movements (see Han and Kamber (2005) for details on clustering algorithms). Note, that we do not use all the data from all time steps as input to the clustering algorithm because this would only group globally similar multivariate profiles together. The resulting clusters would not reflect changes between two time individual steps (analysis of pattern over all time steps).

In a first step we compute an initial clustering for every time step. An important question in analyzing high-dimensional data spaces is how many clusters are in the data space. This is not known a priori and there might be no unique answer. We use cross-validation in our approach to determine the number of clusters. In general, the idea of cross-validation is to divide the overall data into a number of training sets and one test set. Then the same type of analysis is successively applied to the data items belonging to the training samples, and finally, the results of that analysis are applied to the test set to compute a measure of accuracy. We identify the notion of accuracy with that of distance in our cluster analysis using the k-means and EM algorithm. The idea here is to apply the cross validation to a range of cluster numbers, and compare the resulting average distances of the data points to their cluster centers (because the resulting clusters are convex, for further readings see Han and Kamber (2005)).

Figure 4 shows the strength based on the distance to the cluster center of each individual cluster members. Note, in contrast to SOM, each time step may have a different number of clusters.

Figure 3 shows the initial clustering of the InfoVis Contest data set for the year 1992 from the computer hardware industry perspective. The clusters are automatically sorted according to the selected theme (sales) using the sales values of the individual cluster centers. This idea is based on the rank-by-feature framework proposed by Seo and Shneiderman (2004), Seo and Shneiderman (2006). The user can choose a ranking criterion and sort 1/2-D projections according to that criterion. The outcome of the cluster analysis can be easily added into the Data Warehouse. We suggest to use a star schema that link the individual clusters to the selected multivariate profiles. This allows the data analyst to use the described OLAP functionality to explore the clusters. For example, the initial clustering in figure 3 shows no significant difference in the sales attributes of the first three lower sales attribute (see central theme with a boxed display).

**figure 3 should be placed here (top of page)**
It is not clear from this perspective why those clusters are not grouped into a common cluster.
Our system allows the data analyst to explore more details of the clusters along alternative attribute dimensions. Figure 3 shows further the employee distribution of each cluster. The clusters are sorted according to the selected theme (sales) and the sales clusters are colored according to the number of employees. It is clear now, that the number of employees is one of the splitting criteria within the grouping process of the clustering algorithm. The data analyst can further explore the clusters at different aggregation and summarization levels (see also multi-scale visualization in section 3.6).

### 3.2 *Visualizing Temporal Dynamics*

The objective of this research is to develop new methods and techniques to discover the dynamics of spatio-temporal behavior of patterns. The complexity of the visual analysis boosts with the time dimension, and our approach aims to combine visualizations with automatic analysis methods in a fruitful way.
We highlight the spatio-temporal information by showing the contribution of each cluster at time step $t_i$ to all clusters at $t_{i+1}$ (dynamics in the cluster membership) as well as the overall trend of that contribution in the attribute space (dynamics in the attribute space). The strength of which each region is associated to a cluster allows us to identify the core object of the clusters. Core objects are data points in the multi-dimensional data space that are around the centroids local neighborhood. The changes of core objects shift the meaning of a cluster and we show these changes by using different color intensities. Clusters that do not change in time (stable) are visualized using the same color as in the previous clustering. The maps for each time step emphasize the stability of clusters and allow the identification important periods.

### 3.3 *Strength and Uncertainty*

Starting with a given multivariate profile, many different classifiers can be learned depending on the choice of the used procedure or the parameter settings. That means the data analyst must be able to evaluate the different clusters in order to associate a meaning with them. It is useful to analyze the distance to the cluster center as an important characteristic of the resulting clustering (see also cross-validation in section 3.1).
Let $C = \{c_1, \ldots, c_n\}$ be a multivariate cluster in the attribute space. Let $r_i(c_i)$ be the geo-spatial location associated with that cluster member.

**figure 4 should be placed here**
The cluster centroid $Centr(C)$ is defined as the average over all cluster members in the attribute space with

$$Centr(C) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |c_i, c_j|$$

The strength for each region $Str(r_i)$ is based on the distance to the Cluster's Centroid and is defined as follows:

$$Str(r_i) = \frac{1}{|c_i, Centr(C)|}$$

That means, data points in the centroids local neighborhood are very strongly associated to the clusters than distant points. The strength describes the contribution of a cluster member to the cluster since it is more likely that distant points change their cluster memberships in the next time step. Therefore, we call very strong associated data points the core objects of the cluster $C$. The membership change of core objects to a different cluster clearly shifts the meaning of the cluster and it is of high importance for decision makers. In application scenarios that involve a large number of time steps, core objects are very effective for the visual analysis of spatio-temporal behavior. Thus our highlighting approach focuses in such scenarios on emphasizing the visual encoding of the core objects.
Figure 4 shows the strength in the initial clustering for each individual region for the year 1992 from the computer hardware industry perspective. The clusters are sorted according to their centroids sales coordinates (left upper corner to right lower corner). We can observe that some clusters are more compact than others. For example, the clusters with the smallest sales amount are very compact (8 core objects and 3 distant points). In contrast to the compact clusters 1 and 3, The analyst may now select a proper subset of the dimensions to refine the clustering, e.g. ti create more meaningful cluster results. Note, cluster 5 and 7 are special cases because they have just 1 or 2 cluster members.

### 3.4 *Dynamics in Cluster Membership*

A very important issue in understanding temporal dynamics are the changes in the cluster memberships. To understand spatio-temporal patterns, we are interested in multivariate phenomena that are defined over a time period $t_i, \ldots, t_j$ with some minor changes in their core objects.

**figure 5 should be placed here**
We can describe this task more formally.
Let $C_{t_i}$ be a multivariate cluster in the attribute space. We show all contributions of the cluster $C_{t_i}$ to the clusters at time periods $t_i, \ldots, t_j$ with $C_{t_i} \cap \ldots \cap C_{t_j} \neq \emptyset$. Figure 5 shows the contribution to the clustering for year 1993 starting from the initial clustering at 1992 (from the computer hardware industry perspective). The clusters are sorted according their centroids sales coordinates based on the clustering at 1993. In order to improve the understanding of spatio-temporal properties we use the strengths of the data objects to weight their importance along the changes.

### 3.5 *Dynamics in Attribute Space*

The second important issue in understanding dynamics in spatio-temporal behavior are the changes in the attribute space. The visual analysis of the dynamics in cluster memberships alone is not sufficient for the understanding of spatio-temporal patterns. The tight integration with the dynamics in the attribute space provides the meaningful views to spatio-temporal behavior.
**figure 6 should be placed here (top of next page)**
The data analyst is not interested in single informations at certain time steps, he/she is more interested to understand the general direction in which the cluster members are moving. The movement in our highlighting approach is shown by trend maps. Our trend maps show the movements in magnitudes of the change in sorting of the model (e.g. clusters). The change in the sorting describes the global trend of cluster members in terms of raising or dropping according the selected sorting criterion.
Let $C^n_{t_i}$ be multivariate clusters in the attribute space at time step $t_i$. Let $n$ be it's position in the central theme arrangement at $t_i$ (e.g. in the sorting according to sales dimension). The temporal behavior of each cluster member $c_{t_i}$ of the cluster $C_{t_i}$ between the time steps $t_i$ and $t_{i+1}$ can be described using a individual mapping function $f_k(c_{t_i})$. Let $C^m_{t_{i+1}}$ be multivariate clusters in the attribute space at time step $t_{i+1}$ and $f_k(c_{t_i}) \in C^m_{t_{i+1}}$. Let $m$ be it's position in the central theme arrangement at $t_{i+1}$. We define the trend by the change $\delta(C^n_{t_i}, C^m_{t_{i+1}})$ in their arrangement from $t_i$ to $t_{i+1}$

$$trend(c_{t_i}, f_k(c_{t_i}) = \delta(C^n_{t_i}, C^m_{t_{i+1}}) = n - m$$

Figure 6 shows the effect of the cluster membership movements in the attribute space (from the initial clustering perspective). One can easily see that the global trend shows rising or even constant sales trends. One can see,

in 1993 the USA came out of a recession, most, but not all places did better than in 1992.
**figure 7 should be placed here (on top of this page in the final manuscript)**

### 3.6 *Multi-Scale Analysis*

Since Data Cubes represent hierarchical aggregations of the underlying data, it is necessary to take these hierarchies into account, when analyzing Data Warehouses. Our approach allows an automated as well as an interactive drill down to investigate details on certain aggregation levels.
Let $O = \{o_1, \ldots, o_n\}$ be a set of data points. Let $r$ be the radius of the query ball $Q$. The radius $r$ is given by the user and can be adjusted to different application scenarios. We define $Q(o_i, r) = \{o : |o - o_i| \leq r\}$ where $o$ is the dynamic query point in the attribute space.
The analysis and highlighting algorithm goes recursively into the next lower geographic scale, in which the data points $o_i$ are embedded (this corresponds to a SQL-Query to the Data Warehouse).

## 4 Application Scenario – Spread of Product Sales Patterns

This section presents some interesting findings of our space-time highlighting approach applied to the Italian sales data warehouse (product sales data). The goal was to identify any obvious sales patterns over the time and space dimension, e.g. are there regions that show similar sales characteristics and how do these characteristics change over the year. Furthermore it is important to analyze, why regions show similar characteristics.
**figure 8 should be placed here (on top of this page in the final manuscript)**
Therefore we took the product and product sales dimensions in our analysis into account, and did analyze their space and time properties using the proposed methods. Figure 8(a) shows the result of the initial analysis. The discovering of the global spread of two major sales patterns is easy. The sales pattern of the first product type (*red cluster*) starts in January and ends in April and is located in the south of Italy (Sicilia, Campania, Lazio). It shifts to the northern regions Lombardi and Emilia-Romagna in the last three months of the year.
We can observe that the *red cluster* has a strong sales period in the first three months of the year in entire Italy. After that first months we can observe that there is no domestic market anymore for this product type. The same can be

observed for the (*blue*) and (*violet*) product types. Furthermore, we can observe two small interesting sales patterns. The (*orange*) cluster is only active in the north of Italy from May till July. We know that these months are the major tourist season in the north (alpine hiking season).
We can identify two important sales periods of this product type and it may allow a data analyst to adjust sales policies. We can see that the *green cluster* sales pattern (second product type) has a vice versa behavior. To identify the core objects of the relevant cluster, the algorithm can now perform the analysis on the next lower geographical level, and identifies Italian provinces (see figure 8(b)) that are the core objects of each cluster. An analyst may use this knowledge for an adapted sales promotion or refinements of the product lines in certain regions.

## 5 Related Work

Exploring and analyzing large spatio-temporal data sets is a challenging task because of the data complexity and scalability issues. A number of sophisticated techniques have been proposed visualizing geo-related data in an appropriate way. Most of these techniques do, however, focus on the visualization of few single statistical parameters or single attributes over a geo-spatial context, and therefore do not take the multivariate, space and time dimensions of the data into account.
The Polaris system (Stolte et al. (2002)) for example, was proposed to analyze multivariate data sets over geo-spatial context. It supports the analysis of data stored in Data Cubes and provides different visual metaphors to explore the data from different viewpoints. Other successful systems, like the CommonGis tool by Andrienko et al. (2000) employ multiple linked views to analyze the temporal, spatial and multivariate aspects of the data for effective spatio-temporal reasoning.
Recently an interesting approach was proposed by Chen et al. (2006), Guo et al. (2006). The authors introduced an inquiry system for exploring space-time patterns. They construct overviews of the data by using both computational (self-organizing maps) and visual methods (reorder able matrix and map matrix), and present the results of the analysis using multiple linked views. Some efforts have been made in visually mining spatio-temporal patterns, with focus on spatial distribution of temporal behavior Andrienko and Andrienko (2005). With the increasing amount of information that need to be handled, it will become more and more important in the future result to be supported by automated methods when analyzing large data sets. Thus successful analysis of multivariate space-time patterns data requires the tight integration of the user into the exploration process. In Seo and Shneiderman (2004), Seo

and Shneiderman (2006) the authors proposed a framework to enable interactively mining for multivariate patterns by ranking 1-D/2-D projections of high-dimensional data spaces using features of the projections as ranking criterions. Another novel approach to discover high-dimensional data spaces was proposed by Wilkinson et al. (2005). The idea is to use graph theoretic measures on 2-D projections to understand properties of the data.

## 6 Conclusions

This research studies how to adequately analyze large multivariate data sets with a space and time dimension, such as sales data, to support interactive decision making. The newly available complex, high-dimensional data sets pose a challenge on Visual Analytics techniques that integrate automated methods and interactive visualization, in order to face the demand of analyzing these space-time patterns in today's applications.
We proposed a method that is able to detect patterns in multidimensional data sets. Furthermore our method is able to analyze the changes of these patterns over time. Our techniques follow the Visual Analytics Mantra in terms of combining automated and interactive exploration methods. We showed that our approach works on real world business data. An important issue in our future work is to determine the probability of events by using a model (e.g. cluster) and their projection to their near future (try to detect annual sales periods).

## Acknowledgement

The authors thank the anonymous referees for their comments toward the improvement of this report. The authors further thanks Jason Dykes and John P. Lewis for lots of fruitful discussions that greatly improved our research. Special thanks to Andre Skupin for the great support and helpful comments at VASDS 2006 Workshop. Thanks to Stefano Rizzi for providing Italy's product data.
This work was supported by the Max Planck Center for Visual Computing and Communication.

## REFERENCES

Andrienko, G. L. and Andrienko, N. V.: 2005, Visual exploration of the spatial distribution of temporal behaviors., *9th International Conference on*

*Information Visualisation, IV 2005, London, UK*, pp. 799–806.

Andrienko, G. L., Andrienko, N. V. and Gatalsky, P.: 2000, Towards exploratory visualization of spatial temporal data., *3rd AGILE Conference on Geographic Information Science, Helsinki, FI*.

Andrienko, N. V., Andrienko, G. L. and Gatalsky, P.: 2003, Exploratory spatio-temporal visualization: an analytical review., *Journal of Visual Languages and Computing* **14**(6), 503–541.

Bedard, Y., Gosselin, P., Rivest, S., Proulx, M. J., Nadeau, M., Lebel, G. and Gagnon, M. F.: 2003, Integrating gis components with knowledge discovery technology for environmental health decision support, *International Journal of Medical Informatics* **70**(1), 79–94.

Carr, D. B. and Pierson, S. M.: 1996, Emphasizing statistical summaries and showing spatial context with micromaps, *The Computing and Graphics Newsletter* **7**(3), 16–23.

Chen, J., MacEachren, A. M. and Guo, D.: 2006, Visual inquiry toolkit - an integrated approach for exploring and interpreting space-time, multivariate patterns, *AutoCarto 2006, Vancouver, WA, 2006*.

de Oliveira, M. C. F. and Levkowitz, H.: 2003, From visual data exploration to visual data mining: A survey, *IEEE Transactions on Visualization and Computer Graphics* **9**(3), 378–394.

Dykes, J. A. and Mountain, D. M.: 2003, Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications, *Computational Statistics & Data Analysis* **43**(4), 581–603.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H.: 1997, Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals, *Journal Data Mining and Knowledge Discovery* **1**(1), 29–53.

Grinstein, G., Cvek, U., Derthick, M. and Trutschl, M.: 2005, Technology trends in the united states.

Guo, D., Chen, J., MacEachren, A. M. and Liao, K.: 2006, A visualization system for space-time and multivariate patterns (vis-stamp), *IEEE Transactions on Visualization and Computer Graphics* **12**(6), 1461–1474.

Han, J. and Kamber, M.: 2005, *Data Mining: Concepts and Techniques*, second edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Keim, D. A.: 2005, Scaling visual analytics to very large data sets, workshop on visual analytics, darmstadt, germany, 2005. `http://infovis.uni-konstanz.de/events/VisAnalyticsWs05/pdf/03DanielKeim.pdf`.

Keim, D. A., Sips, M. and Ankerst, M.: 2004, Visual data-mining techniques, *in* C. Johnson and C. Hansen (eds), *Book Chapter in: Visualization Handbook*, Elsevier Science Publishing, pp. 813–825.

Monmonier, M.: 1989, Graphic scripts for the sequenced visualization of geo-

graphic data, *GIS/LIS'89*.

Seo, J. and Shneiderman, B.: 2004, A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections, *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, IEEE Computer Society, Washington, DC, USA, pp. 65–72.

Seo, J. and Shneiderman, B.: 2006, Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework, *IEEE Transactions on Visualization and Computer Graphics* **12**(3), 311–322.

Shekhar, S., Lu, C., Tan, X., Chawla, S. and Vatsavai, R.: 1999, Mapcubes: A visualization tool for spatial data warehouses.

Skupin, A. and Hagelman, R.: 2003, Attribute space visualization of demographic change, *GIS '03: Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, ACM Press, New York, NY, USA, pp. 56–62.

Stolte, C., Tang, D. and Hanrahan, P.: 2002, Polaris: A system for query, analysis, and visualization of multidimensional relational databases, *IEEE Transactions on Visualization and Computer Graphics* **8**(1), 52–65.

Thomas, J. J. and Cook, K. A. (eds): 2005, *Illuminating the path: Research and Development agenda for Visual Analytics*, IEEE Computer Society.

Wilkinson, L., Anand, A. and Grossman, R.: 2005, Graph-theoretic scagnostics, *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, IEEE Computer Society.

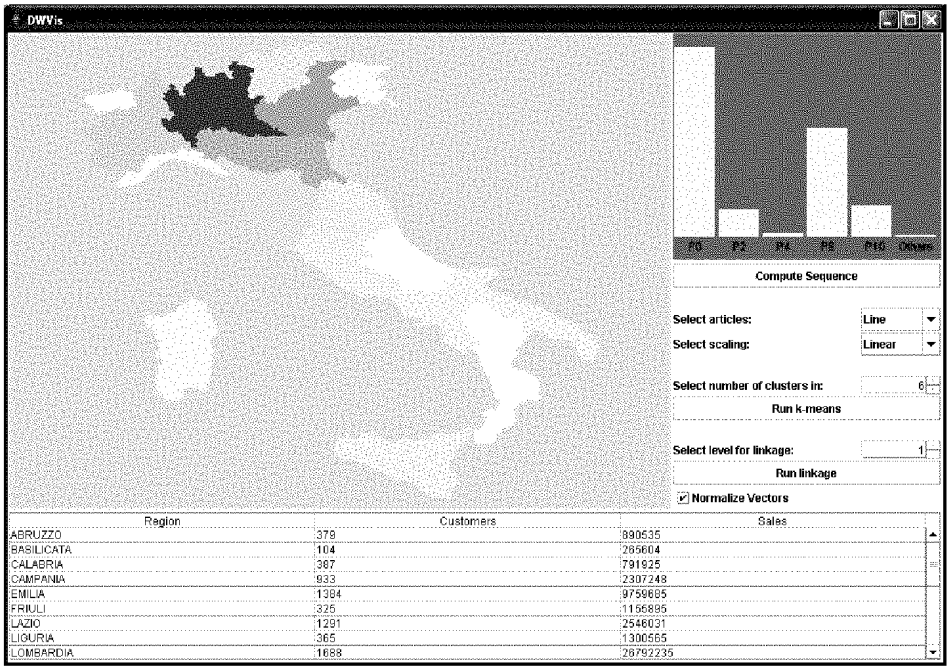



Figure 1. *Warehouse Interface* – Visual Exploration of the data cube through interactive selections of interesting subsets. The interface allows the user to select certain geographic levels in the underlying data cube structure via drill down/roll up functionality

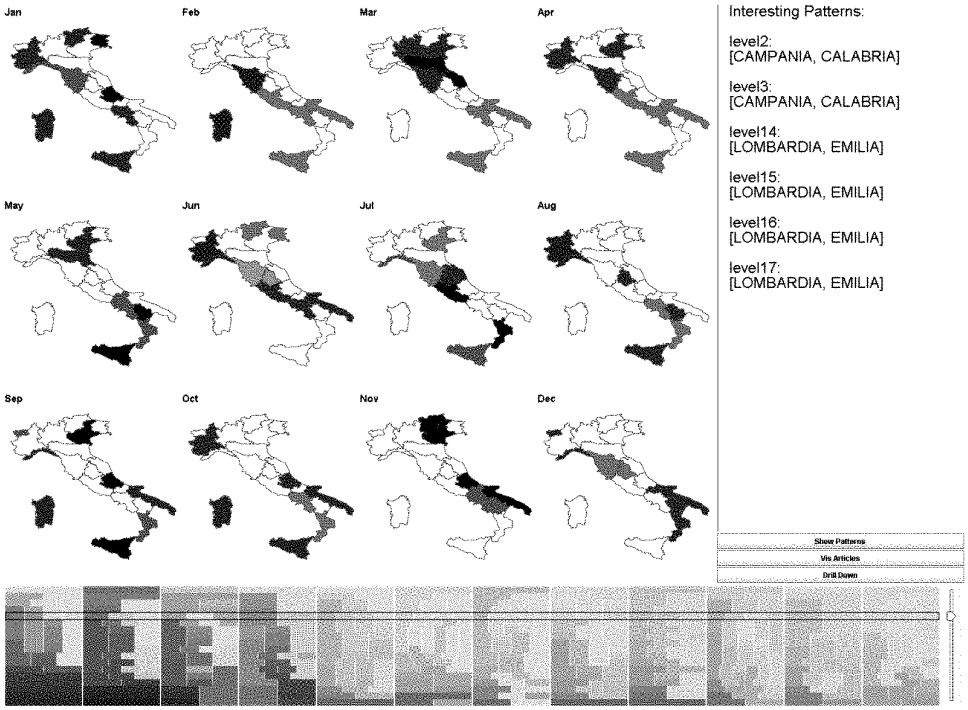



Figure 2. *Analysis Interface* – Highlighting the temporal dynamics of geo-spatial pattern. For example, the regions Sicilia, Puglia, Campania and Lazio are grouped together in a common cluster *(orange cluster)* that is defined from February until April (some little changes in March). The cluster can be seen as a stable sales periods.

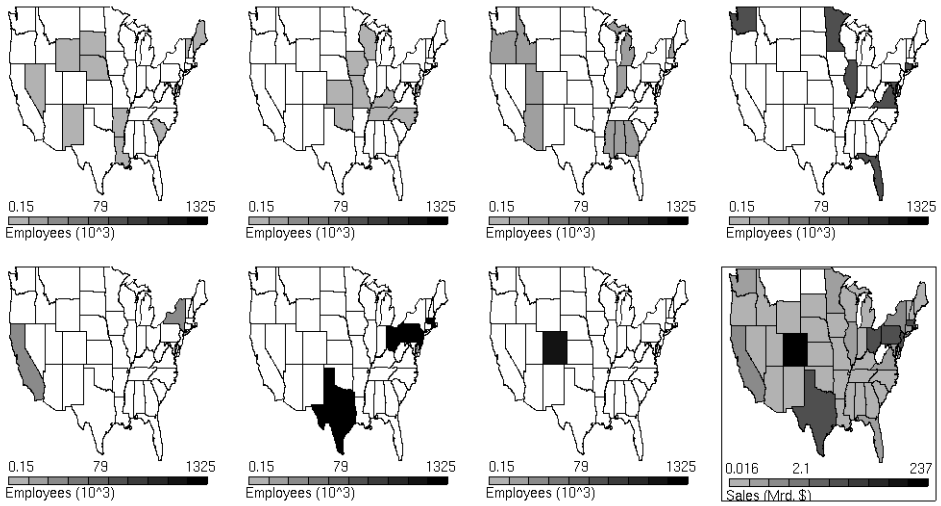



Figure 3. **Multivariate Clustering and Abstraction** – the initial clustering shows diverse sales levels (central theme with boxed visualization) and for each sales level the number of employees of the computer hardware industry in 1992. One can easily see three low sales clusters that show no significant differences in the sales records, but with differences in their number of employees. Further one can identify Colorado with the highest sales of computer hardware.

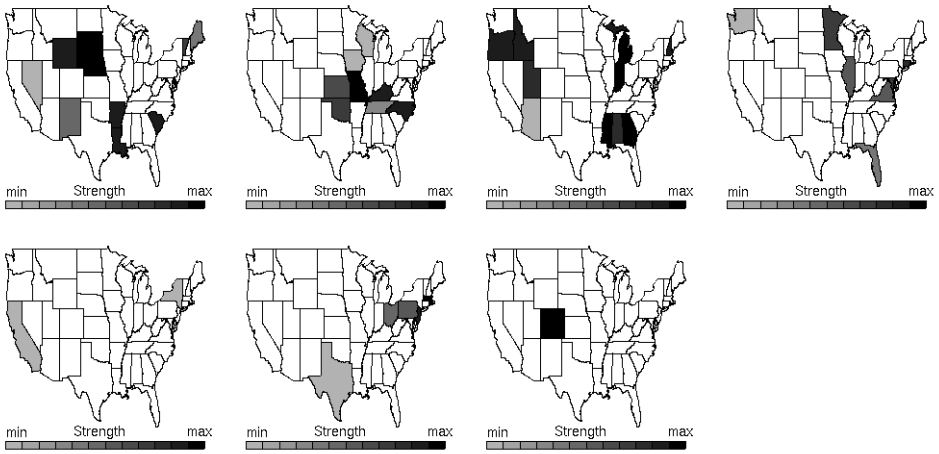



Figure 4. **Pattern and Uncertainty** – our strength is a measure describing the contribution of a cluster member to the cluster. Data points can be core objects of a cluster (high strength) or distant objects. The visual analysis of the core objects supports the data analyst to understand the structure and meaning of the cluster. The figure shows the strength for each region in the initial clustering at 1992 from the computer hardware industry perspective. The visual output is organized around the selected central theme sales level, which means, the maps are sorted from left to right according to their centroids sales amount. We pick a individual color for each cluster because the maps can be merged together in further analysis steps.

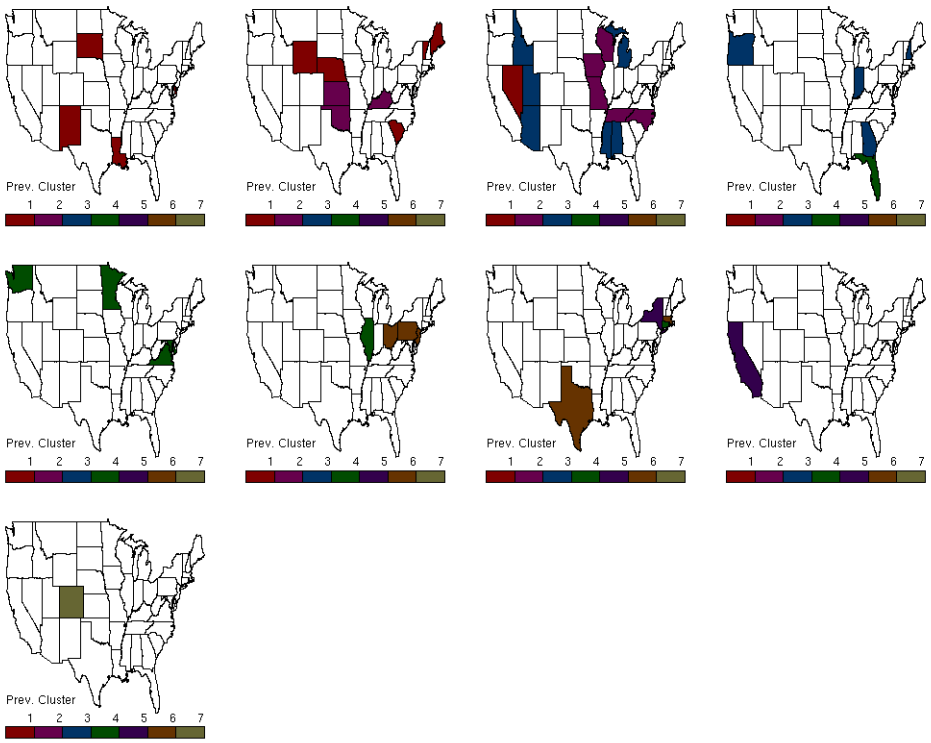



Figure 5. **Dynamics in Cluster Membership** – shows the contribution from the initial clustering for year 1992 to the clustering for 1993. It is easy to see that second and third low sales clusters (previous cluster 1 and 2) are coming together in a common cluster (cluster 3); and additionally one cluster member from the lowest sales cluster. Note, in larger scenarios the highlighting focuses on the core objects of the cluster. The visual output is organized around the selected central theme sales level, which means, the maps are sorted from left to right according to their centroids sales amount. We pick a individual color for each cluster because the maps can be merged together in further analysis steps.

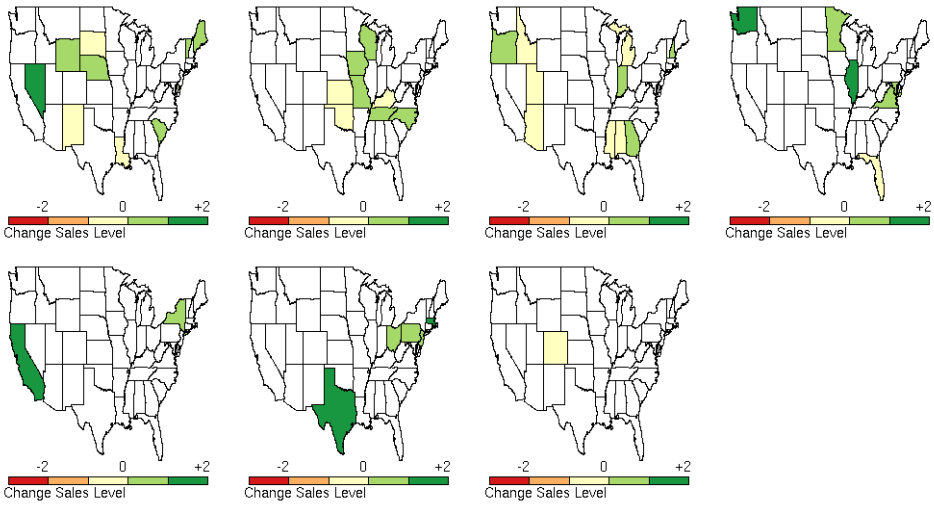
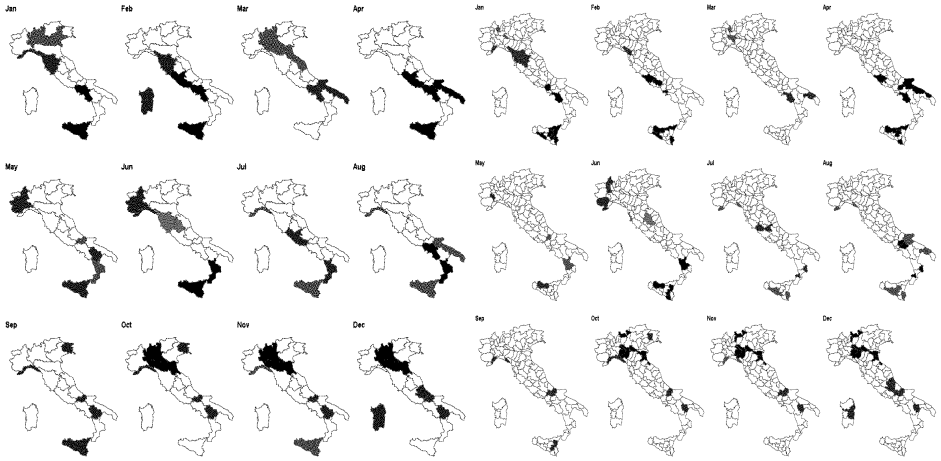



Figure 6. **Dynamics in Attribute Space** – shows the effect of the cluster membership movements in the attribute space. Each region of the initial clustering at 1992 shows the trend. The trend is a measure to show the nature of the cluster membership movements (e.g. to higher or lower sales clusters). One can see that no region drops down to a lower sales cluster. The visual output is organized around the selected central theme sales level, which means, the maps are sorted from left to right according to their centroids sales amount.

Figure 7. **Enhance interesting clusters automatically** – patterns in neighboring regions around the query ball in the multi-dimensional data space are automatically shown in the next finer geographical scale (from state into county level) to support the understanding of clusters along different hierarchy levels.




(a) Highlighting based on regions (Overview) (b) Drill down into provinces (Detail)

Figure 8. **Spread of Product Sales Pattern** – Sales pattern of the first product type starts (*red cluster*) in January and ends in April and is located in the south of Italy (Sicilia, Campania, Lazio). It shifts to the north regions Lombardi and Emilia-Romagna in the last three months.