

Exploratory Data Visualization for Reliable Integration

Eva-Maria Stephan

University of Zurich, 8057 Zurich, Switzerland

stephan@gis.geogr.unizh.ch

Abstract

Integrating different kinds of data is a typical feature and basic requirement of Geographic Information Systems (GIS). In particular, the modeling of environmental data and processes with GIS requires integration of many heterogeneous data sets. Required compatibility for integrating such data is often not given a priori. Data integration therefore often requires preprocessing of data, such as up- and downscaling, or interpolation, to establish compatibility. Such preprocessing, i.e., transforming of data sets may introduce error and uncertainty to the data. This paper focuses on visual exploration of heterogeneous, incompatible data for integration in order to arrive at greater reliability. Interactive investigation and visualization techniques for multiple data sets in a user-friendly environment are discussed.

1 Data Integration and Integrated Analysis

1.1 Introduction

Integrating different kinds of data is a typical feature and basic requirement of Geographic Information Systems (GIS). In particular, the modelling of environmental data and processes with GIS requires integration of many heterogeneous data sets. Driven by the quest for an increased understanding of the working and interrelationships of ecosystems, for example under the conditions of global climate change, combination of data from various sources is called for to derive at new data, e.g., data that is hard or expensive to measure. Through the combination of multiple data sets, *integrated analysis*, e.g., based on some deterministic or probabilistic model (Isaaks & Srivastava, 1989), new information (data) about other phenomena can be modelled. Integrated analysis based on multiple data sets allows one, for example, to calculate measured data of originally sparse resolution, into higher spatial or temporal resolutions.

1.2 Heterogeneity of Data

While GIS are suitable tools for bringing together data from various sources and different thematic fields they are usually devoid of the intelligence for *how* these complex (multidimensional and heterogeneous) data should be properly combined (Burrough & Hazelhoff, 1994). Heterogeneity of data sets involved in an integrated analysis need to be resolved beforehand. Those and problems of incompatibility arise

from different characteristics and data quality of the data sets involved, i.e., different spatial and temporal references, various preprocessing of the data, validity regarding time, different classification of values, etc. Integrated analysis hence often requires transformation, a *pre-modelling* of data, which requires knowledge about the different data set's characteristics.

1.3 Virtual Data Set Approach for Integration

The Virtual Data Set approach, as presented by Stephan et al. (Stephan *et al.*, 1993), offers a flexible framework for data integration. Its basic idea follows an object oriented concept which guarantees a high flexibility and reliability for data integration. The concept proposes an enhancement of original data sets to so-called *Virtual Data Sets (VDS)* with persistently stored methods for prediction and data quality assessment. Those methods can be applied to transform a data set's spatial or temporal resolution, or attributes, i.e., to derive or predict new data values. A VDS thus contains in addition to the original data values, information that is not persistently stored, so-called *virtual data*. This data shall be derived at any time, e.g., to meet the requirements of a particular integrated analysis. Accordingly, the data set is able to adapt itself easily and flexibly to any requirements asked by the user, by performing the prediction methods which are stored jointly with the original data set.

The Virtual Data Set concept relies heavily on a sophisticated enhancement of the data set, i.e., on selecting and associating a data set with suitable methods for prediction and data quality assessment (see also (Bucher *et al.*, 1994)). One approach to determine the appropriate methods for the enhancement of data sets is supported by *visual data exploration*.

1.4 Data Enhancement through Visualization

The VDS concept requires methods of data visualization to gain rapid insight into the characteristics of spatial data, prevent false interpretation, and correlate data in order to promote their correct integration. Interactive visualization techniques that support simultaneous viewing of multiple data sets, correlations, trends, and outliers, allow visual exploration of data prior to their use and lead to new insights which may avoid wrong crucial decisions. For example, there are situations where false interpretation of data leads to false decision making and model-building, e.g., false assumptions about stationarity. The following paper will concentrate on visual data exploration for the enhancement, e.g., for reliable integration of data sets.

2 Visualization

2.1 Data Visualization

As mentioned above, natural science researchers are often confronted with large and heterogeneous collections of data. In order to analyze and integrate these, visual presentation of the data can be used for giving an overview, validating and understanding complex data sets. Combining the ability to both, visualize and interact with a spatial data set allows analysts to see complex patterns and trends associated with the given phenomenon. Data visualization can be applied for two main purposes:

- First, visualizing data can yield additional knowledge about the data's characteristics and its underlying natural phenomenon. Evaluating the relevant data characteristics and their quality prior to their modelling leads to more reliable evaluation of an appropriate model. The additional insight can be used in an exploratory way to improve and validate accurate modelling of data and natural processes.
- Second, graphical depiction of data and its quality is an immediate way to communicate the uncertainty. It can communicate errors inherent to data and propagated in subsequent processing, e.g., integrated analyses.

2.2 Explorative Data Analysis

Data often are collected as observations of the real world rather than as the result of experiment. In the absence of experiment, appropriate probability models against which to test and evaluate the data are not well defined (Haining, 1990). Before any formal model can be set up, e.g., before integrated analysis can be carried out, it is useful to examine data for patterns, outliers, symmetry, homogeneity and distributional properties. In this sense *exploratory data analysis (EDA)* (Tukey, 1977) has attracted attention particularly for analysis of spatial data (Hasslett *et al.*, 1990). Graphical tools can serve to initially explore the data through hypothesis generation, modelling and analysis of residuals (Tukey, 1977). For spatial data, graphical methods of data portrayal are enhanced by recent advances in computer graphics systems, especially the facility to link windows dynamically and permitting several views of the data to be active at the same time (Hasslett *et al.*, 1990), (Aspinall, 1993).

2.3 GIS and ViSC

Although cartography has a long tradition in graphical depiction and analysis of spatial data (Bertin, 1967), (Tufte, 1983), GIS and computing environments offer a completely new working environment for geographical research. Emerging trends from computer graphics and scientific visualization have led to a whole new research topic of *Visualization in Scientific Computing (ViSC)*. These trends try to bring the measured and numerically modelled data to our eyes through the mediation of computer graphics and through use of the human visual sensory system and its ability to steer and interpret complex processes, to analyze trends and extract patterns in graphs and images. Actually, the goals are similar to the cartographic goals, sometimes cartography is referred to as the root of ViSC.

Unfortunately today's GIS visualization techniques are often designated for presentation purposes and designed to produce maps as output, rather than to provide functionality to interactive data visualization. Dynamic graphical data representations are mostly shortcoming and often require bigger efforts programming efforts. Data structures were primarily aimed at recreating the digitized map and could not respond to questions about the elements of the database (Wood & Broodlie, 1994).

Usually GIS offer visualization techniques for single scalar functions in 2 dimensional (sometimes 2 1/2 dimensional) space under stationary or time-dependent conditions. But the real world isn't that simple, and neither are the simulations that attempt to model it. Exploring the complexity and interactions of natural systems more appropriately can be achieved by combining an interplay of multidimensional visualization techniques. Multiple levels of presentation could facilitate the study of

large data sets by providing, first, a quick look and then, if desired, a more detailed look with related position and scalars to compare their characteristics and quality in spatial, temporal and attributional dimensions.

With respect to providing a easy to access and environment for explorative data analysis, the following aspects of GIS visualization functionality need to be further improved:

- better user interface design with graphical interactive access to data and functionality,
- graphical programming and application building,
- 3-dimensional viewing of data coupled with the ability to interactively manipulate viewing angle and display parameters,
- providing visual techniques for presentation and interactive manipulation of data in *realistic* views, i.e. rendered images and complex scenes of the environment,
- providing techniques for continuous and contiguous data, such as surface rendering and animation,
- presenting data at multiple resolutions and allowing real-time up- and down-scaling,
- visualizing data quality,
- combined viewing of multiple data layers,
- linking surface and map-based viewing and statistical plots.

The next section will present a new concept of data visualization and example from a visualization tool, which was developed as an interactive working environment for EDA and user-friendly data viewing of environmental data. Its development was oriented along the above listed items, and aims at improving GIS' data viewing capabilities.

3 Interactive Visual Data Exploration

3.1 Data Scaping

*Data Scaping*¹ is a new concept for *interactive visual data exploration* of environmental data. It is based on the combination of scientific and cartographic presentations of data. The principal idea is to exploit possibilities offered by interactive graphic systems to support the search for patterns and peculiarities in environmental data. It is designed to support integrated analysis in a GIS computing environment.

This section will outline the requirements and specifications of this new visualization concept and present some examples from a first implementation of a Data Scaping-tool.

¹The term was derived from *landscape viewing*. But rather than photo realistic images of landscapes, visualization techniques of Data Scaping present *data values* in rendered scenes and graphical plots.

3.2 Two Categories of Visualization

Two different categories of data visualization techniques help to gain insight into data and to explore data for environmental modelling.

First, exploratory data analysis, based on (geo-)statistic plots shall yield at sampled data's characteristics, and thus increase the ability to reconstruct the spatio-temporal behavior of the phenomenon and consequently estimate attributes at un-sampled locations. I.e., exploratory data analysis supports and improves the interpolation of data for applications where continuous and contiguous data are lacking.

Second, once a data set represents continuous information, i.e., is interpolated on a regular grid or on a contiguous time-scale, then different techniques of *surface visualization and animation* of the data set can demonstrate the spatio-temporal behavior of the real-world phenomenon. Surface rendering, for example, shows the data in such a way, that they corresponds very closely to their real-world appearance, and thus allow for better understanding of its real world behavior. If multiple surfaces are stacked, each representing a different attribute layer, correlation and interdependencies of different attributes and model parameters can be explored.

These two categories of visualization form the bases for exploring the characteristics implicit to data in images. One can either represent the *numeric values* of measured point data using techniques of EDA, or visualize continuous data in their very natural appearance (scaling real world dimensions, space and time on regular grids), by using so-called *realstic* techniques. The level of abstractness (MacEacheren *et al.*, 1992) is very low when data is presented in an image or photo realistic scene, and it is very high, when presented in graphics symbolic form. Representing the two extreme points on an abstractness scale, the combination of both approaches helps to exploit the full range of data presentation.

3.3 Requirements to Data Scaping

The requirements of interactive data visualization are closely tied to explore the characteristics of environmental data. They are in particular:

Real time interaction: a graphical user interface which allows interactive selection and manipulation of

- display parameters, i.e., axes, colors, height, shading, surface texture, legend, etc.,
- viewing angle and object position,
- objects and attributes to be displayed
- resolution and version of an object.

Interactive probes: interactive querying of data values by selecting points on a map or surface with the mouse indicate the numerical values which are color or symbol coded.

2 dimensional views of data: contour plots, images, graphical plots, histograms, scatterplots, boxplots, scattermatrices.

3 dimensional views of data: representing gridded data as color surfaces with different display options (color, shading, transparency), as wire frames, or as combination of both; for surface views the x and the y dimension are always fixed by the data's geographical reference; the z dimension however, also can be given by other attributes than elevation; digital elevation models usually determine the basic shape (x, y and z dimension) for rendered surfaces, and can be overlaid (color-coded) by various attribute layers. The surface's actual z expansion in the display needs to be scalable. This is necessary to view multiple data surfaces with different value ranges in one display.

Animation: viewing time series or rotations of complex scenes;

Multiple linked windows: enables different views of data to be active at the same time and data manipulations in one window imply immediate updates of the representations in other windows.

Accepting different data types: regular or irregular grids, continuous and point data, time series and georeferenced data.

Open architecture: the visualization tool shall provide a port to a GIS or database system.

3.4 VisMint Data Scaping Tool

The following images (screen shots) from the explorative Data Scaping tool VisMint (Visualization Methods for Data Integration) are to illustrate how different visualization techniques of 3 dimensional surface visualization may help to reveal structures and correlations of dynamic natural processes. The data used for the illustrations are gridded (environmental) data sets, various attribute layers and a digital terrain model. The data sets represent environmental information about Switzerland and have different resolutions. The interface and visualization tool were implemented as application framework with the advanced visualization system IDL (RSI, 1993).

Various data techniques supported by VisMint and an efficient user interface provide viewing and interaction of regular gridded data in two to four dimensions. In the first figure one can see parts of the user-interface, where data sets can be selected. Further it shows two windows with different views of the *altitude* (elevation) data layer. One is showing a color coded image and one presents the data as rendered surface. In this case the surface representation was chosen to be a wire frame, with 81 per cent of the points of the full resolution. Many other choices of display parameters and and manipulations are possible with the shown surface viewer tool. Choice by mouse and direct feedback is provided.

Figure 2 shows a combined view of the yearly temperature distribution. In the above window, an image view, wire frame surface and a contour plot are combined. Below the same data set is represented as a surface and as a histogram. The windows can be linked. Selecting data points in the histogram plot will highlight the selected points in the other images, so that their spatial distribution can be looked at.

The third screen shot shows a combination of the two data sets, mean annual temperature and yearly precipitation sum. Both data sets are combined as shaded and as wire frame surface in the above window with the so-called multiple surface viewer. Sliders and buttons allow the user to rearrange and manipulate the display, e.g., moving and scaling objects, viewing angel and surface representation. The profile

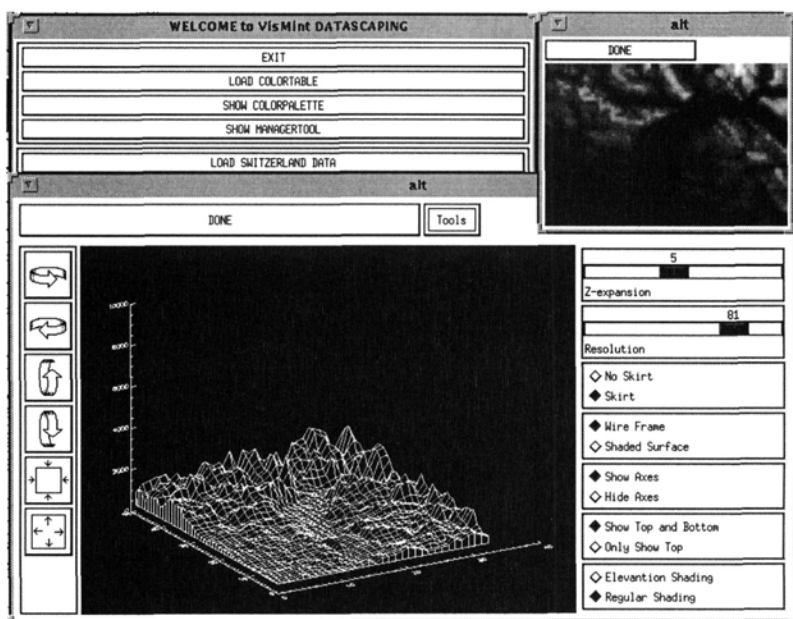


Figure 1: VisMint Data Scaping: main menu and surface viewer

plot in the window below provides a linked viewing of the two surfaces, The two profile plots show a vertical cut through the two surfaces and changes dynamically when the mouse is positioned in the window with the two surfaces. The simultaneous viewing yields at the different characteristics of the two data layers. The two smaller windows show again a color image and a shaded surface of the temperature data.

The program provides many more selected data visualization techniques for displaying one or more geographically referenced variables and some techniques for animating time-series. These techniques enable many interrelated natural phenomena to be shown in a format that fosters comparison. By offering data visualizable in images that correspond very closely to the data's real world appearance one can better explore spatial phenomena and validate and compare preprocessed data. This kind of visualization, in the sense of real world simulation is particularly interesting, as the visualized data first have to be mathematically computed. As they can be altered, different scenarios can be carried out and compared with one another.

Usually, all images can interactively be altered according their viewing angle, color, and scaling, which help to perceive 3 dimensions (and time) on a 2 dimensional screen. Interactive graphics has to be experienced, static prints of sceendumps can only give an impression of their potential.

The presented visualization tool VisMint Data Scaping is still under development and being further completed. It currently being use in an environmental modelling project and the evaluation of its significance to data integration is further investigated.

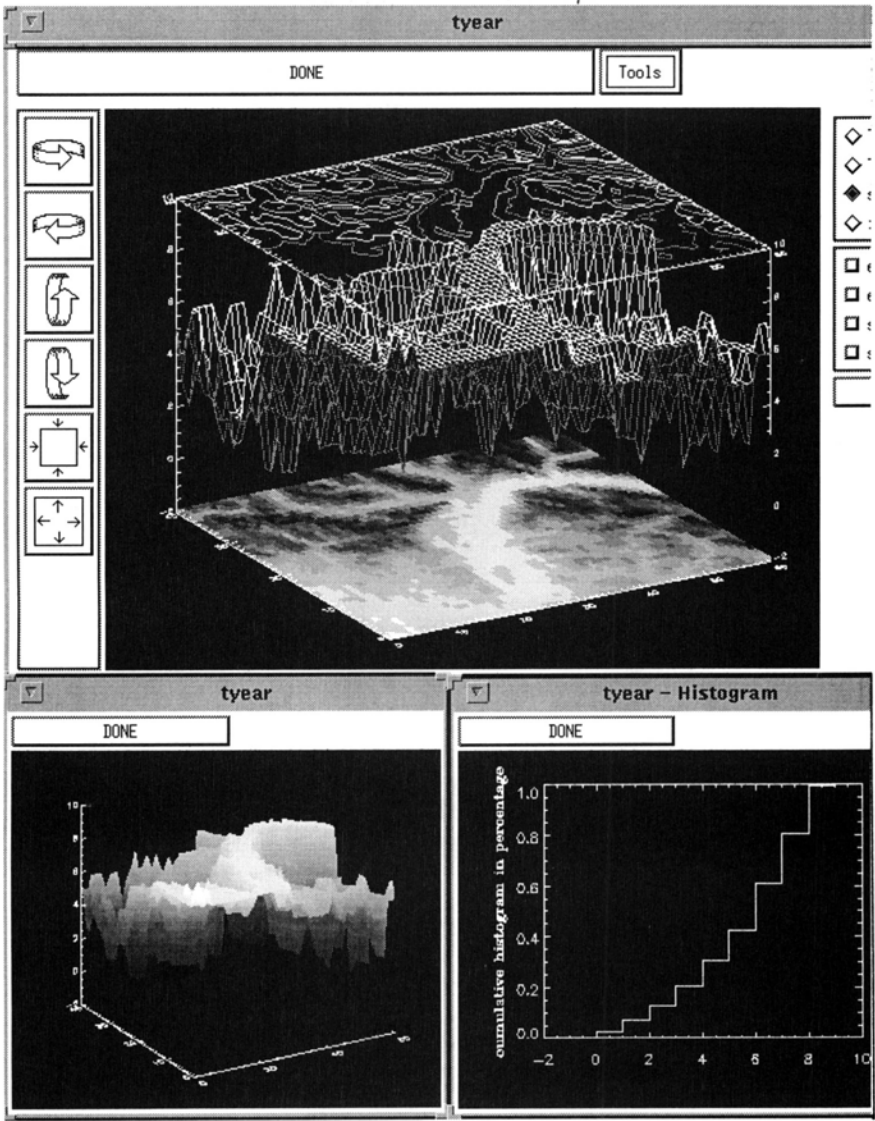


Figure 2: VisMint Data Scaping: combined viewing of a data set

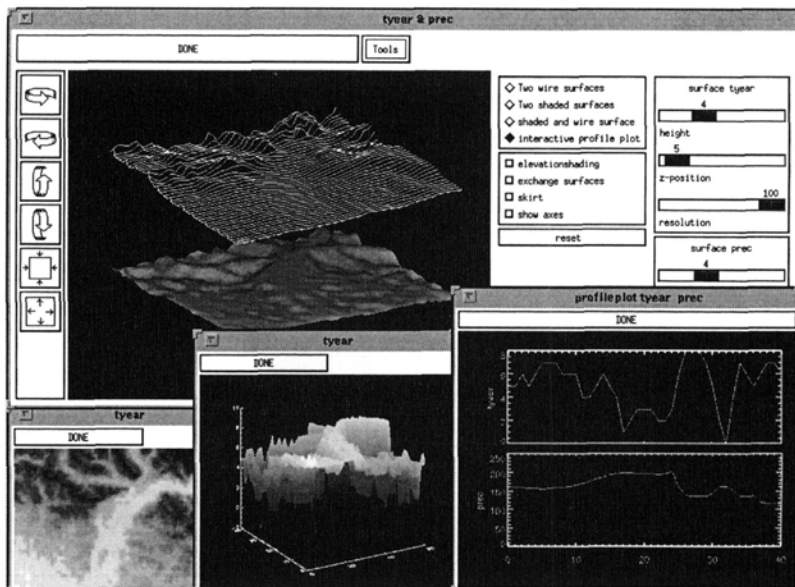


Figure 3: VisMint Data Scaping: two surface viewer in interactive profile plot mode

4 Conclusion

The previous sections seek to stress the use of interactive multidimensional visualization techniques from the computer graphics and scientific visualization community for *interactive explorative analysis of environmental data in GIS*. The concept of *Data Scaping* which was introduced, relies on the idea of heavily linking data and their display to gain more insight into data to study their *natural behavior and interrelations*. A implementation of a viewing tool showed that interactive techniques for the interactive examination of environmental data sets can help to better understand data and its underlying natural phenomena.

Data Scaping is to be seen in connection with the VDS concept. The concept requires a profound understanding of data for defining suitable methods for prediction and data quality assessment, which can obtained by visual data exploration. Visualization techniques can serve for exploring uncertainties in two ways, multidimensional exploratory statistical analysis and a multidimensional viewing of geographic surfaces. This paper outlines both approaches and presents visualization techniques such as graphical statistics, dynamic graphics and visual cue techniques for EDA. Using methods for exploring environmental data can contribute much to reveal uncertainties and to gain better insight into accurate modelling.

References

- Aspinall, R. J. 1993. Exploratory Spatial Data Analysis in GIS: Generating Geographical Hypothesis Describing Farm Types from Spatial Census Data. *Pages 517–526 of: Conference Proceedings EGIS'93.*
- Bertin, Jacques. 1967. *La Semiologie Graphique.* Gauthiers-Villars.
- Bucher, Felix, Stephan, Eva-M., & Vckovski, Andrej. 1994. Standardization and Integrated Analysis in GIS. *Pages 67–75 of: Proceedings EGIS'94.*
- Burrough, Peter, & Hazelhoff, L. 1994. *Environmental Assessment and GIS.* EGIS'94 Workshop.
- Hasslett, J., Wills, G., & Unwin, A. 1990. SPIDER - an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems.*
- Isaaks, E.H., & Srivastava, M.R. 1989. *An Introduction to Applied Geostatistics.* Oxford University Press.
- MacEacheren, A.M., M., Alan, Buttenfield, Barbara P., Campbell, James B., DiBiase, David, & Monmonier, Mark. 1992. *Visualization.* Rudger's University Press.
- RSI. 1993. *IDL - Interactive Data Language.* Research Systems, Inc.
- Stephan, Eva-M., Vckovski, Andrej, & Bucher, Felix. 1993. Virtual Data Set: An Approach for the Integration of Incompatible Data. *Pages 93–102 of: Proceedings of the Eleventh International Symposium on Computer Assisted Cartography.*
- Tufte, E.R. 1983. *The Visual Display of Quantitative Information.* Graphics Press.
- Tukey, J.W. 1977. *Exploratory Data Analysis.* Adisson Wesley.
- Wood, M., & Broodlie, K. 1994. *ViSC and GIS: Some Fundamental Considerations.* West Sussex, England: Wiley.