Visual Data Exploration: Tools, Principles, Problems

GENNADY L. ANDRIENKO and NATALIA V. ANDRIENKO
Fraunhofer AIS – Institute for Autonomous Intelligent Systems, Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany
WWW: http://www.ais.fraunhofer.de/and/
E-mail: gennady.andrienko@ais.fraunhofer.de

Our 1999 IJGIS paper (Andrienko and Andrienko 1999) was about maps as tools for exploratory analysis of spatially referenced data. After the paper was published we continued working on tools and techniques for exploratory analysis of spatial and, more recently, spatio-temporal data. We have written many other papers concerning this topic and have just finished writing a book, in which we try to approach the topic in a systematic way and to lay theoretical and methodological foundations for exploratory data analysis.

Exploratory data analysis (EDA) means open-minded looking at data with the aim to detect and describe patterns, trends, and relationships and to generate hypotheses, which can then be tested using mathematical methods. The very nature of EDA implies that data visualisation plays a crucial role. As is defined in a dictionary, "visualise" denotes "to make perceptible to the mind or imagination". In EDA, it is the mind of a human explorer that is the primary tool of analysis. It is the task of the mind to detect and describe patterns and to generate hypotheses. However, the mind can fulfil its mission only if the data to be explored are made perceptible to it. No thinking is possible without prior perception. Hence, data visualisation is the most important supporting tool (i.e. supplementary to the human mind) in EDA.

For the perception and further reasoning to be valid, a visual representation of data must be structurally similar (isomorphic) to the pertinent features of the underlying phenomenon (Arnheim 1997). A specific implication of this requirement is that geographically related data can be properly explored only with the use of maps as isomorphic representations of the geographical space. It was therefore natural that we started our research into tools for exploratory analysis of spatial data with maps as primary means of visualisation of such data. In this way, our research interests became very close to those of the Commission on Visualisation of the International Cartographic Association (see http://kartoweb.itc.nl/icavis/).

In fact, what we did was fully in line with the research agenda of the Commission, with its "emphasis on the role of highly interactive maps <…> at hypothesis generation, data analysis, and decision-support" (MacEachren and Kraak 1997). Therefore, the community of researchers in geovisualisation clustering around the Commission enthusiastically accepted our work and adopted us as its integral part. It is interesting that this community does not consist purely of cartographers but also of scientists from other disciplines including geography, computer science, psychology, and statistics. This creates a fertile ground for a synergy of expertises and experiences, ideas and approaches.

Like many other geovisualisers, we spent certain time in experimenting with the possibilities provided by the computer screen, a so plastic and so responsive representation medium, and thereby contributed to the establishment of the concept of "interactive map". In particular, our paper shows that an interactive map allows, in

addition to zooming, querying, and brushing, also "playing" with the parameters of the cartographic representation method applied in it. Such "playing" can increase map expressiveness and expose initially hidden features, patterns, trends, and relationships.

Our paper differed from contemporaneous works of other researchers concerned with the exploratory use of maps: our primary focus was the investigation of the properties and capabilities of maps as such while the others mostly considered maps in combination with other tools, for example, statistical graphics or computational techniques. This does not mean, however, that we denied the utility of tool integration for the exploratory data analysis. We just wanted to gain a deep understanding of maps before extending the scope of our research to other tools.

It is generally recognised that maps alone can seldom be sufficient for a comprehensive exploration of spatial data. In most real cases, data complexity requires interactive maps to be used in combination with other exploratory and analytical tools. Therefore, it is natural that the geovisualisers keep an eye on what is going on in related disciplines such as information visualisation, statistics, and data mining and adopt techniques and approaches from these disciplines. Our work after publishing of the paper also developed in this direction. We have designed and implemented quite a number of exploratory tools and investigated their capabilities. Our particular interest was to discover synergistic effects from joint uses of several tools, especially in combination with interactive maps. The development and further investigation of the tools was often actuated by practical needs, i.e. encountering datasets that could not be properly explored with the available tools. In this manner, helped by our colleagues, we have gradually built a quite powerful toolkit for exploratory analysis of spatial data, with a variety of tightly integrated tools that can be used in diverse combinations.

Among others, we have developed a range of tools supporting the exploration of spatio-temporal data, i.e. data with both spatial and temporal components. There are different kinds of spatio-temporal data requiring different sets of exploratory tools. Thus, discrete events such as earthquakes have to be analysed in a different way than movements of objects or temporal sequences of measurements made in a set of spatial locations. Our focus on spatio-temporal data was quite in line with the general trends of the research in geovisualisation and geoinformation science.

Hence, after publishing the paper about interactive maps, our research developed mostly in breadth, by embracing other types of exploratory tools and extending to other types of space-related data. The same can be said about the geovisualisation research in general. In the recent collective book presenting the state of the art in this area, most of the contributions deal with new visualisation techniques, new tool combinations, or new technologies (Dykes et al. 2005).

What was the matter of our concern is the absence of theoretical or methodological framework both for doing comprehensive EDA and for building tools to support it. Is data exploration an arbitrary sequence of trials, with some of them bringing serendipitous discoveries, or it is (or should be) a purposeful and systematic course of actions? If the latter is true, what is (or should be) the underlying system? What actions are involved and how they are organised? If performing the actions requires specific tools, how can an analyst know which tool is suitable for what action? On the other hand, how can a tool designer anticipate the needs of an analyst and to build a tool combination that covers those needs?

Since our methodical and rationalistic minds could not accept the view of EDA as a haphazard activity, so we decided to uncover the underlying structure, principles, and driving forces. We switched from developing in breadth to moving in depth. However,

the previous in-breadth development was a necessary prerequisite for this movement. We have studied a wide range of tools and tool combinations in application to various datasets and in this way acquired a great deal of valuable knowledge and experience. Now we could proceed by reflecting on these knowledge and experience and generalising them. The resulting theory is presented in our book (Andrienko and Andrienko 2005).

The approach we apply to define the principles of EDA can be characterised as task-centred. We view EDA as consisting of tasks, i.e. finding answers to various questions about data. To find the answers, an analyst needs to apply appropriate tools. The character of a task determines what kind of tool (from a range of tools potentially applicable to the given data) can appropriately support it. Hence, a tool designer that knows the potential tasks of the explorer can create a set of tools capable to cover explorer's needs. Accordingly, the keystone of our theory is the definition of the tasks emerging in EDA. More precisely, we introduce a system that allows one to define the set of tasks pertinent to a specific dataset or a class of datasets with similar structures (since the tasks are not defined in an abstract way but in terms of data components).

Besides designers of tool for EDA and researchers in geovisualisation, we want our theory to be useful for guiding data analysts in choosing and applying exploratory tools and in doing EDA in a systematic, comprehensive way. Therefore, we do not stop after introducing our task framework but make two further steps. First, we catalogue the existing tools suitable for EDA, describe their properties and capabilities (in particular, what tasks they are capable to support) and provide examples of their use. Second, we portray EDA as a systematic procedure composed of analytic and synthetic activities. We describe how high-level tasks are decomposed into lower-level ones and how the results of the lower-level tasks are integrated into comprehensive answers to the high-level questions. We uncover the general principles directing the process of EDA. We do not present these principles as our original invention. On the opposite, we firmly believe that any experienced data analyst follows these principles, at least intuitively.

Since the principles have been explicitly described, they can be used to educate or guide novices. Tool developers can use the principles as an organising basis for building EDA toolkits and designing their user interface. The results may be beneficial not only for novices but also for experts, who will almost certainly find such systematically designed toolkits more convenient and easy to use. Moreover, a novel exploratory technique appearing as a part of such a system may cause fewer problems for an analyst since its purpose is clear from its particular place in the system.

We recognise that the way from the formulation of the principles and the practical application of them for designing toolkits and guiding users may happen to be not very short and not quite obvious. Still, we think that this way worth trying since its promises at least a partial solution to one of the most serious problems of exploratory tools, namely, the Usability Problem. This problem, which is pertinent not only to geovisualisation but also to information visualisation and other sorts and flavours of visualisation, means that data analysts, both novices and experts, are, in general, incapable or unwilling to use the superb tools and techniques created for them.

One of the reasons is the lack of knowledge of the EDA concept and principles (which is true at least with respect to novices) and the unconventionality of the EDA tools and techniques. Another reason is the complexity of the EDA process, which is very demanding of an analyst since it is the human mind that is the primary instrument of analysis. The variety of tasks and corresponding tools, the complexity of real data necessitating the exploration by pieces and slices followed by a laboured synthesis of an

overall picture from multiple tiles makes EDA a difficult job. Intelligence embedded in a toolkit for EDA could be an adequate response to these challenges. Intelligent software could "know" the principles of EDA and help and even prompt the users to act in accord with them. It might "know" the tools and assist the users in choosing and utilising them. It could implement the analysis workflow step by step and relieve the users from the cognitive complexity of the EDA process. It could automatically adapt itself to the particular user, data, tasks, and hardware.

However, there are also other causes for the Usability Problem, and tool designers are, in a great deal, responsible for them. Thus, many researchers tend to implement software prototypes demonstrating particular ingenious approaches rather than toolkits intended to cover the needs of people exploring some type of data. The tools created by different researchers are incompatible and, hence, the needs cannot be covered in the way of combining several tools. Furthermore, data analysts need not only tools that allow them to make discoveries and generate hypotheses but also means to verify these discoveries and to test the hypotheses. Hence, exploratory tools should be linked with appropriate confirmatory techniques, for example, be implemented as extensions to statistical packages.

One more challenge, which seems to be not yet properly realised in the visualisation research community, is related to the fact that observations and discoveries that people make in the course of visual data exploration cannot be conveniently captured for later recall and for communication to others. Primary results of using visual data displays are visual impressions, or mental images, which are hard to verbalize or express in any other form without referring to the displays from which they originate. The difficulty of recording and reporting the findings is a serious obstacle to wide recognition and use of visualisation tools.

Hence, the Usability Problem is complex and multi-faceted and requires systematic joint efforts of many researchers and tool designers to be fully solved. We will certainly try to contribute to solving the problem and hope to see a substantial progress in the near future.

Returning to the 1999 IJGIS paper and our subsequent work, we would like to note that the Web supplement to the paper with a number of interactive maps is still available at a new address http://www.ais.fraunhofer.de/and/IcaVisApplet/. Besides, various demonstrators of EDA tools and tutorials explaining their purposes and usage can be accessed from our homepage http://www.ais.fraunhofer.de/and/.

References

ANDRIENKO, G. and ANDRIENKO, N., 1999. Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, **13**(4), 355–374

ANDRIENKO, N. and ANDRIENKO, G., 2005. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*, Springer, Berlin (in press)

ARNHEIM, R., 1997, *Visual Thinking*, University of California Press, Berkeley 1969, renewed 1997

DYKES, J., MACEACHREN, A.M., and KRAAK, M.-J. (eds.), 2005, *Exploring Geovisualization*, Elsevier, Amsterdam

MACEACHREN, A.M. and KRAAK, M.-J., 1997, Exploratory cartographic visualization: advancing the agenda. *Computers and Geosciences*, **23**, 335-344.