"NVIZN" FEDERAL STATISTICAL DATA ON THE WEB

Jürgen Symanzik*, Utah State University, Lacey Jones, Michigan State University.
*Department of Mathematics and Statistics, Logan, UT 84322–3900,
e–mail: symanzik@sunfs.math.usu.edu

Key Words: Digital Government, Graphics Production Library, Interactive Statistical Graphics, Micromaps, WWW.

Abstract

nViZn, read "envision", the successor of the Graphics Production Library (GPL), is a set of Java class libraries for interactive statistical graphics on the Web. In this paper, we provide a short overview of the Digital Government initiative and present approaches on how to display geographically linked data on the Web. We also describe how nViZn can be used to distribute Federal statistical data on the Web, using hierarchical clickable micromaps.

1. Introduction

Major changes concerning the publication and distribution of statistical data from Federal Agencies are to be expected in the near future. Instead of publishing hard copies of statistical data in tables, it is likely that in the future a large amount of data will be distributed through interactive graphical displays on the Web.

In this paper we describe our efforts to develop user–friendly Web–based interfaces for accessing Federal statistical data. These interfaces should enable the user to make selections that influence the visualization of the accessed data. The main graphical components that will be used are different versions of micromaps.

In Section 2 of this paper, we report on the Digital Government initiative. In Section 3, we focus on efforts that allow one to visualize geographically linked data on the Web. The concept of micromaps is introduced and a new software product, nViZn, is presented. We finish with a discussion in Section 4.

2. The Digital Government Initiative

The Digital Government (dg.o) initiative (`http://www.diggov.org`) is a major research initiative funded by the National Science Foundation (NSF) and Federal Agencies such as U.S. Environmental Protection Agency (EPA), U.S. Department of Agriculture — National Agricultural Statistics Service (USDA–NASS), Census, National Cancer Institute (NCI), Bureau of Labor Statistics (BLS), etc.

> "dg.o is a collaboration among academic researchers, government agencies, and the private sector that promotes National Science Foundation (NSF)–sponsored emergent information technologies by creating research partnerships. dg.o provides a forum in which partners can work, learn from each other, discover new research opportunties, and in which potential collaborators can be matched based upon research domain and other common interests. dg.o seeks to assist in the formation of research collaborations, leverage information technology research and identify financial resources to help build the Digital Government of the 21st Century."
> (`http://www.diggov.org/about/index.cfm`)

The Digital Government initiative addresses multiple aspects related to Federal data such as visualization, access, disclosure, security, etc. One of the proposals funded under dg.o is the Digital Government Quality Graphics (DGQG) project (`http://www.geovista.psu.edu/grants/dg-qg/index.html`).

One of the components of the DGQG project is the conversion of tables to graphs and maps. This includes the development of applications based on the commercial software package nViZn, introduced in Section 3.2, and other computer middleware that can produce a wide range of graphs and maps suited to represent statistical summaries on the Web. Non–commercial software for this task is under development at the GeoVISTA Center (`http://www.geovista.psu.edu/`) at Penn State University. Examples such as HealthVis, a visualization system designed to help scientists working with epidemiological data to explore the data for patterns or trends through scatterplot matrix widgets and dynamic parallel coordinate plots, implemented in TCL (Tool Command Language), can be accessed through the GeoVISTA Web page. Our efforts to display Federal data on the Web using nViZn are described in Section 3.2.

3. Geographically Linked Data on the Web

The distribution of statistical data through printed tables has never been an effective and fast approach for understanding large data sets. Multiple rows and columns, spread across multiple printed pages, make it difficult to locate interesting information in a printed table at a glance. Research for Federal Agencies in converting tables to graphs and maps has been conducted over the last decade (Carr 1994*a*, Carr 1994*b*, Carr & Yang 1994, Carr & Nusser 1995, Carr & Olsen 1996). Due to the current Internet and WWW technology, it is now possible to present interactive tables, graphs, and maps on the World Wide Web. Tables, graphs, and maps can be rearranged by the click of a mouse. The user can zoom into a subset of the data and look at summaries at different geographic levels. Applications using Web–based tools are very helpful to quickly find structure in large data sets, in particular for data sets from Federal Agencies that have a geographic component.

One possible presentation technique for data in a geographic context is to use linked micromap plots, often simply called micromaps (Carr & Pierson 1996, Carr, Olsen, Courbois, Pierson & Carr 1998, Carr, Olsen, Pierson & Courbois 1998, Carr, Olsen, Pierson & Courbois 2000). The main idea behind micromaps is to focus the viewer's attention on the statistical information presented in such a plot. Multiple, small generalized maps are used to provide the appropriate geographic reference for the data. The idea to use micromaps on the Web was first considered for the EPA's Cumulative Exposure Project Web site. Meanwhile, micromaps have also been used by authors other than the main developers and the interested Federal Agencies, e.g., Fonseca & Wong (2000).

Currently, the U.S. Department of Agriculture — National Agricultural Statistics Service (USDA–NASS) Research and Development Division provides a graphical representation using micromaps of data from the 1997 Census of Agriculture. USDA–NASS displays acreage, production, and yield of harvested cropland (Corn, Soybeans, Wheat, Hay, and Cotton) on one of its Web sites (`http://www.nass.usda.gov/research/sumpant.htm`). The user can sort the states by acreage or by yield with respect to the selected crop. It is possible to select another crop type and to access and download the raw data for further analysis. While the end user who accesses this Web site gets the impression of full interactivity, this is not the case. The 10 micromaps (5 crops × 2 arrangements) plus one overview micromap have been precalculated and are stored as jpg images. It is not possible to create any new micromap display "on the fly" on this Web site. However, precalculating all possible micromaps is often not possible or desirable for all data sets as we will see in the next section.

3.1. EPA's Cumulative Exposure Project

In 1998, the U.S. Environmental Protection Agency (EPA) was interested in the development of an interactive Cumulative Exposure Project (CEP) Web site (`http://www.epa.gov/CumulativeExposure/`). Initially, the main intention behind the development of this Web site was to provide fast and convenient access to the EPA's hazardous air pollutant (HAP) data for 1990. Concentrations of 148 air pollutants were estimated for each of the 60,803 census tracts in the 48 contiguous states (Caldwell, Woodruff, Morello–Frosch & Axelrad 1998, Woodruff, Axelrad, Caldwell, Morello–Frosch & Rosenbaum 1998, Rosenbaum, Axelrad, Woodruff, Wei, Ligocki & Cohen 1999). The Web site design was proposed to allow the user to easily move through the data set to find information on different pollutants at different locations and at different levels of geographic resolution.

Three stages had been planned for the CEP Web site release. In the first stage, users should have been given full access to the HAP data in tabular form, but no major interaction was intended. In the next stage, users could have been able to interact with the tables and rearrange the data according to several criteria. A preview page that contains some of the interactive table features can be accessed at `http://www.galaxy.gmu.edu/~symanzik/gpl/CEPstart/DATAstartfull.html`. An interested reader may wish to contact Daniel Axelrad at `axelrad.daniel@epa.gov` for the User ID and password to access the preview site.

In the final stage, users would have had access to the HAP data via hierarchical clickable micromaps paired with graphical representations of the data. It was intended to implement these features through the Graphics Production Library (GPL). The GPL (Carr, Valliant & Rope 1996) is a set of Java class libraries for interactive statistical graphics, developed within the Bureau of Labor Statistics (BLS). It was initially intended to add interactivity, such as drag and drop comparisons, panel reordering and rescaling, and pan and zoom to the row–labeled plots of Carr (1994*b*). The GPL followed recent recommendations on statistical graphics as given in Cleveland (1993) and Cleveland (1994). The design of the GPL

also addressed the display of times series and allowed to incorporate metadata, such as warning flags on the time series and links to articles on the time series adjustments. The library was developed to facilitate Web distribution of statistical summaries from the BLS. Examples of its capabilities can be found at `http://www.monumental.com/dan_rope/gpl/`.

Unfortunately, no part of the interactive CEP Web site was ever published due to concerns that the 1990 data was outdated at the intended release date in 1998. Only a static version of the CEP Web site without tables and micromaps is accessible at `http://www.epa.gov/CumulativeExposure/`. The 1990 HAP data can be obtained upon request directly from the EPA. Due to EPA's decision not to release this data on the Web, work on the third stage of the CEP Web site has never been completed. More details on the work related to the planned interactive CEP Web site can be found in Symanzik, Wong, Wang, Carr, Woodruff & Axelrad (2000), Symanzik, Axelrad, Carr, Wang, Wong & Woodruff (1999), and Symanzik, Carr, Axelrad, Wang, Wong & Woodruff (1999).

3.2. Hierarchical Clickable Micromaps under nViZN

Hierarchical clickable micromaps on the Web would allow the user of a Web site to start at the highest level of a geographic hierarchy, e.g., with the entire United States. Then the user could click on a state to obtain a state micromap. In the next step, the user could click on a county to obtain a census tract micromap. Moving up in the reverse direction should also be possible. Similar hierarchies could be useful for health data, using Health Service Areas (HSAs) as the second stage, or for data from natural resources, using ecoregions as the second stage. To allow such interaction on the Web, sophisticated software is needed.

A new, commercial version of the GPL, called "nViZn" (read "envision") (`http://www.illumitek.com/`), has been developed since 1997 and was released in 2000. It is described in Wilkinson, Rope, Carr & Rubin (2000) where the name GPL is still used. nViZn is a JAVA–based software development kit (SDK). It is based on the formal grammar for the specification of statistical graphics introduced in Wilkinson (1999). In addition to the basic capabilities inherited from the old GPL, nViZn has many additional features. Most useful for the display of data in a geographic context are the capabilities that enable a programmer to create data–linked micromaps in nViZn. Experiences with nViZn, its advantages and current problems, are described in more detail in Jones & Symanzik (2001). In the remainder of this section, we will show how we can use nViZn to display Federal statistical data on the Web.

Figure 1 shows a micromap of the mean (with respect to the underlying census tract estimates) modeled 1990 acetaldehyde concentrations for Iowa from the EPA HAP data set. According to the EPA Web page `http://www.epa.gov/opptintr/chemfact/f_acetal.txt`, acetaldehyde is a HAP that "occurs naturally in certain foods, such as ripe fruits and coffee, and in cigarette smoke. [...] The largest users of acetaldehyde are companies that make acetic acid and related chemicals. Companies also use acetaldehyde to make other chemicals such as pyridine, pentaerythritol, and peracetic acid." Not surprisingly, two of the five highest modeled concentrations of acetaldehyde in the air can be found in highly industrialized counties in Iowa, i.e., in Scott County (with Davenport) and in Dubuque County (with Dubuque). Pottawattamie County (with Council Bluffs) is not as highly industrialized as the previously mentioned two counties, but faces the highly industrialized Omaha in Nebraska. High concentrations of acetaldehyde can also be found in heavily populated counties such as Polk County (with Des Moines), Woodbury County (with Sioux City) Linn County (with Cedar Rapids), Johnson County (with Iowa City), Black Hawk County (with Waterloo), and Story County (with Ames). Muscatine County itself is not heavily populated or industrialized, but is located close to Scott County. Lowest mean modeled 1990 acetaldehyde concentrations can be found in sparsely populated, agricultural counties of Iowa.

Figure 2 shows a micromap of the median (with respect to the underlying census tract estimates) modeled 1990 benzene and lead concentrations for Utah from the EPA HAP data set. Similar to Iowa, highest concentrations of these HAPs can be found in heavily populated and highly industrialized counties such as Weber, Salt Lake, Davis, Utah, Wasatch, and Tooele. However, while benzene originates from many different sources, lead is more closely related to industry. This becomes obvious in the micromap display. There is some correlation visible in these two HAPs, but in particular extremely high values of benzene do not necessarily relate to similarly high values of lead.

4. Discussion

As outlined in more detail in Jones & Symanzik (2001), nViZn is a relatively new software development kit. Very little documentation on its use

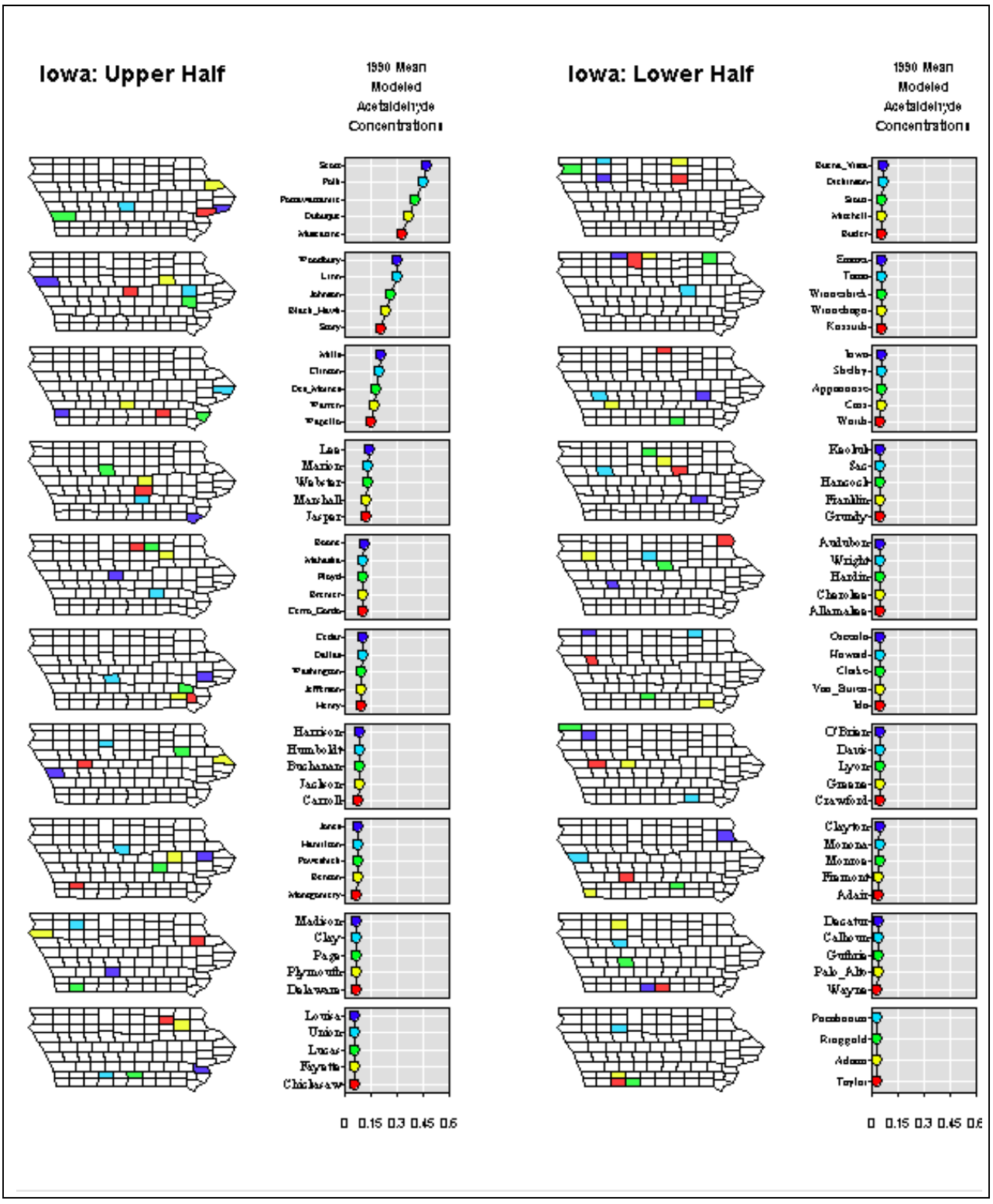

Figure 1: Mean Modeled 1990 Acetaldehyde Concentrations (in $\mu g/m^3$) for Iowa. Highest concentrations can be found in highly industrialized and heavily populated counties.

is available so far. In particular, neither introductory books nor books for advanced users exist. To ease the learning of nViZn, Illumitek, the creators of nViZn, has recently begun to offer training courses

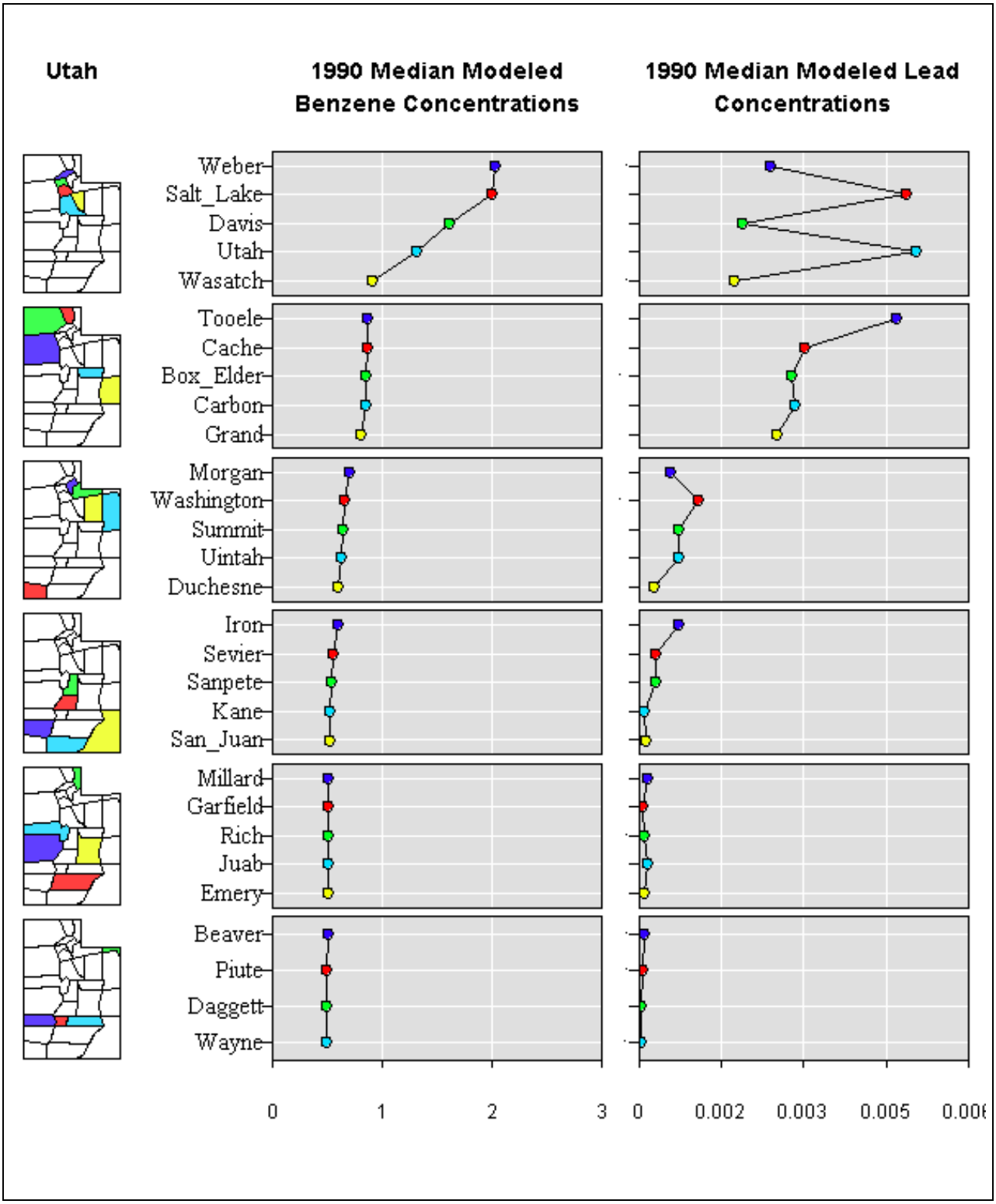

Figure 2: Median Modeled 1990 Benzene and Lead Concentrations (in $\mu g/m^3$) for Utah. Highest concentrations can be found in highly industrialized and heavily populated counties.

on its use. The company also plans to make the nViZn training material available on the Web in the near future.

Based on the existing nViZn documentation and

help obtained from the nViZn developers, we are now able to develop most static tabular and micromap displays that are likely to occur in the context of Federal statistical data. We are still at an early learning stage as far as it involves interaction with these displays, e.g., in order to produce hierarchical clickable micromaps under nViZn.

Based on our experiences so far, nViZn is certainly a powerful software development kit that allows software developers to build Web–based applications for the display and exploration of statistical data from government, academia, and industry. One should mention that an application developer under nViZn needs to have good JAVA knowledge. Similar to most JAVA–based packages, developers need to check carefully whether each application that has been developed using nViZn really works as intended under different Web browsers on different hardware platforms.

Future work with nViZn and micromaps on the Web involves implementation and design issues. First, we plan to finish the prototypic implementation of hierarchical clickable micromaps under nViZn using the 1990 EPA HAP data. Then, we plan to adapt this prototype to current data from Federal Agencies such as NCI, USDA–NASS, or EPA. One question of particular interest is how to link an nViZn application to Federal statistical data bases. Finally, we have to address design and usability issues. While side–by–side micromaps such as in Figure 1 are effective on paper, we do not know yet whether they also work on a computer screen or whether scrollable micromaps are better suited for this medium. Certainly some usability study should also be conducted to answer the question how well non–statisticians understand graphical displays with micromaps.

Acknowledgements

The work of Jürgen Symanzik was supported in part by the NSF "Digital Government" (NSF 99–103) grant #EIA-9983461 and by a New Faculty Research Grant from the Vice President for Research Office from Utah State University. Lacey Jones' work was supported in part by an U.R.C.O. Grant from the Office of the Vice President for Research from Utah State University.

References

Caldwell, J. C., Woodruff, T. J., Morello–Frosch, R. & Axelrad, D. A. (1998), 'Application of Health Information to Hazardous Air Pollutants Modeled in EPA's Cumulative Exposure Project', *Toxicology and Industrial Health* **14**(3), 429–454.

Carr, D. B. (1994*a*), 'A Colorful Variation on Box Plots', *Statistical Computing and Statistical Graphics Newsletter* **5**(3), 19–23.

Carr, D. B. (1994*b*), Converting Tables to Plots, Technical Report 101, Center for Computational Statistics, George Mason University, Fairfax, VA.

Carr, D. B. & Nusser, S. M. (1995), 'Converting Tables to Plots: A Challenge from Iowa State', *Statistical Computing and Statistical Graphics Newsletter* **6**(3), 11–18.

Carr, D. B. & Olsen, A. R. (1996), 'Simplifying Visual Appearance by Sorting: An Example using 159 AVHRR Classes', *Statistical Computing and Statistical Graphics Newsletter* **7**(1), 10–16.

Carr, D. B., Olsen, A. R., Courbois, J. P., Pierson, S. M. & Carr, D. A. (1998), 'Linked Micromap Plots: Named and Described', *Statistical Computing and Statistical Graphics Newsletter* **9**(1), 24–32.

Carr, D. B., Olsen, A. R., Pierson, S. M. & Courbois, J. P. (1998), 'Boxplot Variations in a Spatial Context: An Omernik Ecoregion and Weather Example', *Statistical Computing and Statistical Graphics Newsletter* **9**(2), 4–13.

Carr, D. B., Olsen, A. R., Pierson, S. M. & Courbois, J. P. (2000), 'Using Linked Micromap Plots to Characterize Omernik Ecoregions', *Data Mining and Knowledge Discovery* **4**(1), 43–67.

Carr, D. B. & Pierson, S. M. (1996), 'Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps', *Statistical Computing and Statistical Graphics Newsletter* **7**(3), 16–23.

Carr, D. B., Valliant, R. & Rope, D. (1996), 'Plot Interpretation and Information Webs: A Time–Series Example from the Bureau of Labor Statistics', *Statistical Computing and Statistical Graphics Newsletter* **7**(2), 19–26.

Carr, D. B. & Yang, K. (1994), 'Variations on Row–Labeled Plots for Reexpressing Tabular Summaries', *Computing Science and Statistics* **26**, 436–440.

Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, NJ.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, Summit, NJ.

Fonseca, J. W. & Wong, D. W. (2000), 'Changing Patterns of Population Density in the United States', *The Professional Geographer* **52**(3), 504–517.

Jones, L. & Symanzik, J. (2001), 'Statistical Visualization of Environmental Data on the Web using nViZn', *Computing Science and Statistics* **33**, Forthcoming. (CD).

Rosenbaum, A. S., Axelrad, D. A., Woodruff, T. J., Wei, Y.-H., Ligocki, M. P. & Cohen, J. P. (1999), 'National Estimates of Outdoor Air Toxics Concentrations', *Journal of the Air and Waste Management Association* **49**, 1138–1152.

Symanzik, J., Axelrad, D. A., Carr, D. B., Wang, J., Wong, D. & Woodruff, T. J. (1999), HAPs, Micromaps and GPL — Visualization of Geographically Referenced Statistical Summaries on the World Wide Web, *in* 'Annual Proceedings (ACSM–WFPS–PLSO–LSAW 1999 Conference CD)', American Congress on Surveying and Mapping.

Symanzik, J., Carr, D. B., Axelrad, D. A., Wang, J., Wong, D. & Woodruff, T. J. (1999), Interactive Tables and Maps — A Glance at EPA's Cumulative Exposure Project Web Page, *in* '1999 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 94–99.

Symanzik, J., Wong, D., Wang, J., Carr, D. B., Woodruff, T. J. & Axelrad, D. A. (2000), Web–based Access and Visualization of Hazardous Air Pollutants, *in* 'Geographic Information Systems in Public Health: Proceedings of the Third National Conference August 18–20, 1998, San Diego, California', Agency for Toxic Substances and Disease Registry. `http://www.atsdr.cdc.gov/GIS/conference98/`.

Wilkinson, L. (1999), *The Grammar of Graphics*, Springer, New York, NY.

Wilkinson, L., Rope, D. J., Carr, D. B. & Rubin, M. A. (2000), 'The Language of Graphics', *Journal of Computational and Graphical Statistics* **9**(3), 530–543.

Woodruff, T. J., Axelrad, D. A., Caldwell, J. C., Morello–Frosch, R. & Rosenbaum, A. S. (1998), 'Public Health Implications of 1990 Air Toxics Concentrations Across the United States', *Environmental Health Perspectives* **106**(5), 245–251.