

# XML-based Applications in Statistical Analysis

Yuichi Mori<sup>a,\*</sup>, Tomokazu Fujino<sup>b</sup>, Yoshiro Yamamoto<sup>c</sup>,  
Takafumi Kubota<sup>d</sup>, Tomoyuki Tarumi<sup>e</sup>

<sup>a</sup>*Department of Socio-Information, Okayama University of Science, Okayama 700-0005, Japan*

<sup>b</sup>*Department of Environmental Science, Fukuoka Women's University, Kasumigaoka, 813-8529, Japan*

<sup>c</sup>*Department of Mathematics, Tokai University, Hiratsuka 259-1292, Japan*

<sup>d</sup>*Graduate School of Natural Science and Technology, and*

<sup>e</sup>*Department of Environmental and Mathematical Sciences;*

*Okayama University, Okayama 700-8530, Japan*

---

## Abstract

The increasing use of web-intensive environments gives rise to the need for a treatment of statistical information in a unified format. For this purpose, Extensible Markup Language (XML) technologies may be applied as a useful tool. The present study introduces various XML-based applications that are currently being developed by the authors. Specifically, a database of data sets and analysis stories which are written in XML, an on-line analysis system allowing automatic analyses to be performed with initial parameters provided from XML documents, Web-based statistical graphics and maps using Scalable Vector Graphics and Extensible 3D, and an on-line interactive textbook using XML technologies. These applications will be considered in terms of aspects such as data-description, accessibility to statistical engines, documentation and interactivity.

*Keywords:* XML; SVG; X3D; Web application; databank; interactive graphics; GIS application

---

## 1 Introduction

In the Internet age, many things are becoming increasingly Web-based. Statistics is no exception to this trend: there now exist many sites publishing information on statistical theories and methods; several on-line analysis systems for both general and special uses are available; and on-line textbooks and databanks of data sets, mainly for educational purposes, also exist. However, in order to improve the existing statistical environment it is desirable to provide more useful databanks, which include not only data sets, but also related information on the data, tools for visualizing the data with a high degree of interactivity, and analysis systems for special topics. Given that many tools based on Extensible Markup Language (XML), a language designed for electronic publishing and exchange of data on the Web and elsewhere, are already available, the application of this language provides a potentially attractive means of providing the desired statistical environment described above. There already exist several projects providing standards or foundations to describe data and its related information, and XML-based interactive graphics and mathematical expressions, which are necessary to describe statistical contents, are also being developed. Such projects and studies make full use of the characteristics of XML; ease of presentation using the Extensible Stylesheet Language (XSL) family, easy transformation of XML

---

\* Corresponding author. Tel. & Fax: 81-86-256-96521. *E-mail:* mori@soci.ous.ac.jp.

documents into other documents using XSL Transformations (XSLT), and the availability of tools handling XML documents. Furthermore, since XML has a good relationship with several other Web technologies such as Dynamic HTML, common gateway interface (CGI), Flash, JavaScript and Java, the production of dynamic and interactive web applications can also be facilitated.

The present study introduces various XML-based applications that are currently being developed by the authors, including: a database of data sets and analysis stories which are written in XML, on-line analysis system in which automatic analyses can be performed with initial parameters provided from XML documents, Web-based statistical graphic tools and Geographic Information Systems (GIS) applications using Scalable Vector Graphics (SVG) and Extensible 3D (X3D), and also an on-line interactive textbook using XML technologies. Since these applications have their own original development objectives, we will focus on these objectives as well as how to apply XML technologies to meeting them. We will consider the applications in terms of aspects such as data-description, accessibility to statistical engines, documentation and interactivity.

## 2 XML technologies

In statistical research areas, XML may be used in order to describe data and statistical models in a unified format, and to exchange data between different systems for the purposes of, for example, on-line analysis. This allows us to establish a standard or common format to describe and exchange data and related things in plain text. Furthermore, Web-based statistical tools such as data visualization on the Web can be developed using such XML technologies.

In this section we outline some previous applications of XML including a database of data sets, and also outline XML tools developed to visualize data using SVG and X3D.

### 2.1 Data description in XML

In scientific research fields, there are several XML-based standardizations including, for example, MathML (<http://www.w3.org/Math/>), which is cast as an application of XML and provides a foundation for mathematical expressions for high-quality visual display on Web pages, and OpenMath (<http://www.openmath.org/>), which is an emerging standard using XML for representing mathematical objects to allow them to be exchanged between computer programs, stored in databases or published on the Web pages.

In statistical research areas, there are also many XML-based applications. For example, the Data-oriented Statistical System (DoSS@d, <http://mo161.soci.ous.ac.jp/@d/>, see Section 3.1) and the MD\*Base (<http://www.quantlet.org/mdbase/>) are databanks or platforms for sharing statistical data on the Web, in which data and related information are described in a unified format based on XML. The Data Documentation Initiative (DDI) at the Inter-university Consortium for Political and Social Research (ICPSR, <http://www.icpsr.umich.edu/index.html>) and DandD (Data and Description) project (<http://www.stat.math.keio.ac.jp/DandDII/>) are projects seeking to establish a standard of data description. The former provides an international XML-based standard for the content, presentation, transport, and preservation of documentation for data sets in the social and behavioral sciences, and is already being used by major projects such as Networked Social Science Tools and Resources (Nesstar, <http://www.nesstar.com/>). The latter is a project to publish data and its description jointly, which provides support systems (DandD browser and server) as well as DandD description rules. The Predictive Model Markup Language (PMML) is a mark up language developed by the Data Mining Group (DMG, <http://www.dmg.org>), which provides a way to define statistical and data mining models and to share these models between different applications. StatDataML is another project to establish a standard for the exchange of data between statistical and mathematical packages such as S, R, MATLAB and Octave.

### 2.2 Data visualization with XML technologies

#### 2.2.1 SVG (Scalable Vector Graphics)

Data visualization is a very useful and important technique within statistics. In particular, interactive and dynamic graphics are necessary for exploratory data analysis.

SVG is a two-dimensional XML-based vector graphics format standardized by W3C (<http://www.w3.org/Graphics/SVG/>). W3C released SVG as an integrated format embracing the Microsoft-led Vector Markup Language (VML) and Adobe-led Precision Graphics Markup Language (PGML), which were proposed before SVG. Since SVG has two parts, an XML-based file format, and a programming interface for graphical applications, it can hold not only graphical data, but also its related information. The functions for interactivity and a zooming function are implemented, not by the server, but by the plug-in<sup>1</sup> package for the client browser. A disadvantage of interactive graphical systems based on a combination of raster graphics such as PNG, GIF and JPEG, with Web application development languages such as Perl, PHP, JSP and ASP, is that such combinations require the server-side application to re-generate graphical elements for every client request including zooming and rotation. However, SVG avoids this problem because once an SVG file is created, no further internet connection is required. Furthermore, since SVG files are written in plain text, SVG can be handled easily: SVG files can be provided in XML format as open and standard specifications, they can be generated by any programming language without the need for any libraries for the generation of graphics, and also other XML data files can be straightforwardly converted and loaded to SVG by XSLT and Document Object Model (DOM).

For the reasons outlined above, the advent of this open and standard graphics format was warmly received by many developers, and many tools for SVG, and applications using SVG, have subsequently been developed. Furthermore, existing applications have also developed support to output SVG. Thus SVG may be expected to become widely used in statistical environments in the future.

### 2.2.2 X3D (Extensible 3D)

X3D is an open standard for three-dimensional representation on the Web. Since X3D is a successor of the Virtual Reality Modeling Language (VRML), it is compatible with existing VRML contents. The virtual world written in X3D format based on XML is expressed through the X3D plug-in or the X3D browser<sup>2</sup>.

A further advantage of X3D over XML-based technologies is that users do not have to program to generate two-dimensional representation of three-dimensional data; X3D provides a platform of 3D graphical expression in which virtual space can be turned and moved without programming.

## 3 Examples of XML-based applications

In this section we outline Web-based systems and applications which make use of the features of XML: a database system of real data sets and analysis stories along with an online analysis system, interactive statistical graphics, GIS applications and an online textbook system. We illustrate the original purposes of developing such applications as well as how to utilize XML technologies in them.

### 3.1 DoSS@d (Data-oriented statistical system)

The Data-oriented statistical system **DoSS@d** is a databank located on the Web (Mori et al., 2003) at <http://mo161.soci.ous.ac.jp/@d/>. This databank represents an online database of data sets and documentation describing the processes of the original analyses (we call this kind of documentation “analysis story”), and also incorporates an online analysis system that performs (semi-)automatic analysis based on the analysis story (i.e., using the same parameters as those ones of the original analysis). The reason why the name of this system includes “@**d**” is to reinforce the idea that the system is used for real data. This Web-based system therefore consists of

<sup>1</sup> The Adobe SVG Viewer for Internet Explorer on Windows is a de facto standard as a rendering engine of SVG. The Batik Squiggle which is Java-based SVG Viewer is released by the Apache XML Project as a part of the applications for SVG. The Mozilla Project developing open source Web browser has recently implemented natively rendering SVG to Mozilla (SVG-enabled Mozilla).

<sup>2</sup> Shared virtual worlds written in X3D may be accessed by an X3D plug-in or an X3D browser. MediaMachines Flux X3D/VRML97 plug-in and Venues X3D Viewer plug-in for Windows; FreeWRL for Mac OS and Linux OS.

two functions: a database of typical real-world data sets and analysis stories, and an analysis system with a graphical user interface to allow data sets in the database to be analyzed online. Both systems are based on XML technologies, i.e., descriptions of data attributes and analysis stories are written in XML and online analysis is performed according to the parameters described in the XML file of the analysis story. When used for statistical education, teaching scenarios can be developed easily, giving the students an opportunity to learn various statistical techniques using real data sets, as well as mastering statistical software using the online analysis function. Students can also perform their own analysis to confirm the results of the analysis story through the use of simple operations, and can easily examine the effect of using different parameters.

**DoSS@d** consists of three subsystems; **DoDStat@d** (Data-oriented Database of Statistics), **DoAStat@d** (Data-oriented Analysis System of Statistics), and **DoLStat@d** (Data-oriented Learning System of Statistics). We will focus mainly on the first two of these here.

### 3.1.1 DoDStat@d

**DoDStat@d** is the database system of **DoSS@d**, in which data sets are classified by research subject and statistical method. The user is able to select an interesting or appropriate data set using research subject and/or statistical method as a retrieval key, as well as a keyword (Fig. 1 displays the results of an example search). Each stored data set consists of the data description and the data body (Fig. 2). The former is written in XML to describe information about the data such as the data's name, a brief description, source of the data, research subject, statistical method, case attributes (number of cases and case name) and variables attributes (number of variables, variable name and variable type). The latter is provided in several formats, including tab-, comma- and space-delimited values. **DoDStat@d** also stores analysis stories written in XML which describe the title of the story, goal of the analysis, link to the original data set, source of the story, research area, analysis process, parameters for the statistical method applied in the original analysis, and also links to online analysis and pages describing how to use general statistical packages to obtain the same result as the story (Fig. 3 is an analysis page, Fig. 4 shows a section of the XML source file, and Fig. 5 is a screenshot of a Java application of online analysis).

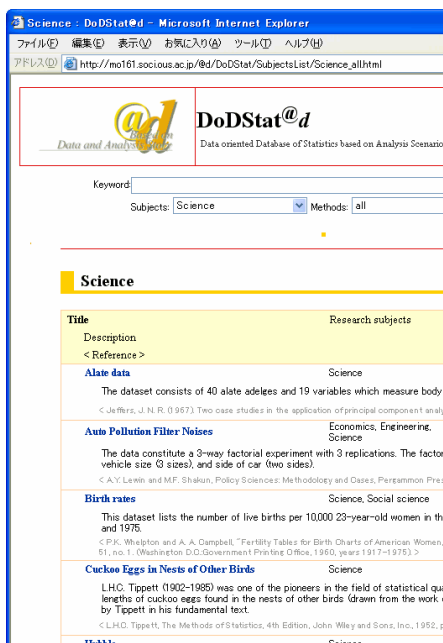


Fig. 1: Result of a search with the key "science" in **DoDStat@d**. All stored data sets are classified by research area and statistical method.

No	Name	Label	Description	Type
1	length	body length	body length	numerical
2	width	body width	body width	numerical
3	fwing	fore-wing length	fore-wing length	numerical
4	hwing	hind-wing length	hind-wing length	numerical
5	numsp1	number of spiracles	number of spiracles	numerical
6	antseg1	length of antennal segment 1	length of antennal segment 1	numerical
7	antseg2	length of antennal segment 2	length of antennal segment 2	numerical
8	antseg3	length of antennal segment 3	length of antennal segment 3	numerical
9	antseg4	length of antennal segment 4	length of antennal segment 4	numerical
10	antseg5	length of antennal segment 5	length of antennal segment 5	numerical
11	numaspi	number of antennal spines	number of antennal spines	numerical
12	tarsus	leg length, tarsus	leg length, tarsus	numerical
13	tibia	leg length, tibia	leg length, tibia	numerical
14	femur	leg length femur	leg length femur	numerical
15	rostrum	rostrum	rostrum	numerical
16	ovipos	ovipositor	ovipositor	numerical

Fig. 2: Data description page. This page illustrates information about the data such as the data's name, a brief description and case and variable attributes as shown in the screenshot. The user can download the data set in several formats from this page.

**[DoSS@d-STORY] Physical measurement of alate adelges**

- Title:** Physical measurement of alate adelges
- Goal:** To find sub groups in which individuals have similar characteristics with each other based on their global scores.
- Data:** Alate adelges
- Source:** Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis, *Appl. Statist.*, 16: 225-236.
- Research area:** Science (Biology)
- Analysis story:** There is a dataset which consists of 40 alate adelges and 19 variables which measure body parts. We wish to find sub groups in which individuals have similar characteristics with each other based on their global scores.

**[PCA]** We apply principal component analysis to this dataset. Looking at the scree plot of the data and a scree graph of them, we determine the number of principal components.

Scree plot of eigenvalues

Here the number of principal components is two because the cumulative proportion of variance is 83% and other eigenvalues are almost the same (less than 1). Then we determine the number of principal components and observe a scatter plot of the first two scores.

As shown in this plot, four sub groups are mainly observed. Therefore it can be classified into four groups based on the observed 19 variables.

Click the [Analysis] button, you can perform on-line analysis of the dataset based on the above story.

Click a button, you can see how to use the software based on the above story using the software.

**Analysis**   **SPSS**   **R**

```

<title>
  Physical measurement of alate adelges
</title>
<goal>
  To find sub groups in which individuals have similar characteristics with each other based on their global scores.
</goal>
<dataset>
  <name>alate</name>
  <title>Alate adelges</title>
  <url>alate_dataE.xml</url>
</dataset>
<area>
  <subject>Science</subject>
  <detail>Biology</detail>
</area>
<source>
  Jeffers, J. N. R. (1967).
  Two case studies in the application of principal component
  analysis, Appl. Statist., 16: 225-236.
</source>
<story>
  <description>
    There is a dataset which consists of 40 alate adelges and 19
    variables which measure body parts. We wish to find sub groups
    in which individuals have similar characteristics with each
    other based on their global scores.
  </description>
  <procedure>
    <subprocedure>
      <storyDescription>
        </storyDescription>
      </subprocedure>
    </procedure>
  </story>
  <execute>
    <location>mol61.soci.ous.ac.jp</location>
    <method name="pca" interactive="yes">
      <option npc="2" matrix="Cov" selectedVar="length, width,
      fwing, hwing, numspi, antseg1, antseg2, antseg3, antseg4,
      antseg5, antspi, tarsus, tibia, femur, rostrum, ovipos,
      ovispi, anal, numhooks" />
      <output showScores="yes" plotScores="yes" plotLoadings="yes"
      plotBiplot="no" />
    </method>
  </execute>
  
```

Fig. 4: A section of the XML file of an analysis story.

Fig. 3: Analysis story page. This page illustrates title of the story, goal of the analysis, link to the original data set, analysis process and so on. The buttons located at the bottom of the page are links to online analysis (a Java application like Fig. 7 appears when clicking [Analysis] button) and pages describing how to use general statistical packages to obtain the same result as the story (currently description for SPSS, R, XploRe and Excel are available; see Fig. 5).

**Analysis with R**

**Physical measurement of alate adelges**

data: [alate.csv](#)

**(0-1) Confirm the goal**

We wish to find sub groups in which individuals have similar characteristics with each other based on their global scores.

**(0-2) Prepare the analysis**

Read the data file.

```
> alate <- read.table("localfolder/alate.csv", sep=",", header=TRUE)
```

**(1) Principal component analysis**

Execute the following codes.

```
> pc.alate <- princomp(alate, cor=TRUE)
> summary(pc.alate)
> plot(pc.alate, type="t", main="Scatter plot")
```

Then the following result is output in the output window and the scree plot of eigenvalues is displayed.

Importance of components:	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Standard deviation	3.7129303	1.5373589	0.86487599	0.71037619	0.52748008
Proportion of Variance	0.7283096	0.1243933	0.03936897	0.02655985	0.01464452

Fig. 5: Description page to illustrate how to use R to obtain the same result as the analysis story. Pages for SPSS, XploRe and Excel are also available.

### 3.1.2 DoAStat@d

**DoAStat@d** is a web-based application for the analysis of any data set stored in **DoDStat@d**, as well as data sets stored on the local computer. Currently this system executes data analysis using a combination of R and the XploRe Quantlet Server (XQS) as a statistical engine. The R Server-based system DoA\_R communicates with the server by CGI, while an XQS-based system DoA\_X (Honda et al., 2004) is programmed in Java and communicates with XQS on the network using a Java communication interface called MD\*Crypt (Feuerhake, 2002; <http://www.md-crypt.com/>). Fig. 8 illustrates the architecture of DoA\_R and DoA\_X.

Users select a data set stored in **DoDStat@d** and a statistical method to be applied in the top page of **DoAStat@d** (Fig. 6). When they execute the system, DoA\_R or DoA\_X starts with the corresponding GUI to the selected method, and reads the data from the server (Fig. 8). Users then perform an analysis by selecting variables and specifying parameters in the same way as in ordinary statistical packages (Fig. 7). In addition to such ordinary online analysis, **DoAStat@d** also provides a function that allows the users to easily obtain the same results as described in the analysis story of the data by automatically importing the parameters stored in the XML document of the analysis story. This function can be used directly from the analysis story page. When clicking the [Analysis] button at the bottom in the story page (Fig. 3), for example, in which principal component analysis (PCA) is applied, DoA\_X starts automatically with the same GUI, but with all initial parameters such as matrix type and number of components for PCA fixed according to the story XML.

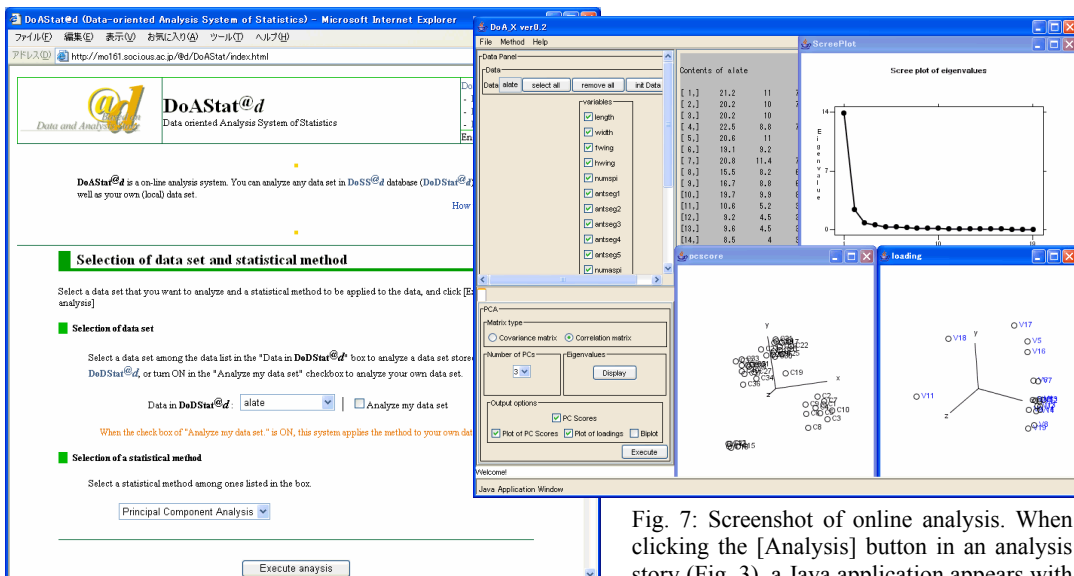


Fig. 6: Top page of **DoAStat@d**. Here users select a data set stored in **DoDStat@d** and a statistical method to be applied in this page.

Fig. 7: Screenshot of online analysis. When clicking the [Analysis] button in an analysis story (Fig. 3), a Java application appears with initial parameters specified in the <method> area of the story XML file. When this application is called from **DoAStat@d**, the same GUI starts with default parameters.

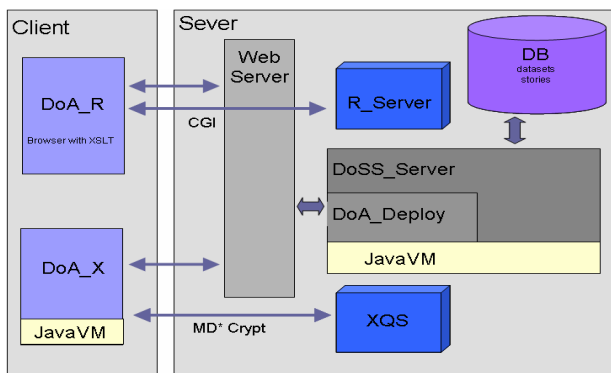


Fig. 8: Architecture of **DoAStat@d**. DoA\_R is controlled by CGI in order to communicate with R\_server and DoA\_X is a Java application to parse XML documents in the database and communicate with XQS through MD\*Crypt to analyze the data.

### 3.1.3 DoLStat@d

**DoLStat@d** is an educational system, in which a variety of educational courses such as “Statistics introductory course” and “Economics course” are provided based on analysis stories stored in **DoDStat@d** according to the study target.

## 3.2 Web-based visualization tools

Making use of the characteristics of XML described in Section 2, we are developing some useful tools and applications for interactive graphics, a library for R, a Web application and three-dimensional graphics. The latter are being developed to facilitate a user-friendly means of plotting interactive graphs even if the user is not familiar with XML related technologies such as DOM and JavaScript, which are necessary to implement the interactivity in the graphics. These tools are now available at <http://www.fwu.ac.jp/fujino/Xg4stat/> in which related information is also provided.

### (1) R library “RInG” (R Interactive Graphs)

RInG is an add-on package of R. The “RSvgDevice”, which was released by T.J.Luciani, is a well-known library to output statistical graphics of R in SVG format. It behaves similarly to the other graphical devices of R, but its output does not allow for the provision of interactivity. To enable R to output interactive graphics, we developed an R library, RInG. For example, R commands to obtain an interactive histogram is

```
library("ringlib")
SvgHist(rnorm(10000)) ,
```

where the RInG package is installed. An SVG file “SvgHist.svg” is then generated, and an interactive histogram appears by calling this file via a browser with an appropriate plug-in (Fig. 9). When a mouse pointer moves over a bar in the histogram, the upper limit, lower limit, class value and frequency of the class of the bar are displayed in the tooltip.

### (2) Web application “WInG” (Web application for Interactive Graphs)

WInG is a Web application to provide interactive statistical graphs using the mechanism of DOM. This is, for example, suitable for users who are not familiar with SVG, enabling them to experience the functions of SVG without the requirement of any special knowledge of this language. The left window in Fig. 10 is an input user interface to generate an SVG scatter plot, by which users enter data and specify parameters for the style of graph such as size, labels and colors. Clicking the [Draw plot!!] button at the bottom of this window, a scatter plot appears as the right window in Fig. 10.

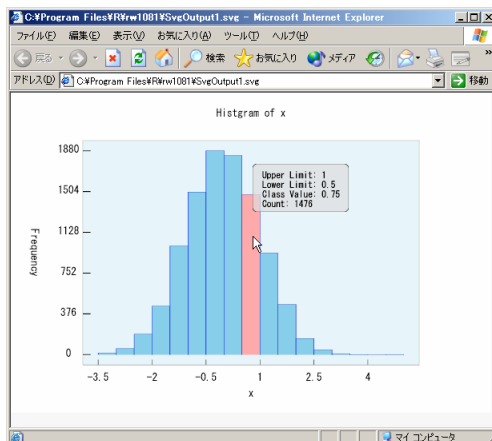


Fig.9: Interactive histogram. This is obtained by displaying a SVG file “SvgHist.svg” generated using RInG on the Internet Explorer with the Adobe SVG Viewer.

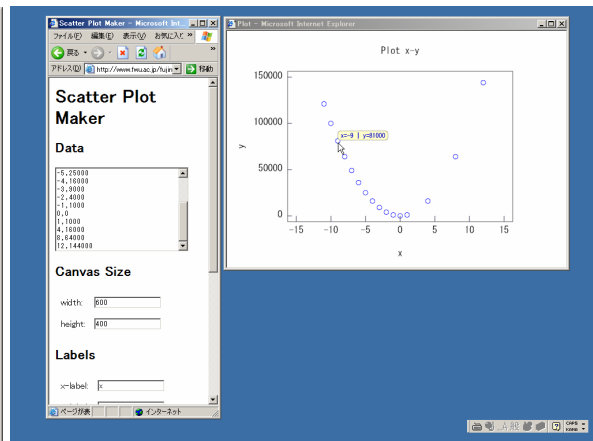


Fig. 10: Screen shot of WInG to draw a scatter plot. The left window is an input interface to enter data and specify parameters for the style of graph, while the right window displays the output of the SVG scatter plot.

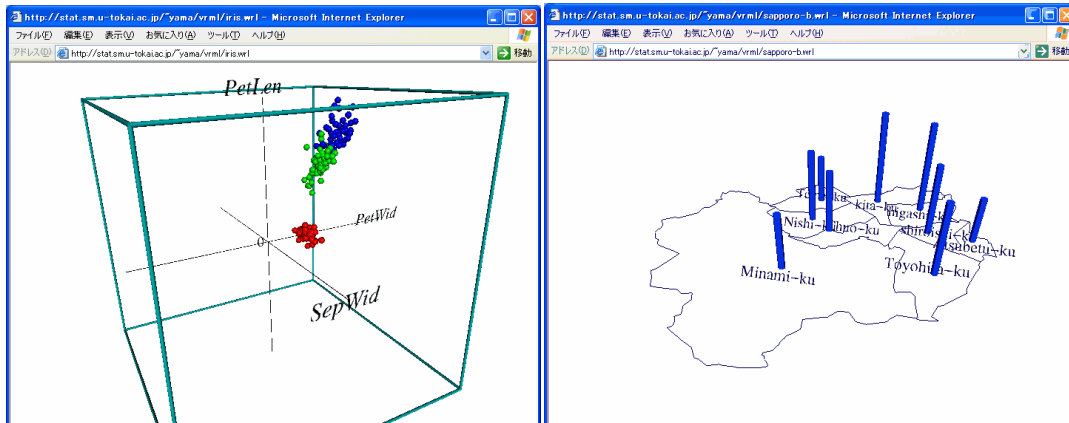


Fig. 11: 3D graphs using X3D technology. The left window displays a 3D scatter plot of Iris data, while the right window displays a 3D bar chart of populations of wards in the city of Sapporo (in northern Japan) plotted on the city map. These plots can be examined by rotation and movement.

### (3) Visualization tool using X3D

Fig 11 indicates screenshots of our X3D application, which provides an easy-to-use means of visualizing three-dimensional data. The graph can be easily rotated and moved using navigation functions in the plug-in software. Our X3D statistical graphics Web site (<http://stat.sm.u-tokai.ac.jp/~yama/x3d/> linked from <http://www.fwu.ac.jp/fujino/Xg4stat/>) provides several input interfaces like the left window in Fig. 10 to generate SVG graphics.

### 3.2.2 GIS Applications

GIS requires special functions in order to, for example, draw rich graphics, or to support vector and raster contents and also to handle a very large amount of data. SVG is well suited to application to these tasks, and many GIS systems have been incorporated into functions to export data/outputs in SVG format.

#### (1) IDS-AP (Interactive Display System of Atmospheric Pollution Data)

IDS-AP is a GIS application using XML and SVG to display the information of the area indicated by the pointer on the map. Fig. 12 is a screenshot of IDS-AP displaying air pollutant data for Fukuoka prefecture (in western Japan). Using the mouse, by simply pointing to one monitoring station on the map, the name of the station pops up at the mouse position, and three kinds of air pollutants (concentration of  $\text{SO}_2$ ,  $\text{NO}_2$  and  $\text{O}_x$ ) and detailed information on the station (name, address, type and location) are displayed in the upper right line plot area and the lower right description area, respectively.

This application consists of five files, one HTML file, two SVG files (to draw a prefecture map and a line plot) and two XML data files (location data of monitoring stations and concentration data of air pollutants). Map data and air pollutant data are obtained easily by converting the numerical national land information provided by the Ministry of Land, Infrastructure and Transport Government of Japan (<http://www.mlit.go.jp/english/index.html>) and the air pollutant observation data provided from Ministry of the Environment of Japan (<http://www.env.go.jp/en/index.html>), respectively (see the right-hand part of Fig. 12, a part of XML file of the converted air pollutant data). Using default functions of the plug-in for SVG, the map can be zoomed in or out, and moved in four directions. Thus users can observe and compare air pollution at any station and in any area, since they can access quickly and visually information of the area of interest without a need for any special programming.

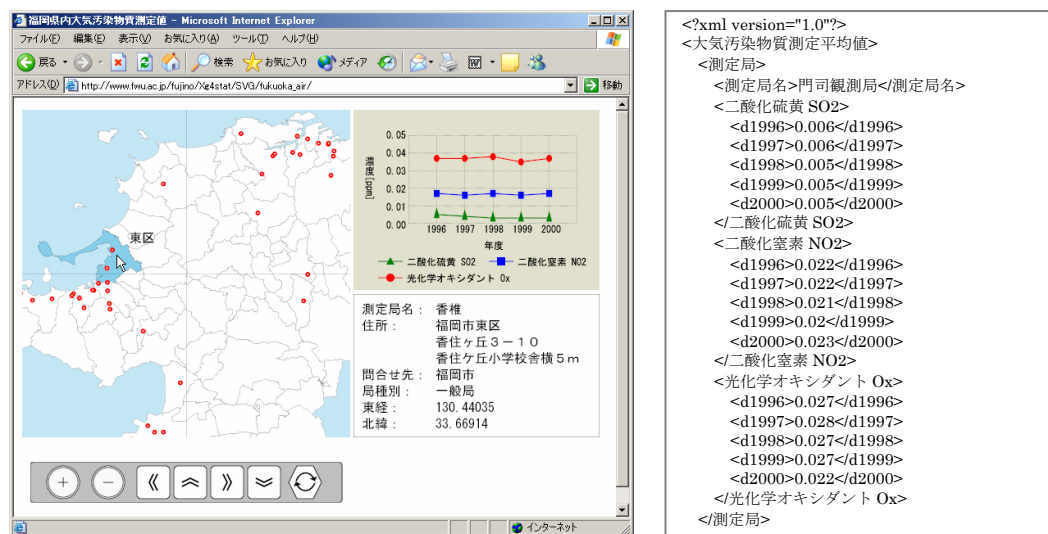


Fig. 12: Screenshot of IDS-AP for air pollutants in Fukuoka prefecture. Upon the mouse moving into a district, the district area is highlighted in light blue with the district name, and line plots of concentration of  $\text{SO}_2$ ,  $\text{NO}_2$  and  $\text{O}_x$  and the information about the monitoring stations in the district are displayed. The list on the right-hand side shows a section of the XML file of the air pollutant data.

## (2) GASWin (Geostatistical Analysis System for Windows)

The specification of a model of variogram, which is an index to measure spatial subordinates, is performed in spatial regression in geostatistics. To select a better model it is necessary to compare and evaluate various models of variogram. There exists an S extension “gstat” (provided either as R package or S-Plus library) as a geostatistics tool, but it has no functions to allow the comparison and evaluation of alternative models, and is based on a character user-interface. We are therefore developing a Web-based GUI application, GASWin, to perform spatial regression including interactive model selection of variogram using “gstat” and SVG technology.

GASWin essentially consists of CGI programs to use R as a statistical engine and an SVG output device to handle SVG files for interactivity. Specifically, Perl programs handle R packages “gstat” for geostatistical computation and “RSvgDevice” for outputs in SVG format, and interactive graphs are implemented on a browser using SVG files, with the latter being output by “RSvgDevice”. GASWin is now available at the Web address [http://face.f7.ems.okayama-u.ac.jp/GASWin/index\\_e.html](http://face.f7.ems.okayama-u.ac.jp/GASWin/index_e.html).

Analysis by GASWin consists of four steps: data entry, basic statistics, variogram analysis and kriging. We will describe a case example of data analysis by GASWin of data consisting of four variables (longitude, latitude, altitude and average temperatures in January, 1999) at 73 observation points in five prefectures (Okayama, Hyogo, Shimane, Tottori and Hiroshima in western Japan) provided by AMeDAS (Automated Meteorological Data Acquisition System) in Japan Meteorological Agency.

The user first inputs data in CSV format into the input form on the top page of GASWin (Fig. 13). Upon submitting the data, the user has to specify three variables to be analyzed. Here we assign longitude, latitude and temperatures for  $x$ -coordinate,  $y$ -coordinate and characteristic value, respectively. The user may then view a location map of observation points (a simple scatter plot of  $x$ -coordinate vs.  $y$ -coordinate) to confirm whether the specified variables are correct (Fig. 14). Since this map (also each graph generated successively) is, of course, written in SVG format, it can be zoomed in or out according to the user’s requests. The next step is one of basic statistics to examine the summary and scatter plot matrix of variables, after which GASWin computes an empirical variogram and draws its scatter plot (see the graph at the left of Fig. 15). To compare

various models of variogram, GASWin provides twelve theoretical models and their sum of squared errors at the right of the same window (see the right of Fig. 15). Upon selecting these models, GASWin quickly overlays its theoretical variogram curve on the scatter plot (Fig. 15 is already overlaid). This step is repeatable.

Based on an examination of empirical and theoretical variograms, the user proceeds to the kriging (prediction) step. Upon specifying parameters for kriging (model number and size of cell to be predicted), a graph whose coordinates are the same as the ones shown in Fig. 14, and whose cells are gradated according to the predicted values for the cell area can be obtained to visualize the levels of predicted temperatures (Fig. 16). When the user moves a mouse onto the graph, the predicted value at the mouse pointer is displayed in the text box below the graph, together with the pointer's  $x$ - and  $y$ -coordinates and the predicted error of the point. The user can also obtain the predicted value by inputting  $x$ - and  $y$ -coordinates into their text boxes from the keyboard.

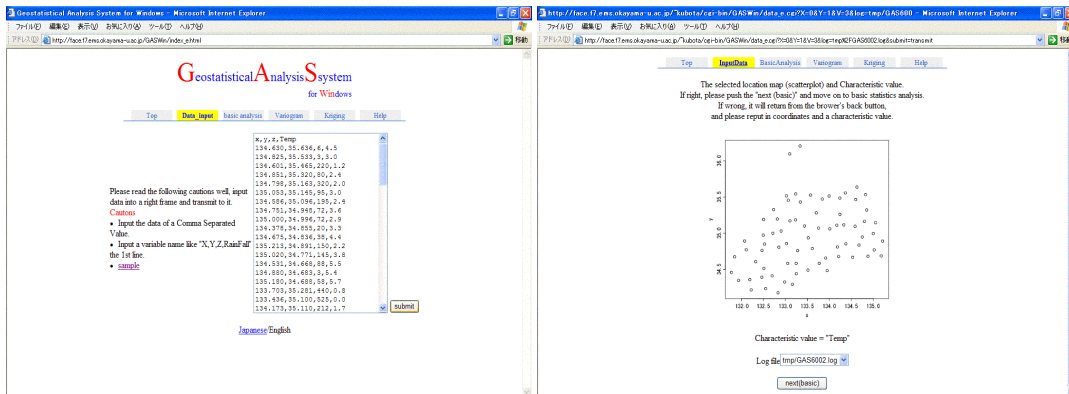


Fig. 13: Top page of GASWin (data input window).

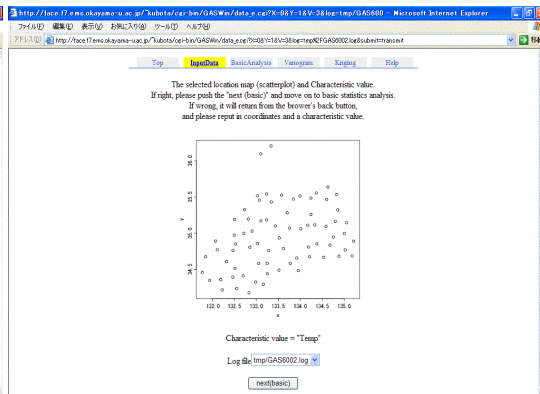


Fig. 14: Location map of observation points.

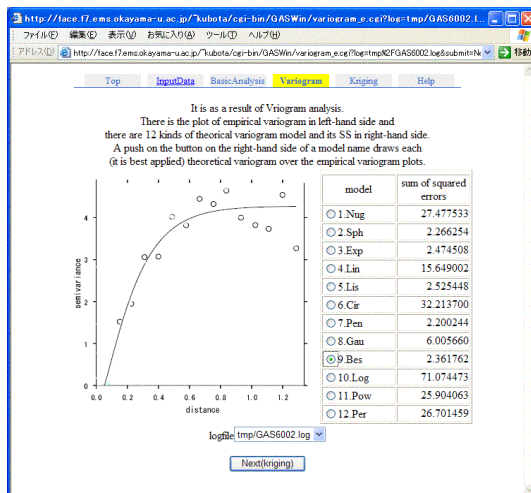


Fig. 15: Scatter plot of empirical variogram (points in the left graph) and twelve theoretical models. The curve of the selected theoretical variogram model is overlaid on the empirical variogram.

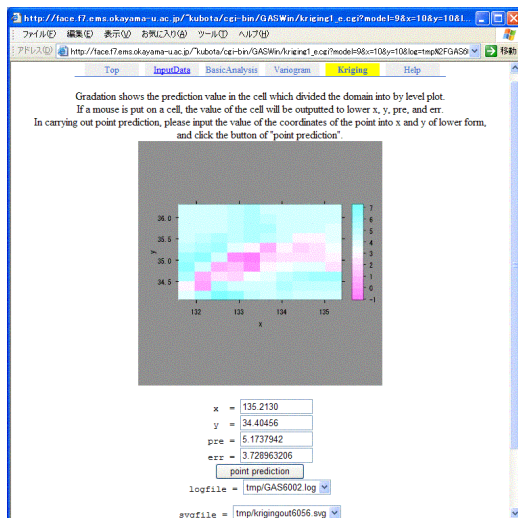


Fig. 16: Visualization of the result of kriging. The coordinates are the same as ones in Fig. 13 and cells are gradated according to the predicted values for the cell area.

### 3.3 Interactive textbook by SVG and MathML

There are several sites providing courseware for statistics on the Web. In most of these sites equations and graphics are generated as image files such as GIF and JPEG, and are embedded in

ordinary HTML texts. To implement the interactivity in the graphics, Web application languages such as Java and Flash have often been employed, but if the size of the created graphics files is relatively large, they may give rise to communication delays.

As a possible solution to overcome these problems, XML technologies may be used, i.e., Web-based interactive textbooks may be created. Each page is written basically in XML to keep a consistency of format through the book, equations are written in MathML or OpenMath, and graphics are provided in SVG or X3D format (Fig. 17). Using appropriate plug-ins or browsers<sup>3</sup>, this kind of interactive page offers a good medium for the provision of educational materials to help learners to understand statistical terms, ideas and theories. Furthermore, since these source files are based on plain texts, their creation and communication times can be expected to be relatively short.

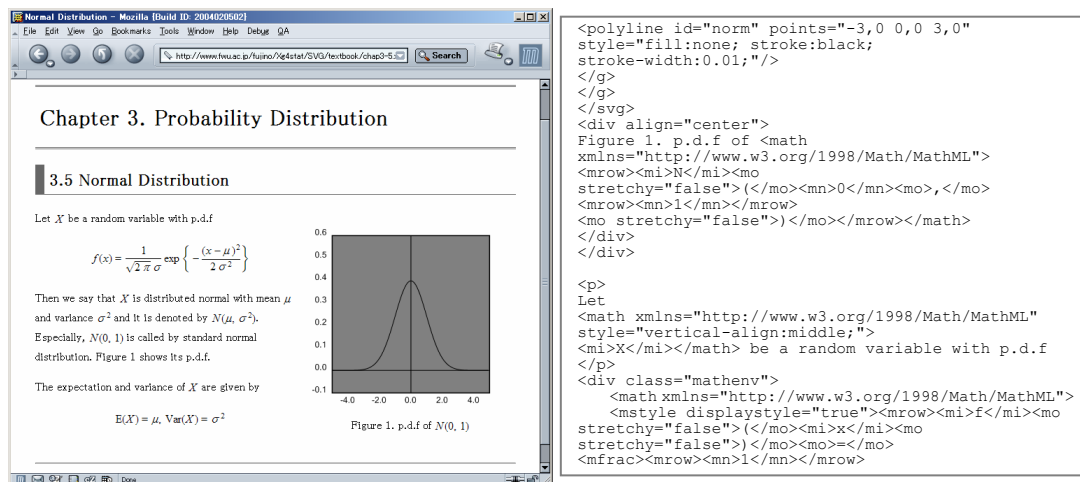


Fig. 17: Example of an interactive textbook of statistics. This page consists of documentation in XHTML, equations in MathML, and graphs implemented in SVG. A sample of the SVG source file is shown on the right-hand side.

#### 4 Discussion and remarks

In this study, we have described various Web-based applications that are being developed by the authors, which in particular use XML technologies such as XSL, SVG, X3D and MathML. Specifically, we have considered a databank of data sets and analysis stories with an on-line analysis system, Web-based interactive statistical graphics tools, GIS applications and on-line interactive textbooks. We focused on their original development objectives as well as how to apply such XML technologies.

Through a series of studies, we became aware that there exist some problems. These include, for example, how to keep security for source codes and data though the Web, how to handle commercial data and closed data, and how to implement more complicated actions in Web browser. However, non-withstanding these issues, it is clear that XML technologies have many significant advantages such as ease of presentation and standardization of data, easy transformation of data from and to XML documents, and the existence of several plug-ins, packages and libraries handling XML files. Furthermore, it is expected that an increasing number of objects, for example map data, may also be described by XML families.

<sup>3</sup> SVG-enabled Mozilla natively supports display of this kind of documentation and Internet Explorer requires the plug-in such as MathPlayer or TechExplorer to display MathML.

### Acknowledgement

The projects described in this report are partly supported by the Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science: Grant#15300094 (2003-2005) for **DoSS@d** project and Grant#15020240 (2003-2004) for others.

### References

- Mori, Y., Yamamoto, Y., Yadohisa, H. (2003). Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). *Bulletin of the International Statistical Institute, 54th Session Invited Papers, Volume LX Two Books, Book 2*, 74-77.
- Honda, K., Mori, Y., Yamamoto, Y. and Yadohisa, H. (2004). Web-Based Analysis System in Data-oriented Statistical System “**DoSS@d**”. In: *COMPSTAT 2004 Proceedings in Computational Statistics*, Heidelberg: Phisica-Verlag. (to appear)
- Feuerhake, J. (2002). XQS/MD\*Crypt as Means of Education and Computation. In: *COMPSTAT 2002 Proceedings in Computational Statistics* (Härdle, W. and Rönz, B. eds.), Heidelberg: Phisica-Verlag, 635-640.