KMITL Sci. J. Vol. 5 No. 1 Feb. 2005

Web-based Data and Story Database for Statistical Education

Yoshiro Yamamoto
Department of Mathematics, Tokai University
1117 Kita-Kaname, Hiratsuka 259-1290, Japan <yamamoto@sm.u-tokai.ac.jp>
Hiroshi Yadohisa
Department of Mathematics and Computer Science, Kagoshima University
Kagoshima 890-0065, Japan <yado@sci.kagoshima-u.ac.jp>
Yuichi Mori
Department of Socio-Information, Okayama University of Science
1-1 Ridai-cho, Okayama 700-0005, Japan <mori@soci.ous.ac.jp>

1. Introduction

Statistical information is widely used in a variety of research areas, and the advent of the computer has helped to simplify the application of statistical techniques to real data sets. Recently, the number of individuals who have to learn and/or are interested in learning how to use data analysis to solve real-world problems has increased at a rapid pace. Therefore, a quality of statistical education comes to be essential. Although there are many good textbooks, statistical classes, and e-learning systems, we often meet the following problems. First, there are often not enough examples, the purpose in collecting the data in the examples is sometimes unclear, and the existing examples rarely relate to the students' interests. Also, traditional textbooks present the material in a systematic order, building upon previous statistical knowledge and methods. These "method-oriented" education causes the following problems: the teachers can teach statistical methods easily using existing textbooks, but have to prepare good data sets if they wish to make the students perform real world problem solving; the students can learn what is the method well through the material, but often don't know how to apply the statistical techniques and methods appropriately in real situations.

Successful data analysis involves the development of problem solving ability through the collection of data, the selection of a suitable statistical method, analysis of the data, interpreting the results and finally making decisions based on the interpretation. The ability to make full use of statistical software is also desirable, involving the specification of parameters based on data attributes and operating the statistical package to obtain the desired information.

In order to mitigate the above problems, a kind of databank should be established to house a large number of data sets with ordinary database function so that anyone can use it anytime and anywhere. Furthermore, it is desirable that the databank provides guidance or documentation on how to practically analyze the data, how to interpret the results and how to use a general package to obtain the desired information (we call the first two documentations "analysis story"). This type of approach is referred to as a "data-oriented" education. The analysis story, which includes a document or record that describes the actual process used in the original analysis, is especially necessary for data-oriented education because this information takes on the role of the manual or chart typical in a similar situation to original one.

Currently, there is an abundance of web sites that provide free or semi-free access to collected data and related information. Chance Database [1], Data and Story Library (DASL) [2], Statlib [6], Data Representation System (DRS, Inoue et al., 2002) [3] and MD*Base [4], among others, could prove useful in providing statistical education. DASL and DRS, in particular, are valuable sites for data-oriented education. DASL contains both a large number of real data sets and the analysis stories, and DRS includes an interactive system that performs an immediate online analysis of any subset of data in the database.

Ideas from these two approaches are being used to develop a data-oriented statistical system over the Internet. This system is a sort of databank, which represents an online database of data sets and analysis stories, and also incorporates an online analysis system that performs automatic analysis based on the analysis story (i.e., using the same parameters as ones in the original analysis). This

web-based system therefore consists of two functions; a database of typical real-world data sets and analysis stories, and an analysis system with a graphical user interface (GUI) to allow data sets in the database to be analyzed online. This environment has been name the "Data-oriented Statistical System" or called **DoSS@*d*** (Mori et al., 2003a, b), where "@d" reinforces that the system is used for real data. When utilized for statistical education (this is regarded as the third possible function of this system), teaching scenarios can be developed easily, giving the students the chance 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.

This report presents an outline of called **DoSS@*d*** for statistical education.

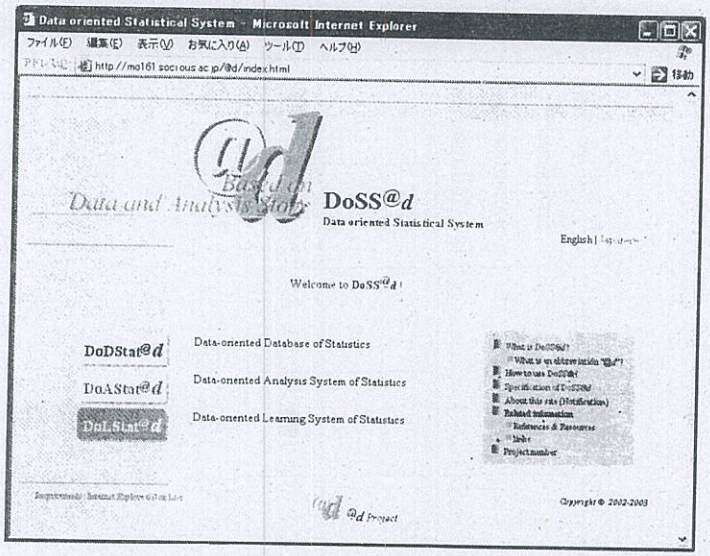

Figure 1. Top page of DoSS@d (`http://mo161.soci.ous.ac.jp/@d/`)

2. DoSS@*d* (Data oriented Statistical System)

DoSS@*d* is located at `http://mo161.soci.ous.ac.jp/@d/` (Figure 1) and 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) as shown in Figure 2.

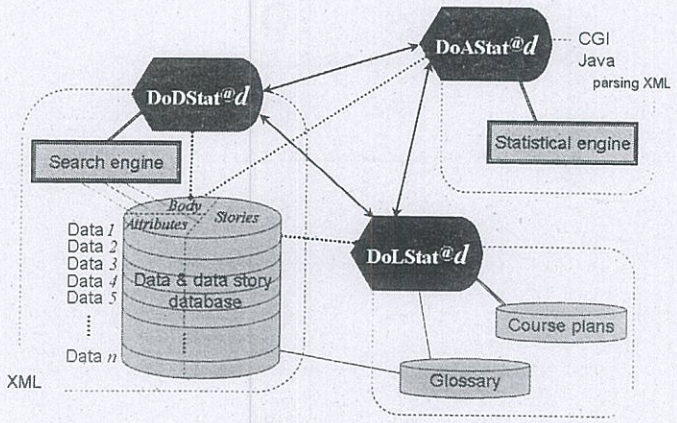

Figure 2. Structure of DoSS@d

2.1 DoDStat@*d* (Data-oriented Database of Statistics)

DoDStat@*d* is the database system of **DoSS@*d***, in which data sets are classified by research subject and statistical method. Each stored data set consists of a data description and the data body. The former is written in XML and describes attributes such as data name, case name, variable names, and variable types (e.g., Figure 4). The latter is provided in several formats, including comma-, tab- and space-delimited values. **DoDStat@*d*** also stores analysis stories written in XML (e.g., Figure 8). The user is able to select an interesting or appropriate data set using a retrieval key such as research subject, statistical method and keyword (e.g., Figure 3 (Left)).

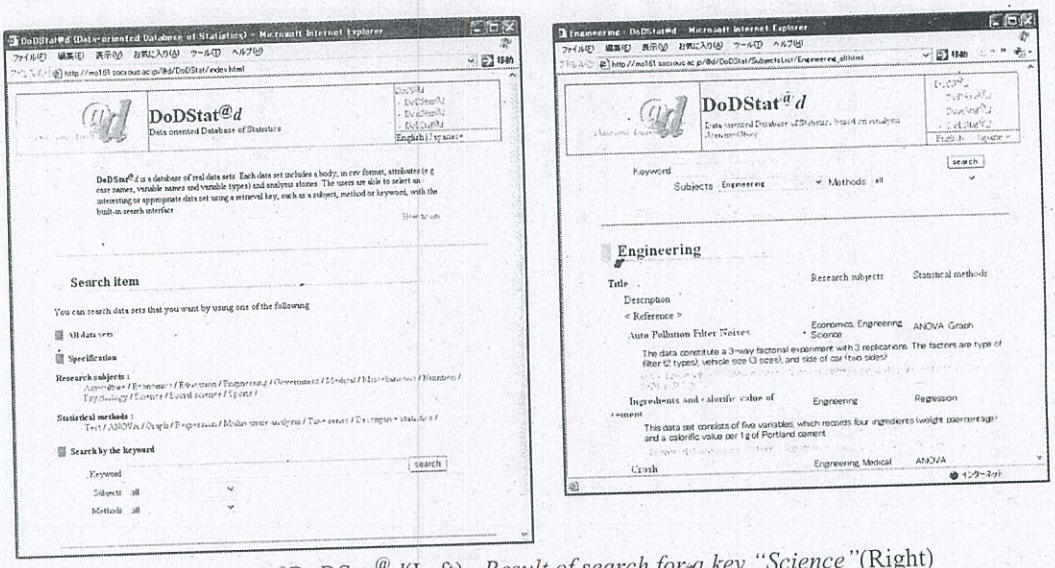

Figure 3. *Top page of DoDStat@d*(Left) , *Result of search for a key "Science"*(Right)

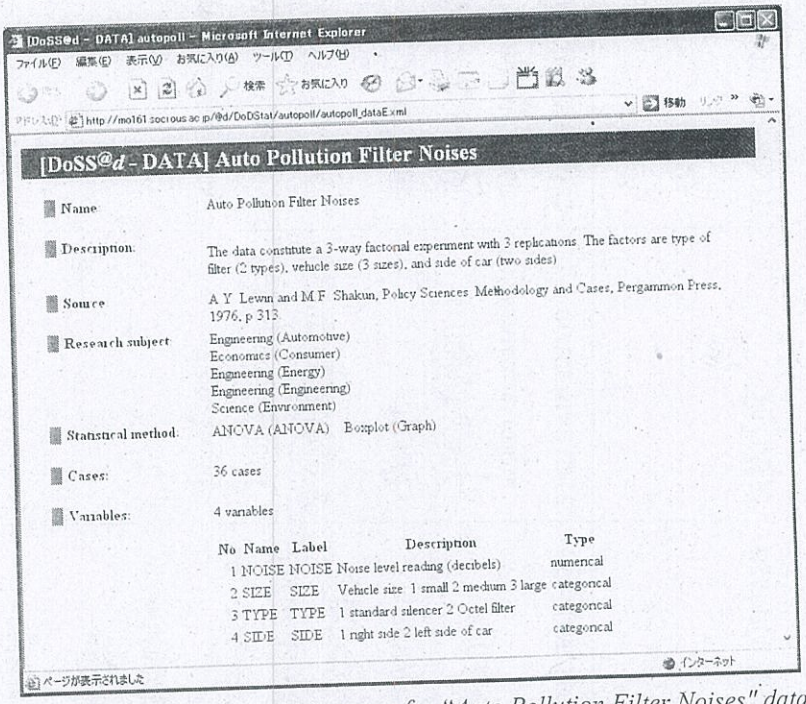

Figure 4. *Data description page for "Auto Pollution Filter Noises" data*

2.2 DoAStat@*d* (Data-oriented Analysis System of Statistics)

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 R or XploRe Quantlet Server (XQS) as a statistical engine. The R Server-based system DoA_R communicates with the server by CGI, and XQS-based system DoA_X (Honda et al., 2003) is programmed in Java and communicates with XQS on the network using a Java communication interface called MD*Crypt (Feuerhake, 2002; [5]).

Users select a data set stored in **DoDStat@*d*** and a statistical method to be applied in the top page of **DoAStat@*d*** (Figure 5). When the [Execute analysis] button at the bottom of the page is clicked, DoA_R or DoA_X starts with the corresponding GUI to the selected method and reads the data from the server. Then users perform the analysis by selecting variables and specifying parameters in the same way as in ordinary statistical packages (Figure 6). In addition to such ordinary online analysis, **DoAStat@*d*** system also provides a function that allows the users to easily obtain the same results as descried 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 (Figure 8). When clicking the [Analysis] button at the bottom in the story page, 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 set based on the story XML.

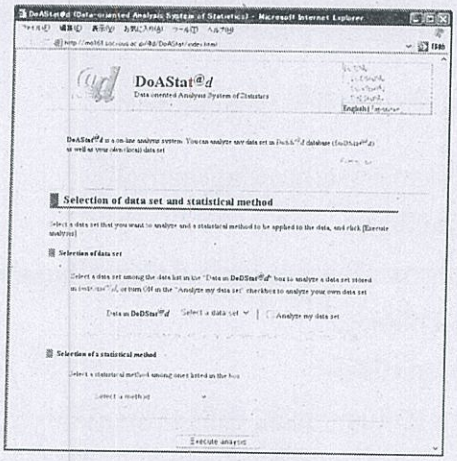

Figure 5. *Top page of DoAStat@d*

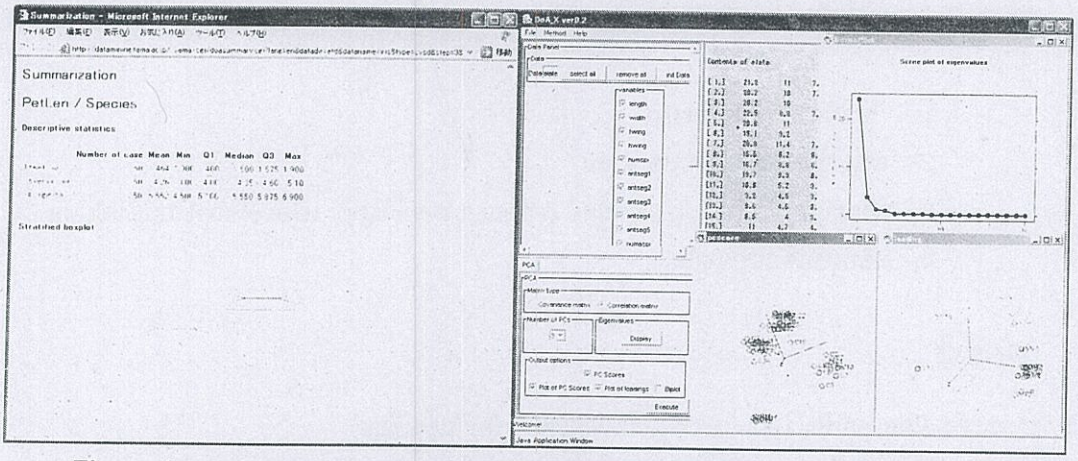

Figure 6. *DoA_R (results of summarization)* (Left) , *DoA_X (GUI and results for PCA)* (Right)

2.3 DoLStat@d (Data-oriented Learning System of Statistics)

DoLStat@d is a learning system, in which a variety of learning courses such as "Statistics introductory course" and "Economics course" are provided based on analysis stories stored in **DoDStat@d** according to the study target, and the computational functions in **DoAStat@d** is employed. Each course has a specific educational purpose and contains five to ten analysis stories arranged in a suitable educational order, according to the lesson's purpose. Using this system users are allowed to learn

- the analysis purpose and the practical analysis procedure (which are provided in the story XML page as in Figure 8),
- the original data (which is obtained from the data XML page as in Figure 4),
- how to execute the analysis based on the analysis story using online analysis system (the analysis can be executed by clicking [Analysis] button at the bottom of the story XML page as mentioned in section 2.2),
- how to use a general statistical package to analyze the data according to the analysis story ([SPSS], [R], [XploRe] and [Excel] buttons at the bottom of the story XML page link to explanation page illustrating how to use the corresponding package to analyze the data with macros or functions).

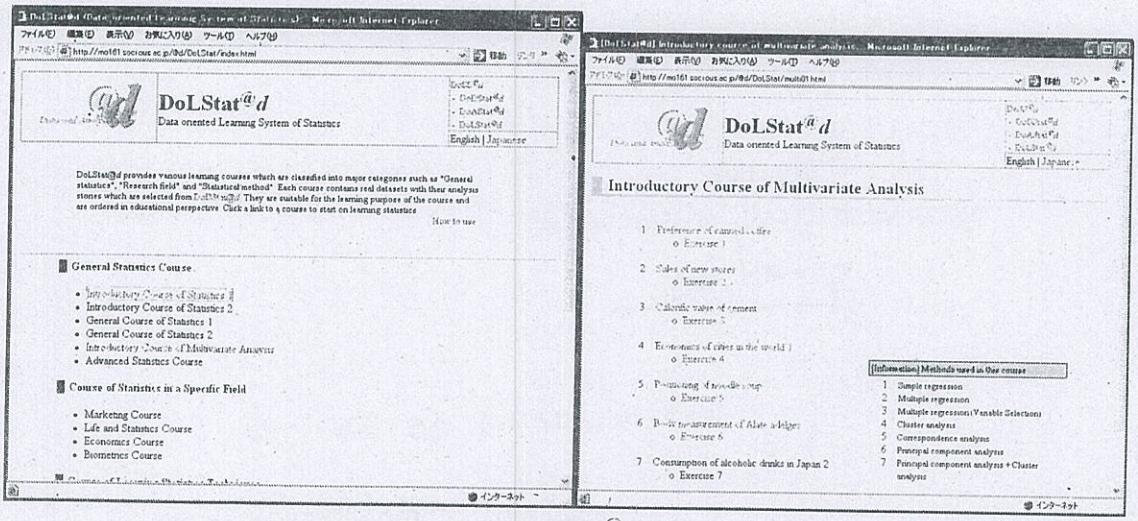

Figure 7. DoLStat@d

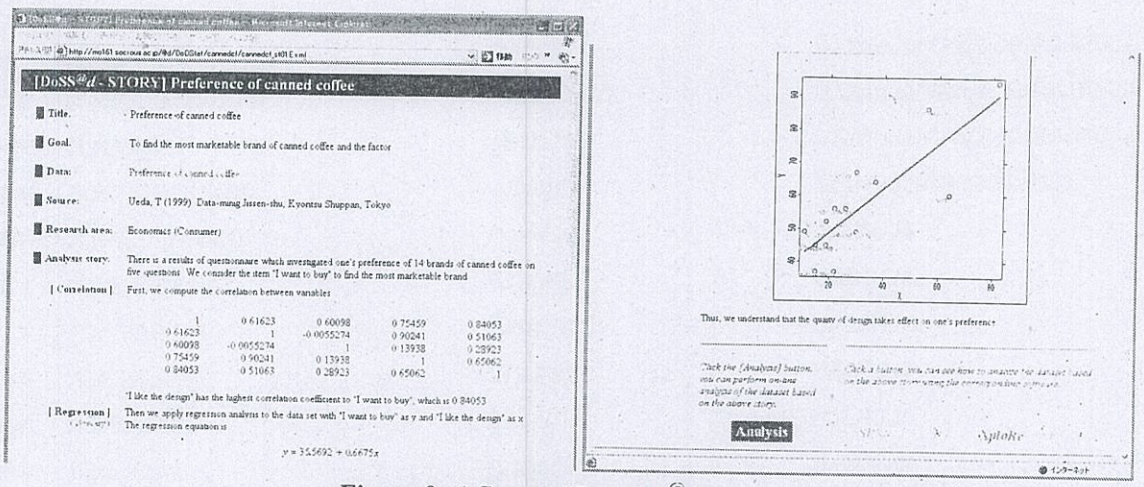

Figure 8. A Story of DoLStat@d

3. Concluding Remarks

DoSS@*d* is a web-based statistical (educational) system that combines real-world data and its analysis story with an online interactive analysis function. This system is constructed assuming to be used in the following cases: those who are concerned in statistical education find appropriate data sets in **DoDStat@*d*** to prepare their classes; those who are concerned in statistical education use courses in **DoLStat@*d***, make students use an online system in **DoAStat@*d*** and make them analyze data sets in **DoDStat@*d*** for exercise in the classes; students who learn statistics use data sets and analysis stories in **DoDStat@*d*** as self-teaching materials, study statistics according to the courses in **DoLStat@*d*** or use an online system **DoAStat@*d*** for their self-teaching; those who do data analysis use some data set in **DoDStat@*d*** to evaluate the statistical method used in their analysis.

There are a number of advantages in constructing a web-based system, such as universal accessibility, easy maintenance, and instant updates. This system was developed mainly to educate people in the area of statistical analysis using a data-oriented approach, however, method-based education also continues to provide an important role in statistical education.

Future plans are to provide a database system of analysis story, to cover more analysis methods, to create upload/remove system and evaluation system, and to collect more data sets and analysis stories.

ACKNOWLEDGEMENT

This project is supported by the Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (Grant#15300094, 2003-2005).

REFERENCES

Feuerhake, J. (2002). XQS/MD*Crypt as Means of Education and Computation. In Härdle, W. and Rönz, B. (eds), *COMPSTAT 2002 Proceedings in Computational Statistics*, 635-640, Phisica-Verlag, Heidelberg.

Honda, K., Yamamoto, Y., Yadohisa, H. and Mori, Y. (2004), Web-based analysis system in data oriented statistical system, COMPSTAT2004 Proceedings in Computational Statistics, 1209-1216.

Inoue, T., Asahi, Y., Yadohisa, H. and Yamamoto, Y. (2002). A statistical data representation system on the Web. *Computational Statistics*, 17(3): 47-63, Springer, Heidelberg.

Mori, Y., Fujino, T., Yamamoto, Y. and Tarumi, T. (2004). XML-based Applications in Statistical Analysis, Proceedings of Interface 2004: Computational Biology and Bioinformatics, 36th Symposium on the Interface.

Mori, Y., Honda, K., Yadohisa, H. and Yamamoto, Y. (2003a). Data-oriented Statistical System "DoSS@d" for supporting statisitical education. *Proceedings of 2003 ISM project "Highly-use of Information in Statistical Science"*, Jan. 22-24, 2004, 85-90. (in Japanese)

Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003b). Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). *Bulletin of the International Statistical Institute, 54th Session Proceedings, Volume LX Two Books*, Book 2, 74-77.

[1] Chance Database Home Page, `http://www.dartmouth.edu/`

[2] Data and Story Library (DASL), `http://lib.stat.cmu.edu/DASL/`

[3] Data Representation System (DRS), `http://www.sci.kagoshima-u.ac.jp/`

[4] MD*Base (Statistical methodology and interactive data analysis), `http://www.quantlet.org/mdbase/`

[5] Md*Tech. MD*Crypt, `http://www.md-crypt.com/`

[6] Statlib, `http://lib.stat.cmu.edu/`