



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

THE STATA JOURNAL

Editor

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnewton@stata-journal.com

Associate Editors

Christopher F. Baum
Boston College
Rino Bellocco
Karolinska Institutet, Sweden and
Univ. degli Studi di Milano-Bicocca, Italy
A. Colin Cameron
University of California–Davis
David Clayton
Cambridge Inst. for Medical Research
Mario A. Cleves
Univ. of Arkansas for Medical Sciences
William D. Dupont
Vanderbilt University
Charles Franklin
University of Wisconsin–Madison
Joanne M. Garrett
University of North Carolina
Allan Gregory
Queen's University
James Hardin
University of South Carolina
Ben Jann
ETH Zürich, Switzerland
Stephen Jenkins
University of Essex
Ulrich Kohler
WZB, Berlin

Stata Press Production Manager

Stata Press Copy Editor

Editor

Nicholas J. Cox
Department of Geography
Durham University
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com

Jens Lauritsen
Odense University Hospital
Stanley Lemeshow
Ohio State University
J. Scott Long
Indiana University
Thomas Lumley
University of Washington–Seattle
Roger Newson
Imperial College, London
Marcello Pagano
Harvard School of Public Health
Sophia Rabe-Hesketh
University of California–Berkeley
J. Patrick Royston
MRC Clinical Trials Unit, London
Philip Ryan
University of Adelaide
Mark E. Schaffer
Heriot-Watt University, Edinburgh
Jeroen Weesie
Utrecht University
Nicholas J. G. Winter
University of Virginia
Jeffrey Wooldridge
Michigan State University
Lisa Gilmore
Gabe Waggoner

Copyright Statement: The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publicly accessible web sites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

Stata and the WeeW information system

Pierpaolo Vittorini, Stefano Necozone, Ferdinando di Orio
Department of Internal Medicine and Public Health
University of L'Aquila
Piazz.le S. Tommasi, 1
67100 Coppito - L'Aquila, Italy
pierpaolo.vittorini@cc.univaq.it

Abstract. The need for timely collection and analysis of epidemiological data is becoming of primary importance, e.g., for bioterrorism detection or epidemiological surveillance. Web-based information systems (WISs) may provide the needed technological support. Thus we present the WeeW (workflow-enabled epidemiological WIS) system—i.e., a WIS that helps epidemiologists, through workflow management, to effectively select remote centers, collect, and process the received data to produce conclusive technical reports. In detail, we show the functionalities of the WeeW system, its architecture, and particularly Stata's role in executing statistical analyses and producing graphs. Furthermore, we discuss the performance of the WeeW–Stata interface. Finally, we outline short conclusions regarding the advantages and drawbacks connected with the proposed solution.

Keywords: pr0027, WeeW, web-based information systems, epidemiology, web development, benchmarks

1 Introduction

To support their strategies and decisions, public-health practitioners use information technology, statistical, and operational research methods (Raghupathi and Tan 2002; Hauck, Smith, and Goddard 2004; Eubank et al. 2004) usually applied to massive databases (e.g., Capocaccia et al. 2003).

Recently, the need for real-time data availability and for restricting the flow of unclear or notoriously wrong data has become critical, e.g., for bioterrorism detection or epidemiological surveillance (Bravata et al. 2004). In such a context, adopting a web-based information systems (WISs) technology might be useful (Deshpande and Hansen 2001; Armoni 2002), since a WIS supports the creation, integration, analysis, and distribution of information, particularly for management processes, entirely through the web.

Thus we developed the WeeW (workflow-enabled epidemiological WIS) system, which uses workflow management (Van Der Aalst and Van Hee 2001; Fischer 2005) to efficiently guide epidemiologists in organizing and executing their research through the following phases: remote center selection, data collection and processing, technical report generation, and discussion forum management. In particular, the WeeW system leans on Stata's powerful statistical and graphical capabilities to effectively process the collected data.

The article is organized as follows. Sections 2 and 3 present the WeeW workflow and architecture. Then sections 4 and 5 describe how WeeW incorporates Stata. Finally, section 6 concludes.

2 Workflow and architecture

WeeW uses workflow management (Van Der Aalst and Van Hee 2001; Fischer 2005). As known, workflow management consists of the automation of (usually business) procedures during which documents, information, or tasks are passed from one participant to another in a way that is governed by rules or procedures. The following are key benefits:

- Efficiency is improved, since automating processes eliminates unnecessary steps.
- A process is kept under control, so it can correctly reach its conclusion.
- Users are helped to take the right actions, thus reducing the possibility of errors.
- In concurrent applications, a workflow prevents discordant and/or potentially harmful actions.

Therefore, WeeW uses workflow management to implement “automated” epidemiological research processes that take place with accuracy and by also reducing the possibility of introducing errors.

Adopting workflows also limits the time needed to acquaint users with the system, since only those actions that are possible in a certain phase of the process are available.

The adopted workflow is highlighted in figure 1. The following three types of users interact with the workflow: epidemiologists, remote centers, and users. An epidemiologist using the WeeW system initially organizes his study by selecting the remote centers. Then he develops the questionnaire(s) needed for the data entry. The remote centers enter the data. Hence, the epidemiologist defines the statistical processing necessary to perform the investigation. By giving comments to the most interesting results, the epidemiologist allows the WeeW system to produce the technical report. Finally, users may discuss the available results by either reading the reports or participating in a discussion forum moderated by the epidemiologist.

(Continued on next page)

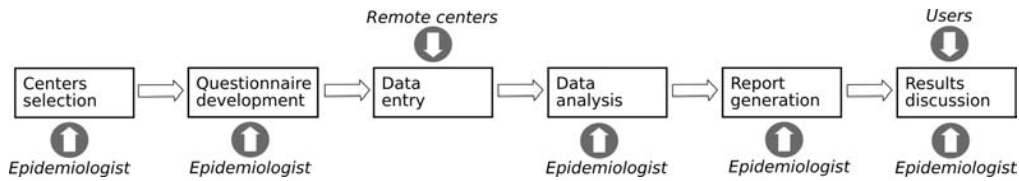


Figure 1: WeeW workflow

The WeeW system follows the application model depicted in figure 2. As shown, WeeW is accessible on the Internet through web-based interfaces. Thus WeeW is compatible with every operating system provided with an HTML-compliant web browser (including Windows, Macintosh, and Linux). WeeW uses a database to store all the data (Kofler 2005), relies on Stata to execute the statistical analyses, interfaces with the \LaTeX typesetting system to produce the technical reports (Lamport 1994), is made up of a collection of PHP scripts (Sklar 2004) to implement all its functionalities, and uses the Apache web server (Laurie and Laurie 2002) to communicate with clients.

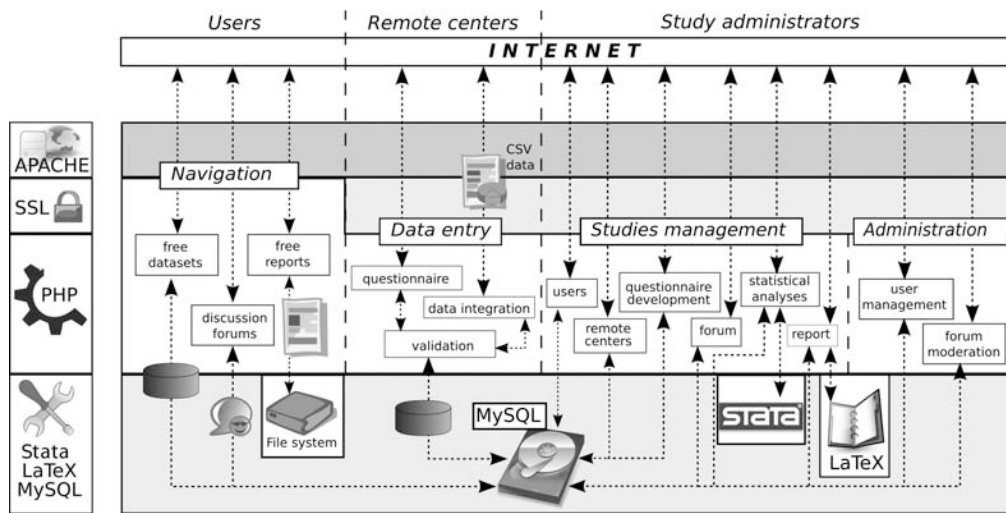


Figure 2: WeeW application model

The available functionalities can be divided into four sections: navigation, data entry, studies management, and administration.

The navigation section is opened to the users and offers the download of free datasets and reports and the chance to debate the results of the available investigations through a discussion forum. The free datasets are generated on the fly by a PHP script that extracts the proper tuples from the underlying database. The free reports are available as PDFs stored in the file system.

In the data entry section, the remote centers can fill in the questionnaire(s) and/or integrate into the underlying database any old dataset if available as CSV data. The system does not tolerate errors in incoming CSV files: data must be comma separated and formatted using English rules, and strings must be enclosed by double quotes. The entered data are validated in light of several constraints, defined by the epidemiologist, which regard

- whether a question must be considered a primary key;
- an allowed range for a numeric answer;
- the associations of integers with text;
- when a missing value can be accepted; and
- the definition of relationships between questions; i.e., a question can be placed if a certain statement is true. The WeeW interface currently allows only the development of statements in which a question is placed in connection with the value of a previous answer.

The studies management section allows an epidemiologist to organize her studies in the following areas: (i) users, (ii) centers, (iii) questionnaires, (iv) statistical analyses, (v) reports, and (vi) forums. The users area offers the possibility to involve other epidemiologists in the investigation. In the centers area, remote centers can be included or excluded from the data entry. The questionnaires are organized as a list of questions created through a visual interface and can be developed to honor the above-mentioned constraints. Furthermore, the following data types can be used:

- Free text (limited to 80 characters)
- Integer values (4 bytes long)
- Floating values (8 bytes long)
- Sets of values (stored as integers)

To process the data, the epidemiologist defines a list of analyses by selecting the desired one among a set of predefined statistical analyses and graphs, by writing a Stata command, or by creating new variables. Then Stata, provided with a dataset and a do-file containing the needed commands, is invoked in the background. As output, Stata produces (i) a log file with the results of the analyses and (ii) a collection of EPS files containing the graphs. The collected data can also be exported in the CSV format—readable by most programs like Excel, Epi Info, Stata, and SAS—to be analyzed with external software. We will discuss this interface in detail in [section 4](#).

The technical reports are created from a list of comments regarding the most significant results: the WeeW system automatically “assembles” all the given comments or results, creates a `.tex` file, then invokes \LaTeX to produce the PDF, and finally displays such a file as the resulting technical report. A report contains (i) a title page, (ii) a page

with a note and the list of study participants, (iii) the table of contents, (iv) a chapter with a description of the epidemiological study and a note preceding the results, (v) a central chapter with all comments and results regarding the analysis, (vi) the description of the questionnaire and of the dataset, (vii) the list of tables, and (viii) the list of figures.

Furthermore, in the administration section, a discussion forum regarding the results of a study can be opened and moderated. Finally, a user-management facility is also provided.

Figure 3 briefly depicts (i) the user interface used to build up the list of the statistical analyses and graphs, (ii) the user interface used to add the comments, and (iii) a sample page of the produced report.

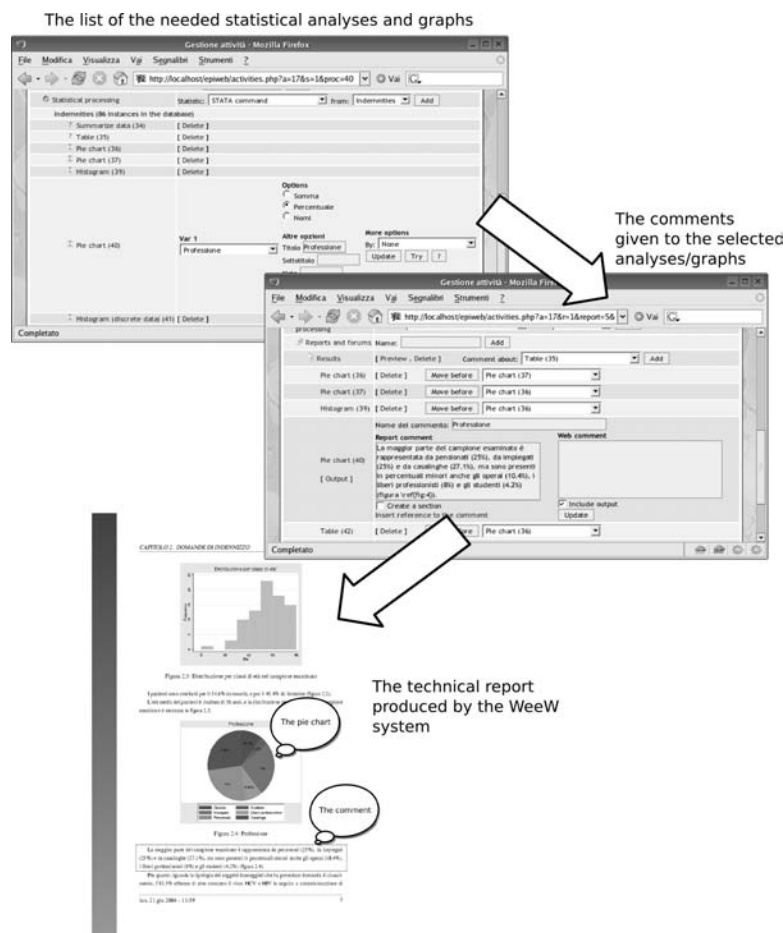


Figure 3: Analysis interface, comment interface, and sample report

3 Database organization

Figure 4(a) depicts the portion of the E/R diagram showing how the WeeW system organizes its datasets. A dataset comes from a **questionnaire**, whose **name** is stored in the database. A **questionnaire** is made up of a list of **questions**, which are organized as a double-linked list, where the **next** and **prev** attributes point to the next and previous questions, respectively. The **type** attribute indicates the data type of the expected answer and the **opt** attribute stores its possible options (for the sets of values data types, the same syntax of the **label define** Stata command is used, e.g., 0 “F” 1 “M”). The **null** attribute indicates whether a missing value might be accepted, and the **text** attribute contains the text of the question. The **ic** field contains a Boolean statement that asserts whether this question has to be asked. To a certain question, an **answer** is given; the value is stored in the proper attribute (**int**, **float**, or **str80**). The flag **missing** represents a missing value.

Figure 4(b) depicts the portion of the E/R diagram showing the organization of the statistical analyses/graphs. A statistical **processing** is the actual application of a certain **model** in a given analysis. Therefore, a **model** has a **name**; a **type**, which specifies whether it is a statistical analysis or a graph; and a **stata_cmds** attribute, which contains its abstract Stata implementation. A model becomes “tangible” by means of the proper **varlist**, **by** prefix, **if** qualifier, and **opt** options, stored in the processing entity. Hence, the models represent all the available statistical analyses and graphs, whereas the processings are the actual analyses executed by an epidemiologist in her investigation.

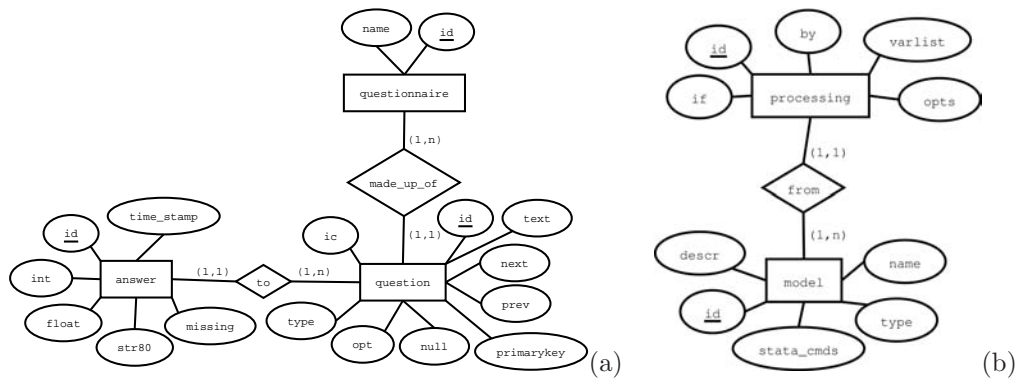
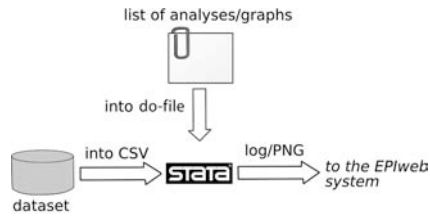


Figure 4: E/R diagram relative to organization of (a) dataset and (b) analyses/graphs

4 The WeeW–Stata interface

Figure 5 highlights the four steps taken by WeeW to interface with Stata. The description of the corresponding detailed approach, shown in algorithm 1, follows.



1. The dataset is converted into a CSV file.
2. The chosen analyses/graphs are transformed into a do-file.
3. They are both passed to Stata.
4. The results are “captured” by WeeW.

Figure 5: Stata and the WeeW system

Step (1) The CSV file is created by extracting from the database the answers given to the questions composing the selected questionnaire [lines 1–6, see E/R diagram in figure 4(a)].

Step (2) The proper do-file is created as follows. Initially, a preamble is added, made up of an `insheet` command, which loads the CSV file; a set of `recast` commands; and a set of `label` commands, which improves the readability of the output. Furthermore, for each statistical processing mode chosen by the epidemiologist, the respective model is selected. Then a Stata command is generated by instantiating the `stata_cmds` attribute with the proper *varlist*, *by* prefix, *if* qualifier, and `opts` options [see E/R diagram in figure 4(b)]. If the model is a graph, a further `graph export` command is added (lines 7–13).

Step (3) Stata is launched to execute the do-file (line 14).

Step (4) Each line of the produced log file is parsed. If the parsed line suggests that the output is a graph, the external image generated through the added `graph export` command is stored in the database. Otherwise, only the textual output is stored (lines 15–21).

Algorithm 1 Procedure used to interface with Stata

```

    {create the weew.csv file}
1: for each question  $q$  do
2:   for each  $i$ th answer  $a$  do
3:     CSV[ $i$ ][ $q$ ]  $\leftarrow a$ 
4:   end for
5: end for
6: store the CSV matrix in the weew.csv file
   {create the weew.do file}
7: add the proper insheet, recast, and label commands
8: for each defined statistical processing do
9:   insert the proper Stata command into the do-file
10:  if command is a graph then
11:    add the proper graph export command
12:  end if
13: end for
   {launch Stata with stata -q -b do weew.do}
14: launch Stata in batch mode
   {parse the log file to acquire the output}
15: for each line of the weew.log file do
16:  if command is a graph export then
17:    store the external image
18:  else
19:    store the line
20:  end if
21: end for

```

Extracting observations from the database, converting them into CSV, and then importing such a file into Stata legitimately raises questions about the correctness of the processed dataset. The processed dataset's correctness is ensured because (i) variables are stored in the database with the same storage type of the respective "Stata equivalent", (ii) numerical variables (either integer or real numbers) are converted into strings (those making up the CSV file) by using the maximum available resolution, and (iii) CSV files are imported through the **insheet** command.

Metadata and labels issues are taken into account as follows. Each variable is "recast" as a **str80** if stored in the database as a string, as a **long** if stored as an integer or a set of values, and as a **double** if stored as a float. The labels are initially set to the questions' texts and eventually modified by the epidemiologist. Finally, for each "set of values" variable, two further label commands are added: **label define** to associate integers with text and **label value** to attach the value label to the variable.

In algorithm 2, part of a do-file created by the WeeW system is shown. This file produces a histogram of the age range of students in an investigation of middle school students' nutritional habits in the Municipality of L'Aquila (Cesare et al. 2005).

Algorithm 2 Part of a do-file created by the WeeW system

```

insheet using "/var/www/html/weew/tmp/csvd3lwoU.csv", delimiter(";")
...
label var d4 "How old are you ?"
...
recast long d4, force
...
histogram d4, discrete frequency title(Distribuzione del campione per eta')
    xtitle(Eta') ytitle(Frequenze) addl xlabel() saving(56, asis replace)
graph display, xsize(10) ysize(8)
graph export 56.eps, replace

```

5 Benchmarks

This section evaluates the WeeW–Stata interface performances¹ in relation to the following dimensions: (i) the dataset size, S_D ; (ii) the number of graphs, N_G ; and (iii) the number of statistical analyses, N_S . The performances are the overall execution time, T_{TOT} ; the CSV creation time, T_{CSV} ; the do-file generation time, T_{do} ; the Stata execution time, T_{Stata} ; and the log file parsing time, T_{log} .

To evaluate the WeeW performances, we proceeded in the following experiment. We initially selected an epidemiological investigation conducted throughout the WeeW system (Cesare et al. 2005). The investigation is made up of the following:

- $S_D = 171$ kb (318 observations, each ≈ 0.54 kb and made up of 67 variables)
- $N_G = 12$ graphs
- $N_S = 10$ statistical analyses

Then we randomly extracted 200 “subsets” of the investigation, i.e., 200 triples:

$$\langle s_D, n_G, n_S \rangle, \text{ where } \begin{aligned} 0 &\leq s_D \leq S_D (=171 \text{ kb}) \\ 0 &\leq n_G \leq N_G (=12) \\ 0 &\leq n_S \leq N_S (=10) \end{aligned}$$

Since WeeW runs in a multitasking environment, several unwanted delays might be introduced by concurrently executing processes not directly related to WeeW. Therefore, for each subset, all the above-mentioned performances were computed 10 times, and the corresponding median (which is not affected by the farthest values) is reported as the actual measure.

The resulting dataset was analyzed as follows. For each performance, the univariate Spearman’ ρ correlation was computed versus S_D , N_G , and N_S . For each independent variable with $\rho > 0.5$, a multivariate linear regression model was carried out, and the

1. The performances were computed on an Intel Pentium 4 3-GHz CPU, 512 MB of RAM, SATA 80-GB hard drive. OS Linux 2.6.10. Apache 2.0.51. MySQL 3.23.58. PHP 4.3.10. L^AT_EX 2.0.2. Intercooled Stata for Linux 8.2.

respective coefficients were reported. As shown in table 1, the T_{CSV} performance is affected only by the S_D dimension and the T_{do} performance is affected by none of the considered dimensions, whereas the T_{Stata} , T_{log} , and T_{TOT} performances are affected by the number of analyses/graphs.

Table 1: Results of the experiment

	S_D		N_G		N_S	
	ρ	Coeff.	ρ	Coeff.	ρ	Coeff.
T_{CSV}	.9189	.0042858	.0063		-.0210	
T_{do}	.0969		.4742		.4298	
T_{Stata}	.0710		.7590	.7534710	.6852	.7418854
T_{log}	.0197		.7626	.1356075	.6892	.1336584
T_{TOT}	.0612		.7588	.8894704	.6871	.8784634

Figure 6 shows the linear regression graphs of the estimated performances in connection with S_D , N_G , and N_S . As shown, the Stata execution is (not surprisingly) the most expensive task, followed by the log parsing procedure. In numbers, the Stata execution takes from 82.7% to 98.7% of the total execution time, whereas the log parsing procedure takes from 1.2% to 17.5% of the total execution time. Consequently, we argue that, because of the log parsing procedure, the WeeW–Stata interface does not introduce a remarkable overload.

(Continued on next page)

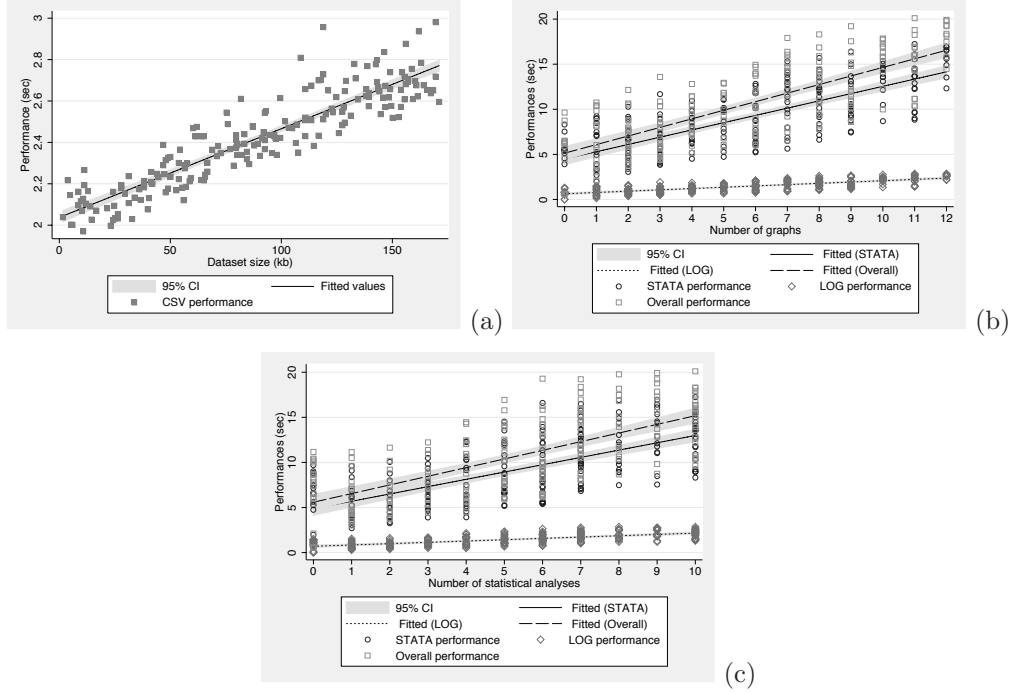


Figure 6: Estimated performances in connection with (a) S_D , (b) N_G , and (c) N_S

6 Conclusions

The WeeW system offers many advantages. In multicenter studies, developing ad hoc information systems usually absorbs part of the given funding and produces software tools that cannot be reused for other investigations. WeeW is instead a versatile system and therefore can be tailored to the specificity of (almost) every epidemiological study. Furthermore, the data are promptly sent to the main center in a digital format, thus reducing the efforts associated with data collection.

In particular, adopting Stata within the WeeW system reassures epidemiologists that they can rely on a powerful and reliable engine that executes both the statistical analyses and graphs. Furthermore, Stata (and the \LaTeX system) acts like a software component reused in a wider project, thus entailing a large reduction in development efforts (Mili, Mili, Yacoub, and Addy 2002). Moreover, the proposed architecture also decreases the deployment costs, i.e., the costs connected with the activities after the initial release. For instance, a software update, which must be applied only on the server, immediately becomes available to clients, too.

The report generation engine takes into account internationalization issues (i18n) as follows. WeeW stores all the texts by using the UTF-8 character set, and a user may select the report's output language. This choice is then used by the report generation

engine, as an option of the `babel` package, used in the `.tex` file to typeset the resulting PDF. The UTF-8 character set caused some characters to be displayed incorrectly in Stata (which instead uses the ISO-8859-1 character set).

The aforementioned Cesare et al. (2005) study regarding the middle school students' nutritional habits was also used to ask epidemiologists to rate the WeeW system. The following comments were reported:

- The workflow was not fully “perceived”, e.g., sometimes the next step was not clear.
- The data collection process was strongly improved (in both timeliness and efficacy).
- The “stereotyped” organization of the reports was sometimes considered limiting, mostly because of the lack of a word processor-like approach. However, the WeeW system follows a \LaTeX -like approach, thus letting the author focus on the logical flow of content rather than document typesetting. However, since reports are made up of a sequential combination of comments (eventually divided into sections), the possibility of freely structuring the text is reduced.

Therefore, the next release of the WeeW system will provide an improved user interface clearly showing the workflow and will allow epidemiologists to develop more flexible reports.

Broader adoption of the WeeW system is ongoing, currently to investigate colon tumors (led by the Surgical Endoscopy Operational Unit of the Hospital of L'Aquila) and as the supporting infrastructure for epidemiological studies in Lezhe (as part of an EU-funded project, called the Albania project).

For benchmarks, we pointed out that—except for the Stata execution, which cannot be controlled by the WeeW developers—the log parsing procedure is the most expensive task. WeeW is forced to concurrently access the file system to combine the log (containing the output of the analyses) and the EPS files (containing the graphs) to reconstruct the complete output. This sharp separation induces a question regarding the need for Stata to adopt a comprehensive and exhaustive log file format that can embrace both textual and graphical output. The XML-based ODT format (OASIS 2005) could be a reasonable choice, which would also let Stata take a further step, similar to that recently taken in Stata 9 for data management, toward XML.

(Continued on next page)

7 Appendix

To run, WeeW needs the following:

- Linux or Windows as the operating system
- Apache or IIS as the web server
- PHP as the HTML-preprocessing engine
- MySQL as the database
- Stata as the statistical engine

WeeW is free software licensed under GPL (<http://www.gnu.org/licenses/gpl.html>). MySQL, PHP, and Apache are also free software, whereas Stata is commercial software. You may request a copy of WeeW from the authors via email.

Since WeeW runs behind an Apache + PHP + MySQL server, a web administrator is qualified to set up and maintain the system. The benchmarks reported in section 5 should be used to estimate, in terms of the expected dataset size and analyses, the required performances and therefore to identify the characteristics of the server that will have to host the WeeW system. However, to support a medium-sized study (i.e., thousands of questionnaires coming from about 10 remote centers) an entry-level server suffices.

8 References

- Armoni, A., ed. 2002. *Effective Healthcare Information Systems*. Hershey, PA: IIRM Press.
- Bravata, D. M., K. M. McDonald, W. M. Smith, C. Rydzak, H. Szeto, D. L. Buckeridge, C. Haberland, and D. K. Owens. 2004. Systematic review: Surveillance systems for early detection of bioterrorism-related diseases. *Annals of Internal Medicine* 140: 910–922.
- Capocaccia, R., G. Gatta, P. Roazzi, E. Carrani, M. Santaquilani, R. De Angelis, A. Tavilla, and the EURO CARE Working Group. 2003. The EURO CARE-3 database: Methodology of data collection, standardisation, quality control and statistical analysis. *Annals of Oncology* 14 (Suppl. 5): 14–27.
- Cesare, B., P. Vittorini, E. Sallusti, V. Bontempo, M. Graziani, S. Necozone, and F. di Orio. 2005. Le abitudini alimentari dei ragazzi di scuola media nel Comune di L'Aquila: Risultati di uno studio descrittivo. In *IX Conferenza Nazionale di Sanità Pubblica*. In Italian.
- Deshpande, Y., and S. Hansen. 2001. Web engineering: Creating a discipline among disciplines. *IEEE Multimedia* 8: 82–87.

- Eubank, S., H. Guclu, A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczka, and N. Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Science* 429: 180–184.
- Fischer, L., ed. 2005. *Workflow Handbook 2005*. Lighthouse Point, FL: Future Strategies.
- Hauck, K., P. C. Smith, and M. Goddard. 2004. The economics of priority setting for health care: A literature review. Washington, DC: The World Bank, Discussion paper no. 28878.
- Kofler, M. 2005. *The Definitive Guide to MySQL 5*. 3rd ed. Berkeley, CA: Apress.
- Lamport, L. 1994. *TEX: A Document Preparation System*. 2nd ed. Reading, MA: Addison–Wesley.
- Laurie, B., and P. Laurie. 2002. *Apache: The Definitive Guide*. 3rd ed. Sebastopol, CA: O'Reilly.
- Mili, H., A. Mili, S. Yacoub, and E. Addy. 2002. *Reuse-Based Software Engineering: Techniques, Organizations, and Measurement*. New York: Wiley.
- OASIS. 2005. Open Document Format for Office Applications (OpenDocument) 1.0. Online resource. <http://www.oasis-open.org/committees/office..>
- Raghupathi, W., and J. Tan. 2002. Strategic IT applications in health care. *Communications of the ACM* 45: 56–61.
- Sklar, D. 2004. *Learning PHP 5*. Sebastopol, CA: O'Reilly.
- Van Der Aalst, W., and K. Van Hee. 2001. *Workflow Management: Models, Methods, and Systems*. Cambridge, MA: MIT Press.

About the authors

The authors are affiliated with the Faculty of Medicine and Surgery at the University of L'Aquila, Coppito, L'Aquila, Italy.

Pierpaolo Vittorini is a research professor in computer science whose research interests include information systems for public health and epidemiology, statistical databases, and XML data management.

Stefano Necozone is an associate professor of hygiene and public health whose research interests include clinical epidemiology and evidence-based medicine.

Ferdinando di Orio is a full professor of hygiene and public health whose research interests include public health and epidemiology.