

Learning Statistics based on the Compendium and Reproducible Computing

Patrick Wessa *

Abstract—This paper explains how the R Framework (wessa.net) and a newly developed Compendium Platform (freestatistics.org) allow us to create, use, and maintain documents that contain empirical research results which can be recomputed and reused in derived work. It is illustrated that this technological innovation can be used to create educational applications that can be shown to support effective learning of statistics and associated analytical skills. It is explained how a Compendium can be created by anyone, without the need to understand the technicalities of scientific word processing (L^AT_EX) or statistical computing (R code). The proposed system allows educational researchers to objectively measure key aspects of the actual learning process based on individual and constructivist activities such as: peer review, collaboration in research, computational experimentation, etc. The system was implemented and tested in three statistics courses in which the use of Compendia was used to create an interactive e-learning environment that simulated the real-world process of empirical scientific research.

Keywords: *compendium, reproducible computing, technology, statistics education*

1 Introduction

The problem of irreproducible research received a great deal of attention within the academic community [2], [10], [12], [6], [5], [7], [3]. The most famous citation about the problem of irreproducible research is called Claerbout's principle: *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures.* [2]. Interestingly, Jan de Leeuw adds the following comments: *First, there is no reason to single out figures. The same "Principle" obviously applies to tables, standard errors, and so on. The fact that figures often happen to be easier to reproduce, does not preclude that we should apply the same rule to any form of computer-generated output. Second, there is no reason to limit the Claerbouts Principle to published articles. We can make exactly the*

same statement about our lectures and teaching, certainly in the context of graduate teaching. We must be able to give our students our code and our graphics files, so that they can display and study them on their own computers (and not only on our workstations, or in crowded university labs). And third, and perhaps most importantly, it is not clearly defined what a "software environment" is. Buckheit and Donoho apply the principle in such a way that everybody who wants to check their results is forced to buy MatLab®. Not Mathematica®, Macsyma®, or S-plus®. Those you may need to buy for other articles. This violates the Freeware Principle... [2].

Several solutions have been proposed ([5], [3], [8]) - most prominently there is an R package [11] called "Sweave" ([8]) where the concept of the so-called Compendium plays a fundamental role. The Compendium is defined as an integrated collection of text, code, and data (that allows the presented science to be reproduced). All the necessary documents that are needed to create the article (with embedded R code) and the data are contained in an archive file (preferably in tar.gz or zip format). A few examples of such Compendia can be downloaded from [1].

While the so-called Sweave package [8] is an excellent tool, there are a few reasons why the Sweave solution may be rather impractical for educational use:

- The reader is required to download, install, and execute software on their client machine which may lead to a variety of technical issues related to compatibility and security;
- The reader is assumed to have a working knowledge of L^AT_EX and R code;
- If the reader reuses the archived computation in derived work then there is no easy way to ensure dissemination (a new Compendium must be created first and sent to all users who might be potentially interested);
- The use of Compendia cannot be measured or monitored which implies that we cannot research the related educational processes.

*Affiliation: K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium. This research was funded by the OOF 2007/13 project of the K.U.Leuven Association. Email: patrick@wessa.net

2 A new solution

2.1 The concept

In order to address the previously mentioned issues we need to redefine the Compendium as: a research document where each computation is referenced by a unique URL that points to an object that contains all the information that is necessary to recompute it. This definition differs substantially from the original one while achieving the same goal. The key difference is the fact that in my proposal there is a complete separation between text and computing - Figure 1 provides a conceptual overview. The computations (incl. data, parameters, and software) are stored in a central, web-based repository that can be referenced from any text. In addition, there is a complete separation between the archived computation (which is described with various types of meta data) and the actual computing infrastructure (so-called R Framework). An example of such a Compendium can be found at [18]. In the next subsection it is explained how such a Compendium can be created based on the proposed solution. The computations that I included in this article to illustrate this are reproducible/reusable. According to my definition, this article is a Compendium.

2.2 Creating the Compendium

The R Framework allows educators and scientists to develop new, tailor-made statistical software (so-called R modules which are based on the R language [11]) within the context of an open-access business model that allows us to create, disseminate, and maintain software modules efficiently and with a very low cost in terms of computing resources and maintenance efforts [17]. Educators may browse the collection of available R modules at the R Framework's website (www.wessa.net) and use them for the purpose of generating research output that can be used to explain statistical techniques and concepts in their courses. If the educator has a working knowledge of the R language then new R modules can be created and published.

There are some unique features of the Compendium Platform that are of particular importance in the e-learning environment that is proposed:

- any computation that is created within the R Framework can be easily archived in the repository there is no need for students to keep track of the data, the model parameters, or the underlying statistical software code;
- any user who visits the unique URL of an archived computation is able to instantly reproduce the computation or reuse it for further analysis only an internet browser (and an active connection) is required to use the repository;

- educators and researchers are able to retrieve data for research purposes.

Suppose that the educator wants to include a histogram about some data set of interest into a course. There are only few steps needed to achieve this. In a first step the histogram is computed with the appropriate R module [19]. The data series of interest (Gold prices in Brussels) are pasted into the submission form of the R module and the computational parameters specified. When the user hits the Compute button a series of events are triggered that cause one of the dedicated R servers to compute and return a result page (more details can be found in [17]). In the result page there are three types of information available: a summary of the computational transaction, an HTML-based table with various statistics about the frequency table, and a picture which is available in png and postscript format (Figure 2).

The computational transaction table contains two hyperlinks that refer to the actual R code that was sent to the R server (raw input), and the R server's reply (raw output). This provides the user with full transparency of what is going on behind the scenes, while non-experts (like most students) can safely ignore this information. If the R interpreter was unable to execute the computational request, the error message (produced by R) is displayed in the transaction table.

In a second step the educator needs to insert the picture of the histogram into the course. In normal circumstances, one would just save the png (or postscript) picture to the local hard drive and insert/import it into the text processing environment. With the new approach however, the educator first creates a permanent archive of the computation by clicking the appropriate hyperlink which is printed above the HTML table and below the picture. A simple submission form allows the user to specify the following optional attributes: a meaningful title, keywords, comments, an e-mail address, a private/public flag, and a moratorium date. With the last two attributes the user is able to specify if (and how long) a computation is hidden from the public eye.

When the submission is completed, a series of events are triggered. All the meta data about the computation is packaged into a serialized string and stored in the web server's local cache. A remote procedure which resides on the Compendium Platform server is invoked and the stored computation is fetched through an HTTP GET callback ("Blog" in Figure 1). Once the computation has been transmitted, the Compendium Platform executes the following tasks: check if the computation is complete, store the serialized string as a physical file (as backup), unserialize the string, parse various attributes to obtain keywords that are indexed, store the computation in the repository database. When everything is completed, a result page is sent to the end-user with a hyperlink of the

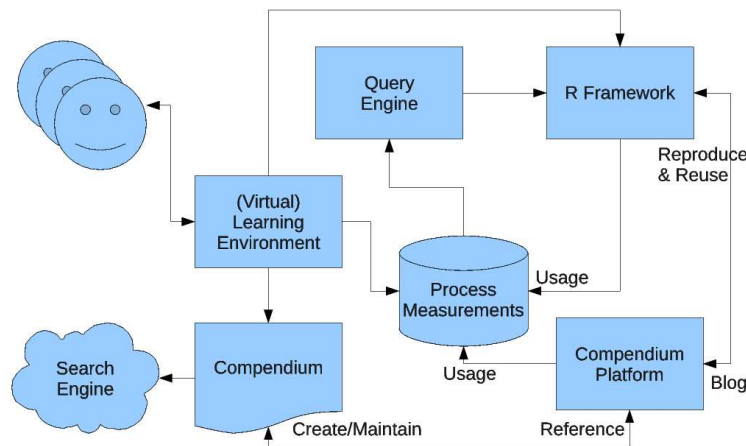


Figure 1: Reproducible Computing in Education

archived computation.

The educator can visit the web page that displays a snapshot of the archived computation and find all the information that is needed to appropriately reference the histogram in a course. The bibliographical reference of this particular computation has been included in references of this paper [13]. In addition, the hyperlink of the web page can be included in the text processing environment. If the educator uses a text processor that is similar to Microsoft® Word or OpenOffice Writer, then the hyperlink can be simply pasted. If a L^AT_EX editor is used then the following code can be used to invoke the package that allows one to create clickable hyperlinks: `\usepackage[dvips, bookmarks, colorlinks=false]{hyperref}`. The actual code that inserts the picture and includes the hyperlink is simply the following:

Listing 1: Including the Histogram in L^AT_EX

```
\begin{figure*}
\centerline{
\mbox{\includegraphics[width=4.00in]{
  histogram.ps}}}
\caption{Histogram of Gold prices (in
  Brussels) }
\centerline{\htmladdnormallink{www.
  freeststatistics.org/blog/date/2008/Jun
  /20/t1213979586y9h8zd8mwz5x7br.htm}{
  http://www.freeststatistics.org/blog/
  date/2008/Jun/20/t1213979586y9h8zd8
  mwz5x7br.htm}}
\label{histogram}
\end{figure*}
```

The above code snippet illustrates how the reproducible histogram (shown in Figure 2) was included in this paper (“Create/Maintain” in Figure 1).

2.3 Reading the Compendium

Anyone who reads the Compendium can click (or copy & paste) the link that is available below Figure 2. This allows the reader (student) to examine all the details of the computation and to challenge the results. For instance, the student may wonder if the gold prices are really bimodal - the conclusion that is drawn from the histogram may be distorted by the choice of the number of bins.

In order to investigate this, the student clicks on the Reproduce button of the web page that displays the snapshot of the original computation. The Compendium Platform retrieves the meta data that define the computation of the histogram and transmits it to the R Framework through a HTTP POST request (“Reproduce & Reuse” in Figure 1). This causes the R Framework to (exactly) recompute the histogram. The result page can be examined by the reader (student) and reused in a new computation (the Kernel Density Estimation module which can be found in the Descriptive Statistics section of the website). The result of this new analysis leads to the unambiguous conclusion that the data are (indeed) bimodal [14]. Now the reader has the option to archive this result and use it in a paper or for the purpose of communicating with others. The computation of the Kernel Density Estimates can be reproduced by using this webpage: <http://www.freeststatistics.org/blog/date/2008/Jun/22/t1214129827ny0eqihaz22u1x8.htm>.

3 Applications

There are various applications that can be built with the proposed system. I briefly discuss a few examples that have been tested in the recent past and which have been (according to several evaluation surveys) highly appreciated by students.

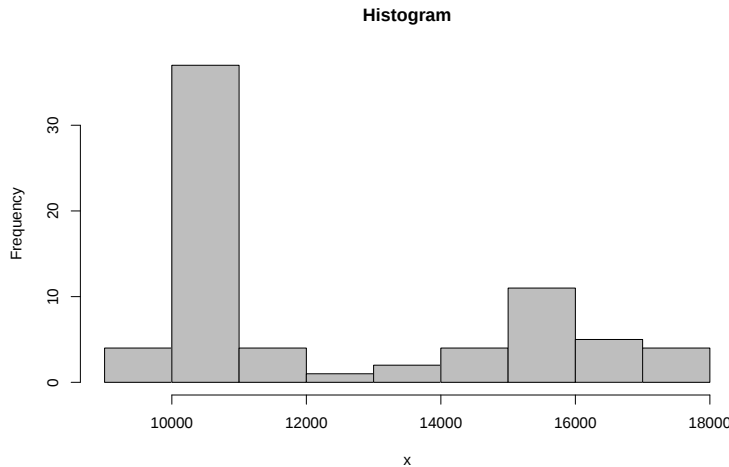


Figure 2: Histogram of Gold prices (in Brussels)
<http://www.freeststatistics.org/blog/date/2008/Jun/20/t1213979586y9h8zd8mwz5x7br.htm>

3.1 Statistics education

It is obvious that reproducible computing and the compendium may be used in various ways to support statistics learning. One approach can be found in [20] where it is explained how an e-learning environment can be built that supports non-rote learning of exploratory and inductive statistics within the paradigm of social constructivism. The key idea behind this constructivist application is that students are empowered to interact with reproducible computations from peers and the educator. Students are required to play the role of an active scientist who investigates problems, presents solutions, and reviews the work of peers. Obviously, reproducible computing is a *conditio sine qua non* that allows students to engage in such peer review activities.

Another approach is to create Compendia based on real-life examples that can be used as case studies. Once they have been created, the educator does not have to worry about the availability of the software and the data. For instance, the Compendium in [18] allows students to learn about descriptive statistics and the linear regression model. This type of Compendium can be used in self-study or in interactive courses where students are involved in hands-on sessions (in small groups).

Yet another way to make good use of the system is when the educator provides examples (in a Compendium course text) while students are asked to write an assignment paper in the form of a Compendium. This allows the educator to monitor the learning process (as explained in section 3.2), and to continuously make improvements through the feedback obtained from students about conceptual and technical problems that might otherwise remain undetected. Let me briefly illustrate this feedback mechanism in order to explain its importance and how it

affects quality.

In two undergraduate statistics courses I introduced univariate time series models (such as ARIMA). Most students perceive the specification of an ARIMA model to be (very) difficult. The problem is that students have to decide which model parameters should be used based on a series of so-called diagnostic tools (Autocorrelation Function, Partial Autocorrelation Function, Spectral Analysis, etc...). In order to make this process easier, I created an R module that estimates a general (possibly over-parametrized) model and applies a parameter reduction technique which is presented by a (convenient) visual chart. In spite of all my efforts to carefully prepare the R module, some students encountered a few problems which I had not anticipated. Fortunately, they were instructed to send me the link of an archived computation as soon as they ran into trouble with using any of the R modules. One example of such a problem can be found in [15] and was encountered by a student who archived the computation in the Compendium repository. When I received the e-mail with the link of the troublesome computation I could easily look at the archived snapshot and detect the source of the problem in the raw output of the R engine. It took me almost no time to reproduce the computation, change the parameters to appropriate settings, and send the student a reply with a solution [16]. Much to my surprise, such problems often have nothing to do with the actual ARIMA model or the estimation algorithm. One example of such a problem occurs when students specify a logarithmic transformation for a time series with negative numbers! As a consequence, this feedback mechanism has gained me a lot of insight into the types of problems students run into - therefore, I can now improve my lectures, course texts, and software in order to help students avoid them.

3.2 Educational research

If the system is configured in such a way that students are required to log on to their account in the Virtual Learning Environment, it is possible to measure all the actions that occur within the system that is portrayed in Figure 1 (the central part of the figure shows that the measurements of all major components are stored in the Process Measurements Database). This allows us to accurately measure activities (that are otherwise unobservable) and attribute them to the individual student: usage of the R-based software modules, creation of archived computations, usage and reuse of archived computations, grading and feedback of peer reviews, collaboration among students, etc. Educators are empowered to query the database for the purpose of improving the learning experiences of students. In addition, educational researchers can engage in data mining, based on the R Framework.

For example, in [20] two statistical hypotheses are tested that show a very strong relationship between the number of submitted peer review messages and objective examination scores, in which the questions assess conceptual understanding rather than rote memorization. This has important consequences for our understanding of the role and importance of the often used method of peer assessment: the study suggests that the focus should be on the review of peer's work based on verbalization, not the actual grading.

Another example of how objective measurements can improve our understanding of the educational processes is provided in [21]: a comparison between actual and reported data, demonstrates that reported learning process measurements are highly misleading in various and unexpected ways. However, reproducible computing and objective measurements of actual learning behavior, reveal important guidelines that allow us to improve the effectiveness of learning through cognitive, behavioral, and constructivist incentives.

It is important to note that the infrastructure (incl. R Framework and Compendium Platform) is available free of charge for the purpose of academic research. Anyone, with a sincere interest in educational research based on reproducible computing, is invited to make contact.

4 Conclusions and Future Work

The solution that I propose is based on a new system, that satisfies all of the requirements that were formulated by Claerbout and de Leeuw in section1:

- Every reader of a Compendium is able to reproduce and reuse all computations that reference computing objects that have been stored in the Repository of the Compendium Platform;

- All computations can be reproduced and reused - not just the figures;
- A Compendium can be created for articles, presentations (slides), course texts, handbooks, websites, etc...;
- There is no need to download or install anything on the client machine - all computations are performed on a network of dedicated servers and the interpreted source code is free.

In addition, it can be concluded that the proposed solution allows us to create applications that are beneficial for statistics learning and that can be used to engage in educational research based on measurements that are otherwise unobservable. In future, this infrastructure will probably lead to a series of investigations where the effectiveness of statistics learning can be explained/predicted by co-factors that are related to: software usability, pedagogical approaches (constructivism, behaviorism, etc...), collaboration between students, mathematical versus intuitive instruction, measures to prevent fraud and free riding, etc...

The approach that is proposed in this paper may have fundamental consequences for statistics education and educational research. In addition, we are now able to create, implement, and maintain virtual learning environments that are increasingly effective in terms of learning outcomes, and consistently efficient in terms of the effort that is required to improve course materials and computational infrastructure.

References

- [1] "Publications - bioconductor.org", URL <http://www.bioconductor.org/pub>
- [2] de Leeuw, Jan, "Reproducible Research: the Bottom Line", *Department of Statistics, UCLA. Department of Statistics Papers*, nr. 2001031101, 2001, <http://repositories.cdlib.org/uclastat/papers/2001031101>
- [3] Donoho, D. L., Huo, X., "BeamLab and Reproducible Research", *International Journal of Wavelets, Multiresolution and Information Processing*, 2004
- [4] Eggen, P., and Kauchak, D., "Educational Psychology: Windows on Classrooms", 5th ed., Upper Saddle River, NJ: Prentice Hall, 2001
- [5] Gentleman, R., "Applying Reproducible Research in Scientific Discovery", BioSilico, 2005, <http://gentleman.fhcrc.org/Fld-talks/RGRepRes.pdf>

- [6] Green, P. J., "Diversities of gifts, but the same spirit", *The Statistician*, 2003, p. 423438
- [7] Koenker, R., Zeileis A., "Reproducible Econometric Research (A Critical Review of the State of the Art)", *Department of Statistics and Mathematics Wirtschaftsuniversitt Wien, Research Report Series*, Report 60, 2007
- [8] Leisch, F., "Sweave and beyond: Computations on text documents", *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003, Vienna, Austria, ISSN 1609-395X
- [9] NIST/SEMATECH, "e-Handbook of Statistical Methods", <http://www.itl.nist.gov/div898/handbook/>, 2006
- [10] Peng, R. D., Francesca Dominici, and Scott L. Zeger, "Reproducible Epidemiologic Research", *American Journal of Epidemiology*, 2006
- [11] R Development Core Team. "R: A language and environment for statistical computing.", R Foundation for Statistical Computing, 2008, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [12] Schwab, M., Karrenbach, N. and Claerbout, J., "Making scientific computations reproducible", *Computing in Science & Engineering*, 2000, 2 (6), pp. 61-67
- [13] Statistical Computations at FreeStatistics.org, *Office for Research Development and Education*, URL <http://www.freeststatistics.org/blog/date/2008/Jun/20/t1213979586y9h8zd8mwz5x7br.htm/>, Retrieved Sun, 22 Jun 2008 08:48:50 +0000
- [14] Statistical Computations at FreeStatistics.org, *Office for Research Development and Education*, URL <http://www.freeststatistics.org/blog/date/2008/Jun/22/t1214129827ny0eqihaz22u1x8.htm>, Retrieved Sun, 22 Jun 2008 10:17:18 +0000
- [15] Statistical Computations at FreeStatistics.org, *Office for Research Development and Education*, URL <http://www.freeststatistics.org/blog/date/2007/Dec/05/t11968656188agqc20eb70xnxo.htm/>, Retrieved Sun, 22 Jun 2008 14:12:06 +0000
- [16] Statistical Computations at FreeStatistics.org, *Office for Research Development and Education*, URL <http://www.freeststatistics.org/blog/date/2007/Dec/06/t1196935792ab417daashrthaz.htm/>, Retrieved Sun, 22 Jun 2008 14:13:25 +0000
- [17] Wessa, P., "A framework for statistical software development, maintenance, and publishing within an open-access business model", *Computational Statistics*, 2008, www.springerlink.com (DOI 10.1007/s00180-008-0107-y)
- [18] Wessa, P., "A Compendium of Reproducible Research about Descriptive Statistics and Linear Regression", 2008, URL <http://www.wessa.net/download/tutorial.pdf>
- [19] Wessa P., Histogram (v1.0.9) in Free Statistics Software (v1.1.23-r1), Office for Research Development and Education, 2008, URL http://www.wessa.net/rwasp_histogram.wasp
- [20] Wessa P., "How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment", November 2008, submitted to be presented and published in Proceedings of the 7th European Conference on e-Learning
- [21] Wessa P., "Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing", October 2008, submitted to be presented and published in Proceedings of the 3rd International Conference on Virtual Learning