nsm2019
Nordic Statistical Meeting
HELSINKI 26.–28.8.2019

SAILS: an e-learning system using data of official statistics

Violeta Calian, Statistics Iceland, Violeta.Calian@hagstofa.is

Anna Helga Jonsdottir, University of Iceland, ahj@hi.is

Gunnar Stefansson, University of Iceland, gunnar@hi.is

Jamie Lentin, Shuttle Thread, lentinj@shuttlethread.com

Abstract

Statistics Iceland (SI) and University of Iceland (UI) are collaboratively developing a digital ESS - shareable product in the area of statistical e-learning, the "Statistics Adaptive Intelligent Learning System" (SAILS).This will be a freely accessible, open source, web-based educational system designed for high-school and university students, which utilizes open data of official statistics for training and personalised, interactive drills. It is also a valuable tool for training students or for research on web-assisted education.

The new product is part of and based on a successfully tested, more complex e-learning system developed at the UI, School of Statistics, the tutor-web (http://tutor-web.net). It has been a research project for the past 20 years at UI and it relies exclusively on open source computer code with material under the Creative Commons Attribution-ShareAlike License. Software is written in the Plone, CMS (content management system), on top of a PostgreSQL Server. Educational material is mainly written in LaTeX. Examples and plots are mostly driven by R (R: A Language and Environment for Statistical Computing).
The emphasis of the present project is on the novel features added to the existing solution:
1) the software module for the tutor-web which makes possible for the students to use real data for a variety of statistical analyses. Drill/exercise questions will then be based on such data sets.

2) new educational content which will include tutorials and exercises/drills relevant for data of official statistics.

The software module will be flexible enough and designed such that it will easily adapt to various api/data structures corresponding to various NSIs. To start with, it will be able to use data from Eurostat and at least two NSIs. The e-learning system can be enriched continuously with new open data and with new educational content from any NSI, instructors and other open data providers.

***Keywords**: statistical literacy, e-learning, open data, open source*

1. Introduction

The purpose of this paper is to describe the collaborative project[1] of Statistics Iceland (SI) and University of Iceland (UI) which develop together an open source product in the area of statistical e-learning, the "Statistics Adaptive Intelligent Learning System" (***SAILS***). It consists of the enriched Statistics module of the existing *tutor-web* of UI, and it is obtained by adding new learning and training content and by making use of open data provided by official statistics for exercises and drills. With *SAILS*, individualized, real-life data sets are provided for statistical analysis and learning, to each student, something which was not available until now.

Several characteristics of this product are worth mentioning[2]:

(i) it is a supplementary tool for learning and formative assessment, providing step-by-step solutions to assignments

(ii) students and instructors may contribute to the educational material and all contributions are rated

(iii) a chat platform is provided, an offline support and a mobile version of the tutor-web are available

The system is designed for high-school and university students and it could also become a valuable tool for training students for statistical competitions. In addition, the collected data on students' behaviour may be used for research on web-assisted education such as drill item selection methods. Excercises/drills are designed to increase difficulty gradually, training on a given lecture/tutorial/course has no limit in number of excercises but the answering to each such a drill is timed. Rewards are given in a cryptocurrency called Smileycoin.

2. Experience and results

Tests and surveys have shown learning and satisfaction of users (more than 2000 students tested in Iceland, high schools and university, and in Kenya) with the tutor-web (*http://tutor-web.net*)., i.e. the more complex e-learning system created at the UI, School of Statistics, with development by ShuttleThread. This has been a research

[1] This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 825696.

[2] http://presentations.shuttlethread.com/2016-10-03-mareframe-wp7-update.html#/3

project for the past 20 years at UI and it relies exclusively on open source computer code with material under the Creative Commons Attribution-ShareAlike License.[3]

The experiments reported in the first five articles listed as references show the proof of student progress in measurable terms and provide an analysis of their behavior. Such behavior includes stopping times and the way that grading schemes affect student strategy. The relation between grades obtained in final exams and scores given by the e-learning tool has been also analysed in detail[4]. Several in-class surveys showed general satisfaction with the tutor-web but also the fact that students prefer a combination of assessment and learning methods: the traditional one and the web-based one which gives instant feedback especially for practicing and improving new skills and knowledge.

One of the goals of the tutor-web project might become to link appropriate learning material to the drill items so if a student answers an item incorrectly the student will be pointed towards appropriate material to read. These links are made to material within the system but it would also be interesting to allow users to provide links to other Creative Commons licensed material outside of the system resulting in a completely individualized learning path through an entire course within the system or even the entire web.

The main new contribution of SAILS at the present time is a software module designed for accessing open data from within the e-learning system which is flexible enough and designed for being easy to adapt for various api/data structures corresponding to various NSIs. An early version is already at *https://github.com/tutor-web/twstats*. To start with, it will be able to use data from Eurostat[5] and at least two NSIs. We estimate that the most likely NSI data source will be, in addition to

[3] Software is written in the Plone, CMS (content management system), on top of a PostgreSql Application Server. Educational material is mainly written in LaTeX. Examples and plots are mostly driven by R (*R: A Language and Environment for Statistical Computing*). The statistical module can be accessed from any web-browser and the installation code is available at *https://github.com/tutor-web/tutorweb* . In addition, URL-redirecting / forwarding may be used for efficient sharing

[4] https://arxiv.org/pdf/1406.5004.pdf

[5] https://data.europa.eu/euodp/data/dataset/estat-web-services

Statistics Iceland (SI), Statistics Finland[6], Statistics Netherlands or Statistics Sweden, which provide good open data access as well.

The e-learning system will continue to grow continuously by linking to new open data and by contributions of new educational content from any NSI, instructors and other open data providers.

3. Methods

The project is entirely developed using open source code and it consists of several stages described in what follows.

3.1. Harmonizing the coding of SI open data

The first part of our work is concerned with the standardization of the open data of SI, for SILC (statistics on income and living conditions), LFS (labour force survey), census, i.e. preparing and testing open data of official statistics.

The purpose of this stage is to make sure that the open data of SI which will be used by the e-learning system can be automatically queried in a systematic and consistent way. In order for the automatic queries to work, the micro- and meta- data should have standard coding of the values of variables in all data sets. This standardization is an ongoing process inside SI, where most of the time different departments use different coding for common variables.

Other data sources, like Eurostat open-data, have consistently coded values (labels) of variables (dimensions) already. Data sets from Statistics Netherlands (SN), Statistics Finland (SF) and Statistics Sweden (SS) have consistent coding as well, according to our tests and information provided by these NSIs[7].

For this purpose, a Python script parsed all pc-axis tables (data sets on the SI website) starting with Census, SILC and LFS domains as a pilot analysis. It extracted

6 http://www.stat.fi/org/avoindata/pxweb_en.html

7 http://www.statistikdatabasen.scb.se/pxweb/en/ssd/?rxid=6895730d-5478-49be-8b36-dcb0e3f7d5aa,

http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/,https://ec.europa.eu/eurostat/web/json-and-unicode-web-services/getting-started/generate-new-query,

https://opendata.cbs.nl/statline/portal.html?_la=en&_catalog=CBS

information on: variables, values, codes, domains, file identifiers. (*Example*: domain=Census, variable=gender, value=Total, code=T, file_identifier=CEN0001.)

The results were as follows:

(i) We obtained a database with the following characteristics: 97 variables, 1904 values, 1290 codes, 164 files, for the Icelandic version and same numbers of files for the English version. The analysis showed that:

- most of the variable values (69.7%) have codes and 36.3% are not coded at all (but they receive *default codes automatically*, due to *api* requirements). A small percentage appears in both of these categories, both with codes, in some tables, and without, in others
- out of all the values which are coded, most are coded in one and only one way (95%). The rest have two and more codes for the same value of a variable
- the number of values of variables with no codes (receiving automatic default codes) is: 8.6% for Census, 55.6% for SILC, 39% for LFS

(i) the results of the same analysis performed on the whole set of pc-axis files of SI show similar proportions: 36% of values are not coded. From the values which are coded, about 96% have unique codes and the rest have more than one for same value. In total, the whole set contains:1805 files, 674 variables, 45705 values, 17062 codes.

3.2. Variable coding validation process

The second stage of the project consists of the creation of a *new validation process of all web-tables* published by SI, which detects errors in value coding and corrects them automatically, according to a standard set of codes.

The coding errors are currently detected by a Python parser of pc-axis files (web tables) and are part of the database created for the overview analysis of such files. A new procedure, for the correction/harmonization step will be run as a "layer" between production and publication of tables: it replaces all value codes with harmonized codes where needed and always sends a feedback report about error detection and correction to the author of the files.

The main requirement for such a process to work is to have a database with all the standard codes for all values of all variables found in the published data. We are building this "dictionary" using the following strategy: if the values (labels) exist in Eurostat dictionaries[8], then these codes are adopted as standard codes for SI, as well. If the SI values are not found in Eurostat database then SI creates new ones, based on same principles as used by Eurostat, such as having character codes for all grouping and aggregating variables (dimensions). Such a database is subsequently easy to maintain and update.

For the Census, LFS and SILC data sets, this work is in an advanced stage, since most SI Census values already have standard, Eurostat codes already and they represent 62% of the total values of variables. Most of the SILC and LFS values of variables have received automatic codes, the mapping to the standard codes is work in progress. The process of building the complete dictionary, needed for the whole SI published data is expected it to end this year as well.

3.3. Linking open data and e-learning system

The third stage of the project is concerned with the R-package "*twstats*", which allows automatic queries of Eurostat and several NSIs open data.

The following main resources[9] (see the links therein) are exploited when creating the new module:

- the R package eurostat : tools to download data from the Eurostat database together with search and manipulation utilities.

- the R package cbsodataR: access to Statistics Netherlands' (CBS) open data API from R.

- the R package pxweb : access to pc-axis data of several Nordic NSI's

The package *twstats* which can be found at*https://github.com/tutor-web/twstats* is a crucial component of our project and currently already contains the main R - code for automatic queries of open data of Eurostat and Statistics Netherlands for the time being. More sources will be connected in the coming months.

---

8 Efficiently accessed by using the R package http://ropengov.github.io/eurostat

9 https://github.com/SNStatComp/awesome-official-statistics-software

3.4. New educational content

This stage is dedicated to choosing the best *strategy for defining the relation between learning topics and open data* sets and to the accumulation of new content. We have examined two main strategies for training based on:

- problems tailored to data sets and
- general problems and tests defined for arbitrary data set structures

We converged to using the second scenario since more efficient and since the correct identification of the variable classes is not a problem due to variable value coding. This option offers more flexibility in absorbing new content from other contributors than UI and SI as well as other open data sets. Tailored learning paths can be created efficiently in this setting.

4. Conclusions

The *SAILS* project is timely both for UI and SI, as well as the larger e-learning community. The e-learning system can be enriched continuously with new open data and with new educational content from any NSI, instructors and other open data providers.

References

A.H., Jakobsdottir, A. & Stefansson, G. (2015). Development and Use of an Adaptive Learning Environment to Research Online Study Behaviour. Educational Technology & Society, 18 (1), 132–144.

Jonsdottir, A.H., Bjornsdottir, A. & Stefansson, G. (2015). Difference in learning among students doing pen-and-paper homework compared to web-based homework. Journal of Statistics Education, 25:1, 12-20.

Jonsdottir, A.H. & Stefansson, G. (2014). From evaluation to learning: Some aspects of designing a cyber-university. Computers & Education, 78, 344–351.

Stefansson, G.,& Sigurdardottir, A. J. (2011). Web-assisted education: From evaluation to learning. Journal of Instructional Psychology, 38(1), 47–60.

Jonsdottir, A.H., Development and testing of an open learning environment to enhance statistics and mathematics education , Ph.D. thesis, 2015, University of Iceland.

Methodology for Data Validation 1.0 (Handbook), revised edition June 2016, https://ec.europa.eu/eurostat/cros/content/methodology-data-validation-10-handbook-revised-edition-june-2016_en

Leo Lahti, Przemyslaw Biecek, Markus Kainu and Janne Huovari. Retrieval and analysis of Eurostat open data with the *eurostat* package. R Journal 9(1):385-392, 2017). R package version 3.3.32. URL: http://ropengov.github.io/eurostat