INTRODUCING LARGE DATA SETS INTO THE CLASSROOM: A GRAPHICAL USER INTERFACE FOR TEACHING WITH DATABASES

Heike Hofmann, Ulrike Genschel and Danielle S. Wrolstad
Department of Statistics, Iowa State University, United States of America
ulrike@iastate.edu

Analysis of large, complex data sets is increasingly relevant for today's statisticians. To help facilitate training of databases and SQL (Structured Query Language) at the undergraduate level, we propose a graphical user interface allowing for statistical analyses of large databases using subsampling techniques. The example database contains information on 25 variables for over 120 million commercial flights across the United States since 1987, including information on originating and destination airport and temporal information, such as planned flight schedule, actual take-off and landing times and further qualitative variables. Textual output of a session's SQL commands summarizes students' attempts in interacting with the database providing not only feedback to the instructor but also serving as starting points for more complex aspects of the SQL language similar to SAS (Statistics Analysis Software) scripts initiated from JMP sessions.

BACKGROUND

Over the last decade, computing power and storage capacities have tremendously increased the amount of data that can be collected for a given statistical analysis. Companies are investing significant resources to overcome the challenge of sheer data volume (e.g., the recently awarded Netflix challenge) before conducting any statistical analyses. Further, having massive amounts of data is a doubled-edged sword: Computing time and computability become crucial factors in the analysis that can lead to an operational breakdown of standard analytical tools such as Excel, SAS, or R. Even when partial information from the database would already be sufficient, there is often no easy way for sub-setting or aggregating the data at particular levels.

We have also changed our understanding of what we consider large amounts of data. The size of a data set can be defined in terms the number of observations, the number of variables, or both. What is considered large also depends on the subject area (e.g., microarray analyses typically deal with thousands of observations while clinical trials usually involve a few hundred observations). In general, the transition from normal to large takes place whenever classical tools and procedures no longer work properly when performing analyses (Theus & Urbanek, 2008).

In order to help students become more familiar with large data problems and to give them some experience before entering the workforce, we need to both introduce and facilitate the use of large data sets in a classroom setting.

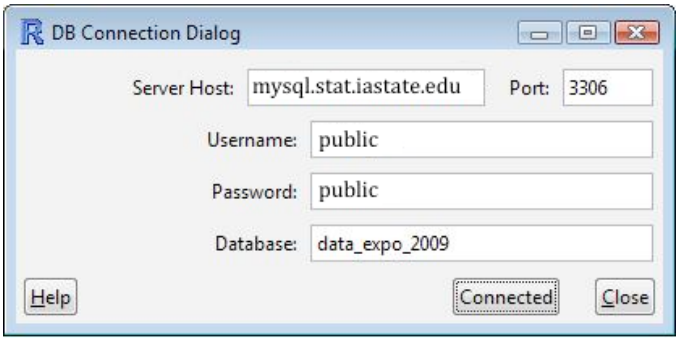


Figure 1. Connection browser to establish database connection

DATABASE CONNECTIONS

We are using MySQL (a relational database management system) for efficient storage and manipulation of the data. For increased efficiency, data are often stored to be in normal form (Kent, 1983). This gain in efficiency imposes an additional technical hindrance due to the now non-rectangular form of the data. Additionally, data retrieval from the database requires

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

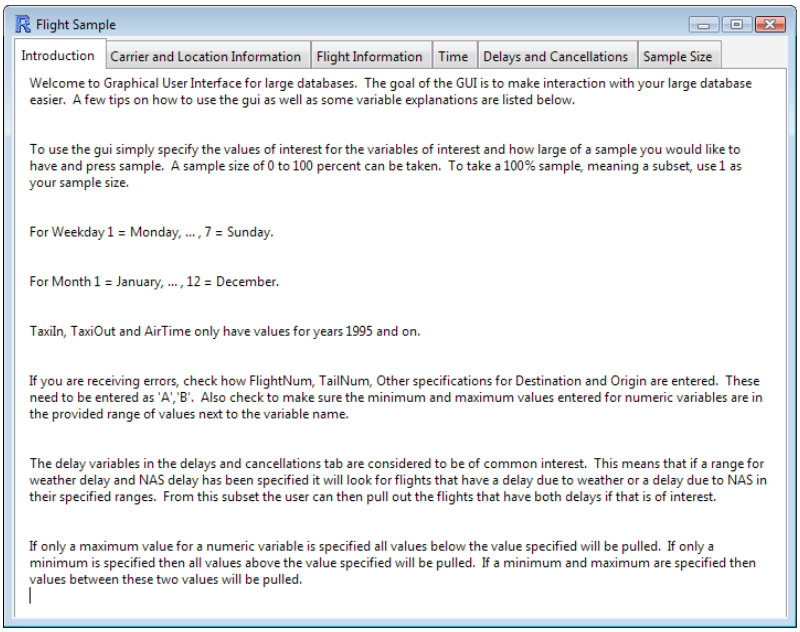
knowledge of SQL, the Structured Query Language. Students tend to find this to be intimidating, whereas a graphical user interface (GUI) allows a gentler introduction to databases due to its more familiar looking interface.

DATA EXAMPLE

To illustrate the use of the GUI, we chose data considered in the 2009 American Statistical Association (ASA) Data Exposition (Wickham, 2009). These data consist of flight arrival and departure details for all commercial flights within the U.S. since October 1987. There are more than 120 million data records taking up 12 gigabytes of hard drive storage uncompressed. We gathered supplemental information from other sources, such as spatial location and runway layouts of airports, as well as operational information provided by the Federal Aviation Administration (FAA) and hourly weather information from the National Oceanic and Atmospheric Administration (NOAA) and Weather Underground (wunderground.com, 2009).

GRAPHICAL USER INTERFACE

The GUI consists of two parts: the connection browser (see Figure 1) and the variable browser (see Figures 2 and 3). Communication with the database is handled through packages DBI (database interface, R Special Interest Group on Databases, 2007) and RMySQL (James & DebRoy, 2008). Figure 1 shows the connection browser that allows the student to establish a connection to the database by choosing a server and a data set. Next, a variable browser dialog window will be displayed. The variable browser consists of six tabs. The first tab (displayed in Figure 2) is an introductory tab that provides the student with tips for working with the GUI. The tab also contains a list of frequently asked questions (FAQ) with answers that provide additional guidance to the students, such as, how to enter tail numbers, origin, destination, or flight numbers. Definitions for levels of categorical variables that are not self-explanatory (e.g., day of the week) are provided as well. The introductory tab also provides the student with an explanation of the interpretation in multiple selections. For example, if the student wants to research all cancellations of a specific aircraft, he/she will need to specify a tail number and set the checkmark for "cancelled" to yes. This extends an SQL statement by "*where TailNum in (' ') AND Cancelled=1*." This tab can be easily adapted according to students' need and feedback.



The introductory tab gives the student a reference for possible questions encountered while using the GUI

Figure 2. Tab Introduction of the variable browser

The other five tabs provide information about and allow the student to interact with carrier and location, dates of flight, time of flight (see Figure 3), delays and cancellation information and sampling.

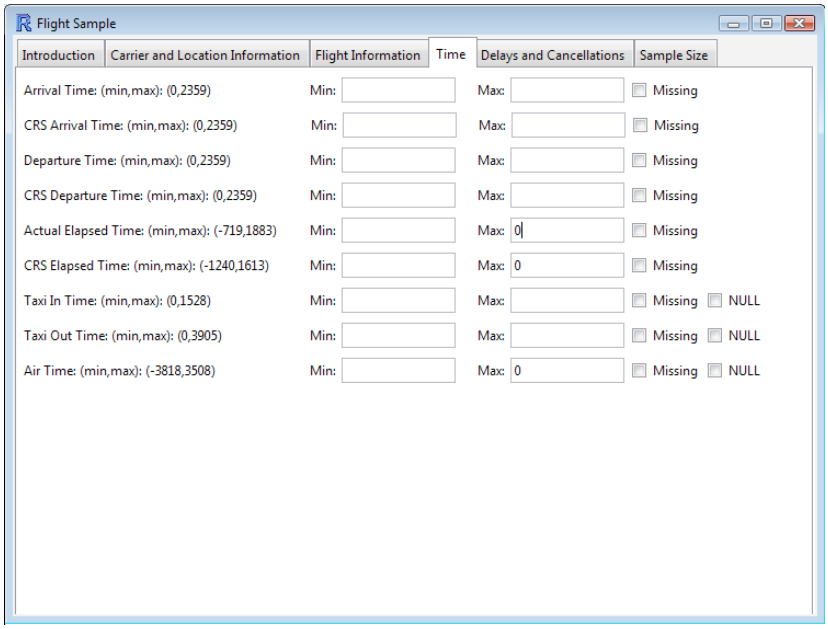


Figure 3. Tab Time of the variable browser for specifying times of flights, amount of taxi-in, taxi-out, and airtime

USABILITY

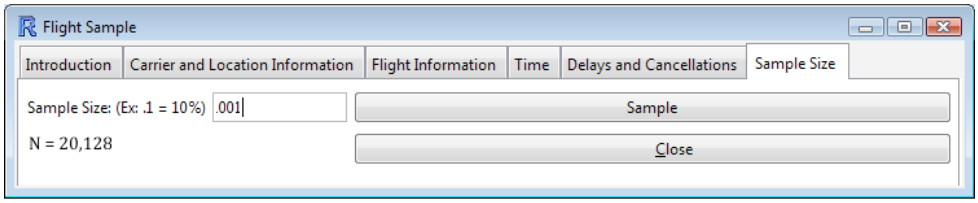
The variable browser serves three main functions:

1. Data verification for easy and immediate verification of data entries,
2. Data retrieval/database sub-setting to prepare the data for statistical analysis,
3. Learning tool for the structured query language (SQL).

Figure 3 shows the type-dependent summary of all variables: the minimum and maximum value for numerical variables and the list of levels for a categorical variable. This information enables the student to screen the database for invalid data entries (e.g., negative minimal values for the actual elapsed time of an airplane between leaving the gate at the originating airport and arriving at the destination airport gate that can theoretically only be positive; see Figure 3). The student immediately gets prompted to further investigate the data before beginning with any statistical analyses that potentially might lead to erroneous results.

The last tab of the variable browser (Figure 4) allows the student to retrieve data from the database and to import it as regular data frames into R. Samples can be taken from the entire database or just from a range of data values/variable levels by specifying the lower and upper endpoints in editable text boxes in the previous tabs (Figure 3). Samples sizes are determined according to the sampling fraction of interest and taken at random from the chosen portion of the database.

The GUI serves also as a learning tool for the SQL as it provides textual output of the session's SQL commands that then can be saved for future reference by the students. This helps students to overcome an often steep and intimidating initial learning curve.



A sampling rate of 0.001 results in a subset of 20,128 observations.

Figure 4. Sampling tab of the variable browser

PUBLIC AVAILABILITY AND DEVELOPMENT OUTLOOK

To achieve broad dissemination of the developed tools ready for classroom implementation, the database is set up for public use on a server of the ISU Department of Statistics.

An accompanying webpage is available at http://www.public.iastate.edu/~hofmann/vldb.html.

This webpage provides the following information and tools:

1. R source code for the GUI ready to download and for further development,
2. information on accessing the SQL database,
3. detailed description of the data,
4. examples and sample analyses for classroom demonstration,
5. contact information for feedback.

We introduced the idea of a graphical user interface that allows an introduction of data sets into classroom that otherwise are too large in size to be managed by commonly used statistical software. We explained some of the technical and pedagogical aspects of the GUI and an extension of the GUI is currently under development. Although the GUI has been used successfully in the classroom a full implementation and evaluation of the GUI is still necessary and currently being prepared.

REFERENCES

James, D. A., & DebRoy, S. (2008). RMySQL: R interface to the MySQL database. *R package version 0.7-2*.

Kent, W. (1983). RMySQL: A simple guide to five normal forms in relational database theory. *Communications of the ACM*, *26*(2), 120-125.

R Special Interest Group on Databases (R-SIG-DB) (2007). DBI: R database interface. *R package version 0.2-4*.

Theus, M., & Urbanek, S. (2008). *Interactive graphics for data analysis: Principles and examples.* London: Chapman & Hall–CRC.

Wickham, H. (2009). *ASA Data Exposition 2009*. Washington, D.C.: American Statistical Association, Section on Statistical Computing and Statistical Graphics. http://stat-computing.org/dataexpo/.

wunderground.com (2009). *Weather underground*. Ann Arbor, MI: Weather Underground, Inc. http://www.wunderground.com/.