MAKING MULTILEVEL DATA IDEAS MORE ACCESSIBLE

Danny Parsons[1], David Stern[2] and Roger Stern[2,3]
[1]Mathematical Institute, University of Oxford, UK
[2]University of Reading, UK
[3]Statistics for Sustainable Development, UK
danny@aims.ac.za

Each year increasing amounts of data are being produced and there are growing trends towards data becoming more accessible, particularly online. Here we present a range of examples where data are conveniently arranged in multiple linked rectangles or data frames. They are often omitted from all but advanced statistics courses. However, they are common in practice, hence their omission leaves graduates poorly prepared for real world problems. The obvious example is a survey that is at multiple levels. Other examples include multiple time series with spatial data, where the spatial information is in a separate data frame; and data sets in a single rectangle (data frame) but where the analyses are on summary data. The statistical software, R-Instat, resulting from the African Data Initiative is designed to make it easy to handle such data.

INTRODUCTION

There is an ever increasing amount of data being generated and being made accessible to everyone (King, 2007). Organizations such as the UN and governments in many countries are recognizing the importance of data, data literacy and statistics to achieve sustainable development, for example UN's Data Revolution (Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG), 2014). There are also shifts towards opening up access to data because of the recognized potential impact the information that could be extracted from data could have for people and communities (Gurstein, 2010).

Multi-level data, also known as hierarchical or nested data are data in which cases are contained in groups i.e. at another level. These groups could be contained in further groups, and in general there could any number of groups or levels. A classic example is students grouped in classrooms, which are in turn grouped in schools, and further into districts (Fidell & Tabachnick, 2007). Many interesting and useful datasets are naturally multi-level and so there is need for people who can work with multi-level data to allow everyone to benefit from the information that could be derived from such data.

Multi-level data is often only taught at post graduate level courses. Many statistics undergraduate courses would typically contain very little multi-level data analysis, or if it did, it would usually be towards the end of a degree. One reason for this is that multi-level data analysis in statistics courses often focuses on modelling aspects, which are algorithmically complex and so would not be appropriate earlier.

However, there are aspects of multi-level data analysis which appear naturally in real world data problems and do not involve complex modelling. We claim that there is value in teaching aspects of multi-level data analysis earlier in statistics courses in ways that are accessible for students.

Statistics software should play an important role in statistics courses so students get experience of practical aspects of statistics. In most software, a single rectangle is for data at one level, and hence multi-level data requires a set of two or more rectangles. Although most statistics software are well prepared to analyze single rectangles, most are less suitable for handling multiple rectangles. To overcome this most statistics software have good facilities for merging data from multiple rectangles and hence data from multiple levels is easily brought to a single level. However, the merging process can be complex and can easily be done incorrectly. An alternative is to leave data in separate data frames, with the addition of meta-data to specify how they are linked.

The African Data Initiative has produced a front-end system to the popular statistics language, R (Stern, 2017). This front-end is called R-Instat and it has taken steps towards managing data in multiple data frames (rectangles). First, it has a spreadsheet like interface for viewing multiple data frames in separate sheets. The software can accommodate multiple rectangles of data in one instance of the software which can then be saved as a single file, similar to a spreadsheet package. Second, it has implemented some features similar to those found in a database structure, such as relations between tables (data frames) and key columns. With these features, particularly the relations or links, which specify how data frames are connected, multi-level data analysis can be done without the need to disrupt the natural position of data at different levels. This is achieved through R-Instat's calculation and summary system. Thus, merging is not a prerequisite for performing multi-level analysis.

The approach could have an impact on how teaching multi-level data analysis is thought of since it could begin to make multi-level data analysis accessible to students at an earlier stage than it is currently introduced.

METHODS

In R-Instat data, metadata and objects (graphs, models, tables) are contained in a single structure in R, called an *instat object*. The *instat object* has features which are similar to structures of a relational database, such as keys and relationships. Keys and links (relationships) are the important features which facilitate working with multi-level data in R-Instat through its calculation and summary system.

Keys

In R-Instat each data frame may contain a set of *keys*. A *key* is a set of columns which uniquely identifies each row. For example, daily time series data may have a date column as a key but also the three columns year, month, day could also form a key, as could other sets of columns. Multiple keys can be defined on a data frame.

The concept of a key is not new in R, for example, the data.table package uses keys to allow extremely fast operations to be done on the data.table (CRAN R Project, 2017). R-Instat does not currently use the data.table package as it does not easily fit into our current implementation, but we hope to incorporate it in the future.

Keys are useful in their own right, for example, checking whether a set of columns are able to define a key is a useful data checking tool since it specifies the uniqueness of the set of columns. However, the real power of keys comes when multiple data frames are used which have relationships between them.

Links

In R-Instat a *link* is a relationship between two data frames where a row from the second data frame can have multiple matching rows in the first. In database terminology links encode many-to-one (and one-to-one) relationships. Hence, a requirement to be a valid link is that the columns specified in the second of the two data frames must define a key so that a many-to-one relationship is valid. For example, two time series data frames, one at the daily level and one at the yearly level, could be linked by the year columns in both data frames, where the year column must define a key in the yearly level data.

Links can be imported into R-Instat from recognized databases structures, be defined manually, or be created automatically through the calculation of column summaries. Once links have been defined (either manually or automatically) R-Instat's calculation and summary system can use the links to identify the *natural* data frame that calculated and summarized data should be placed.

For example, daily time series data can be summarized to the yearly level through the *Column Summaries* dialog. A new data frame will be created for the yearly level data (if it doesn't exist) as well as a link between the data frames. If different summaries were then done to the same level, by default R-Instat will add the new summaries into the existing data frame instead of creating a third data frame. By using the information in the link R-Instat attempts keep data at the same level in the same data frame and hence the number of data frames is kept to a minimum, helping users organize their data.

RESULTS

Here we illustrate three areas where multi-level analysis plays an important role in understanding the data, and explain how R-Instat approaches these analyses.

Survey data

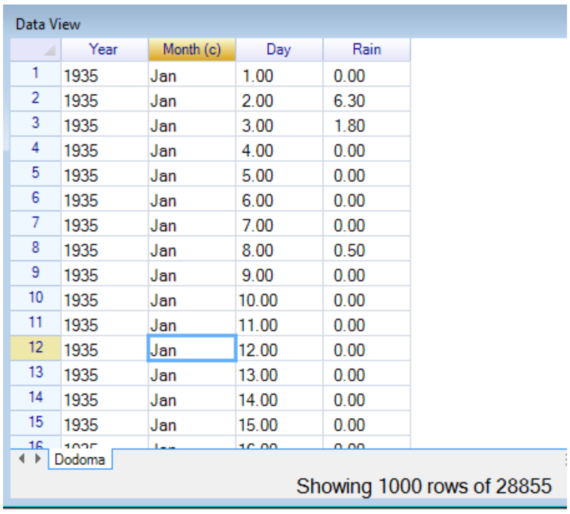
Surveys of households are a common tool used by national statistics offices and development agencies to try to understand various factors about a population, including welfare, health and income levels (Deaton, 2003). Data from such surveys are usually at multiple levels, for example, household level and village level. Data at different levels could be entered in different forms, and entry tools such as CSPRO and ODK can include information about the structure of the data and its links. However, this is often lost when importing into a standard statistics package. With the database structures of R-Instat, these links can be retained and then used in analyses which often involves combining household and village level data.

Multiple response questions are common in many surveys. Data from multiple response questions could be stored with the other responses in a single rectangle of data. This makes analysis straightforward but can make the data unnecessarily wide and affect the speed of doing analysis. Alternatively, there could be a separate rectangle for each multiple response question so that each rectangle has a simple structure, although it then may be harder to analyze across rectangles. However, if there are links between the rectangles than analysis can still easily be done across the whole data while maintaining the simple structure of each rectangle.

Climatic data analysis

Historical climatic data is often available on a daily basis. Figure 1a shows an example of these data for Dodoma in Tanzania obtained from the Tanzania Meteorological Agency (Tanzania Meteorological Agency, 2017). The results shown in Figure 6b were obtained from using a dialog from R-Instat's climatic menu. This data frame in now on a yearly basis, which is often the level of interest for climatic analyses, and a link has also been automatically created between the yearly data and the primary daily data. This means the primary data is not lost and could be referred to later, even though the summarized data is being used for analysis.

Dodoma Data in R-Instat

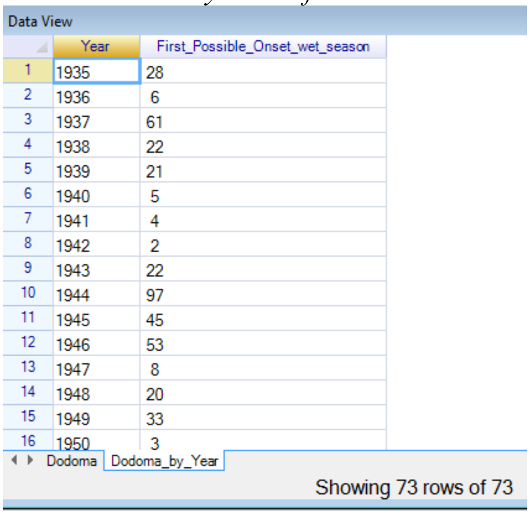


	Year	Month (c)	Day	Rain
1	1935	Jan	1.00	0.00
2	1935	Jan	2.00	6.30
3	1935	Jan	3.00	1.80
4	1935	Jan	4.00	0.00
5	1935	Jan	5.00	0.00
6	1935	Jan	6.00	0.00
7	1935	Jan	7.00	0.00
8	1935	Jan	8.00	0.50
9	1935	Jan	9.00	0.00
10	1935	Jan	10.00	0.00
11	1935	Jan	11.00	0.00
12	1935	Jan	12.00	0.00
13	1935	Jan	13.00	0.00
14	1935	Jan	14.00	0.00
15	1935	Jan	15.00	0.00
16	1935	Jan	16.00	0.00

Showing 1000 rows of 28855

Figure 1a: Dodoma data.

Yearly Start of Rains

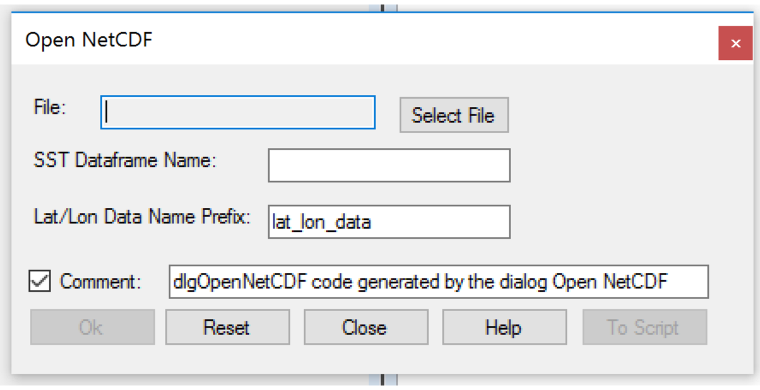


	Year	First_Possible_Onset_wet_season
1	1935	28
2	1936	6
3	1937	61
4	1938	22
5	1939	21
6	1940	5
7	1941	4
8	1942	2
9	1943	22
10	1944	97
11	1945	45
12	1946	53
13	1947	8
14	1948	20
15	1949	33
16	1950	3

Showing 73 rows of 73

Figure 6b: Linked data frame with a yearly basis.

Climatic data often comes together with spatial data, for example, EUMETSAT makes data on radiation freely available in a NETCDF format (European Organization for the Exploitation of Meteorological Satellites, 2017) and Figure 2 shows the dialog for importing these data into R-Instat. They import directly into two linked data frames, with one for the actual measurements and another that provides the location information for each pixel. By making use of the link, the measurement data could be filtered based on the location data, for example, to do an analysis on all the measurement data from stations with a certain radius of a given location. By making use of the link, the two rectangles need not be merged to perform the filter and the data can remain at their natural levels.



File:

SST Dataframe Name:

Lat/Lon Data Name Prefix: lat_lon_data

☑ Comment: dlgOpenNetCDF code generated by the dialog Open NetCDF

Figure 1: Dialogue to input data.

Analyzing data for corruption risks

A recent advancement has been made in measuring corruption risks in government contracting (Fazekas, Tóth, & King, 2016). This method makes use of large datasets on government procurement contracts to identify proxy indicators for corruption risks. R-Instat has a special corruption risks menu which implements the methods of this new approach. The primary data for this analysis is at the procurement contract level. However, there may also be data at the country and year level, such as the data about the financial secrecy of countries, which could be useful when analyzing contracts won by companies based outside the origin country.

Often contract level data is freely available, such as for EU countries (Digiwhist, 2017) and Tanzania (Public Procurement Regulatory Authority (PPRA), 2017). Individuals or organizations, such as government agencies, may have their own additional data which could enhance the analysis but may be confidential data, for example, data on organizations that are involved in the procurement process. These data are at different levels and hence could be linked to the contract level data without the need to merge the confidential data. This would make it easier to share parts of the data or the results from all the data without compromising the confidentiality of certain data.

DISCUSSION

The examples of multi-level data analysis shown in the Results section all relate to real world problems which also have important applications to development issues. There are therefore obvious benefits that could be made by supporting people working in these specific areas. For example, staff in national or regional meteorological services could benefit from R-Instat's special climatic menu to calculate summaries such as when the rain season starts. Here the concept of the start of the rains is simple to understand but implementing this as code is non-trivial and so we believe there could be a role for tailored products that can do specific analyses, to support people who have not been primarily trained in statistics.

If analyses of such data becomes more accessible, this could have implications on statistics education by enabling the teaching and learning of statistics to become more relevant to the real world through more use of real and complex data as examples. By including more real world data in statistics education this will automatically bring in multi-level analysis since this is often the structure of real world data.

This could then lead to opportunities for changes in the statistics curriculum to include ideas of multi-level data earlier than it is currently introduced. Such data may be more complex than is usually used in statistics courses, but methods of analysis used on multi-level data in early statistics courses need not focus on complex modelling techniques and could instead initially focus descriptive analyses such as those outlined in the Results section.

This would leave students more prepared to handle real world data and better able to contribute to solving problems which naturally involve using data from multiple levels.

CONCLUSIONS

In this paper we described how the concept of keys and links in R-Instat were implemented to allow users to analyze multiple rectangles of data (at multiple levels). We also explained how R-

Instat uses these links to ensure that data produced from calculations and summaries can be placed at their natural level, helping users to manage data in multiple rectangles.

Many real world datasets are naturally multi-level, as illustrated in the Results. Hence, R-Instat's approach has implications for education by enabling more use of real world datasets and providing an opportunity for changes in the curriculum which could see multi-level data being introduced earlier than it currently is. This is consistent with current literature which advocates for more use of real world data in statistics teaching. For example, Witt (2014) provides examples of climate science data which could be used to teach introductory statistics and suggests that the use of data on this relevant topic could stimulate an enthusiasm for statistics in students. Ridgway and Smith (2013) have described tools for visualizing large, complex multi-level data and discuss the implications these could have in statistics education.

We have shown that the keys and links structures in R-Instat are useful to facilitate analyses such as those described in the Results section. With these structures in place a next step is to investigate other potential benefits in using links in statistical analyses, in particular, in multi-level modelling. This is a separate topic to the kinds of analyses discussed in this paper but we intend to investigate how these R-Instat structures could also make multi-level modelling more accessible.

ACKNOWLEDGEMENT

We are very grateful to the Tanzania Meteorological Agency for their permission for the Dodoma data to be used for training purposes and allowing the data to be included in R-Instat.

REFERENCES

CRAN R Project. (2017, January 31). *Keys and fast binary search based subset.* Retrieved May 30, 2017, from https://cran.r-project.org/web/packages/data.table/vignettes/datatable-keys-fast-subset.html

Deaton, A. (2003). Household Surveys, Consumption, and the Measurement of Poverty. *Economic Systems Research, 15*(2), 135-159.

Digiwhist. (2017). *Data: Explore diverse procurement datasets from around Europe*. Retrieved May 30, 2017, from Digiwhist: http://digiwhist.eu/resources/data/

European Organization for the Exploitation of Meteorological Satellites. (2017). *Data - EUMETSAT*. Retrieved May 30, 2017, from EUMEtSAT: http://www.eumetsat.int/website/home/Data/index.html

Fazekas, M., Tóth, I., & King, L. (2016). An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research, 22*(3), 369-397. doi:10.1007/s10610-016-9308-z

Fidell, L., & Tabachnick, B. (2007). *Using Multivariate Statistics (5th ed.).* Montreal, Canada: Pearson/A & B.

Gurstein, M. (2010, September 2). *Gurstein's Community Informatics.* Retrieved May 30, 2017, from Open Data: Empowering the Empowered or Effective Data Use for Everyone?

Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG). (2014). *A World that Counts: Mobilising the Data Revolution for Sustainable Development.* Report prepared for the UN Secretary General by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development.

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods, 36*(2), 173-199.

Public Procurement Regulatory Authority (PPRA). (2017, May 30). Retrieved from https://www.ppra.go.tz

Ridgway, J., & Smith, A. (2013). Open data, official statistics and statistics education: threats, and opportunities for collaboration. *Proceedings of the Joint IASE/IAOS Satellite Conference "Statistics Education for Progress".* Macao, China.

Stern, D. (2017). Seeding the African Data Initiative. In A. Molnar (Ed.), *Proceedings of the IASE Satellite Conference, Rabat, Morocco.*

Tanzania Meteorological Agency. (2017). *Tanzania Meteorological Agency*. Retrieved May 30, 2017, from Tanzania Meteorological Agency: http://www.meteo.go.tz/

Witt, G. (2014). Using data from climate science to teach introductory statistics. *Journal of Statistics Education, 21*(1), 1-23.