PRACTICAL PUBLIC ONLINE DATA: INTRODUCING TUVA AND CODAP

Tim Erickson
5269 Miles Avenue, Oakland, CA 94618 USA
eepsmedia@gmail.com

This is a report on a hands-on workshop in which we explored two web-based data analysis tools — Tuva and CODAP — designed for novices. The author has used both systems, and developed curriculum materials for students, ranging from high-school students in the USA to post-secondary learners in Sudan. We will provide links to the online materials we used. However, the tools are changing rapidly, and we don't have control over public data. Therefore, the specific links from this paper and our online document, although containing extensive instructions, will not be maintained. Still, both tools will have easily-accessible introductory material. Rather than describing the tools in detail, this paper briefly lays out the need for tools for novices, gives an extended example, and frames some questions about their features and use.

INTRODUCTION

Consider the democratization of data. What an evocative phrase! We are data educators, after all, so we want everyone to understand data better. At some level, we all believe that understanding data will, in general, improve your life; therefore educating people about data is, at its root, making the world a better place. It's a reason for doing what we do.

The most engaging and relevant data the public needs to understand is, frankly, data about the public. In our classes and workshops, participants are excited when we use data about people and human institutions — our challenges and triumphs, our preferences and our problems. We justify using data about the public partly by saying that you need to understand data to be a good citizen in a democracy: we need to understand the societies we live in, in order to make good choices about our future.

Where shall we get the data, and how will our students engage with it? Other projects from this conference (See, e.g., Desmedt 2016, Engel 2016, Cleveland, Hall & Jeffers 2016 and Lopez and Batanero, all in this volume) are devoted to finding suitable public data and making it freely available. Some is available already from public sources. But we know that simply having data *available* does not make it *useful* — especially to a public with little experience in how to work with data and analyze it.

As an example, let's look at the World Bank. Its public data lives at http://data.worldbank.org/, and includes stats on many countries and regions. Some data are organized according to their relevance to important ideas such as Sustainable Development Goals (SDGs). You can make graphs of data right on their site, but the graphing options are limited, both in which attributes can be displayed and in the types of visualizations.

You can, however, download a suitable file (e.g., a CSV) of the data. If you speak JMP, Python, or R, you can now make whatever graph you want. But our task, in this workshop, was to think about what kind of access we can give to what amounts to lay people: members of the public. That will include many decisionmakers.

Must decisionmakers hire statisticians and data scientists to process data and make visualizations? We hope not. But to what extent can they do it themselves? Assuming they will *not* be learning R anytime soon, this workshop looked at two existing tools to see what they can get out of the data. Because it was a workshop, the participants actively worked with the data. You, dear reader, may well rather use R or your tool of choice. But this was a chance to play with simpler tools in friendly company.

We used two tools —

- Tuva (http://www.tuvalabs.com) and the
- Common Online Data Analysis Platform (CODAP) (http://codap.concord.org),

with which I have been associated.

Here is a link to the Google doc we used in the workshop: http://bit.ly/TEE2016Berlin.

Through hands-on tasks from real lessons, participants learned how to look for data, how to make and modify all sorts of graphs, and how to perform relevant computations. They also

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE),
July 2016, Berlin, Germany. ©2016 ISI/IASE iase-web.org/Conference_Proceedings.php

mastered, first-hand, additional important skills, for example, looking at subsets of data, and actually getting a new chunk of data into each of these systems.

In general, we saw that both systems were reasonably easy to use. People were able to access data and make graphs according to the instructions. In doing so, we ran across interesting situations that are emblematic of using real, as opposed to sanitized, public data.

AN EXAMPLE

Here is an example of the kind of thing we explored:

We looked at data on cereal yield—the number of kilograms of grain that are harvested per hectare under cultivation—for countries in sub-Saharan Africa and in "Latin America," which includes the Caribbean in the World Bank data.

Here is a graph from Tuva showing the region-wide time series. This required filtering to get rid of the rest of the world's data and make the chart more legible. Just this much is enough to provoke discussion about the data.

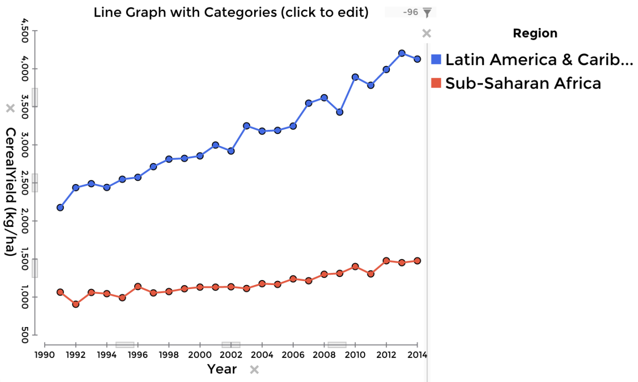


Figure 1: Cereal (grain) yield, in kg/ha, by year, for sub-Saharan Africa and for Latin America.

It's clear that Latin America has higher yields than Africa. When we break that down by country, however, we get this:

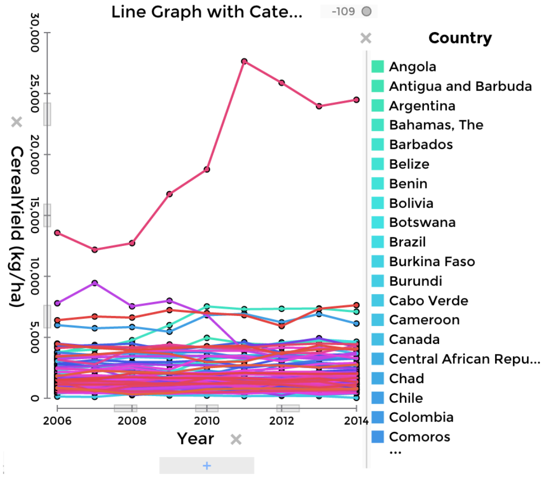


Figure 2. Cereal yield by year for the individual countries in those two regions. Yikes.

What do we want novices to do with a graph like this? If coping with this is part of "data literacy," what skills do they need?

For one, they need to recognize that the graph is a mess. Way too much data and too many lines. One response is to reduce the dimensionality. If we really want data by individual countries, perhaps we should give up on the time series. Then we could make a graph of the distribution of yield by region at one point in time. Figure 3 shows 2014, and shows the median:

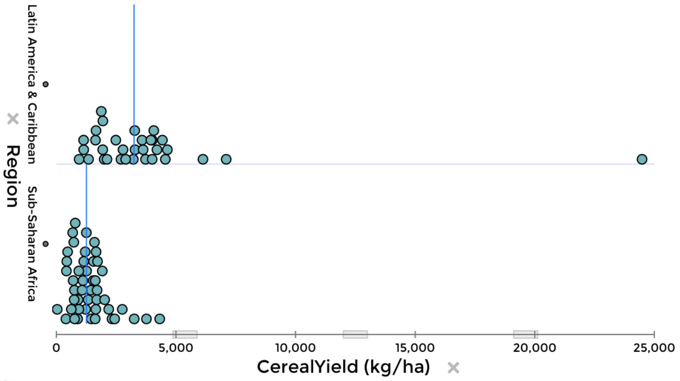


Figure 3. Distribution of yields for the same countries, but only for 2014.

Next, they should attend to outliers. What's going on with the point that's so high? At 20,000 kg/ha, this country is about five times more productive than Latin America as a whole. Users have to recognize an important point in the graph and know how to find out what it represents.

Both tools let us easily "drill down" and see all of the attribute values for any case. That outlier represents St. Vincent and the Grenadines, a small nation in the Eastern Caribbean. Do the Vincentians have a miracle procedure for growing grain? Maybe, but probably not. With further drilling, we learn that they have 35 hectares in production (in contrast, Sierra Leone has about 740,000 hectares and Brazil has almost 22 million). So grain production is not a big part of the Vincentian economy, and the numbers they gave the World Bank might have been rough estimates.

So if we wanted to compare the two regions (Africa and Latin America) using the distributions of yields for the different countries, it might be best to limit which countries we include. In Figure 4, the same graph, but only for countries with more than 10,000 hectares in production:

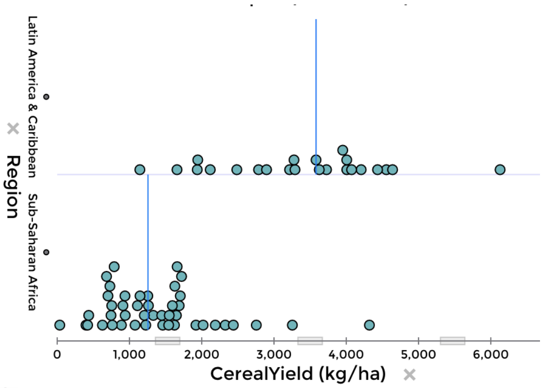


Figure 4. Distribution of grain yields by country, two regions, 2014, but only for countries with more than 10,000 hectares under cultivation.

Now we see clearly that, although the median value for Latin America is much higher than that for sub-Saharan Africa, the distributions by country overlap, and the top-performing African countries (South Africa, Côte d'Ivoire) fit well with the bulk of Latin America.

Think of the skills and "data maturity" that whole procedure entails, and how it's not really part of the traditional statistics curriculum. Do we leave it out because it's too elementary? Or do we always sanitize our data because it's too complex? Figure 2 is a mess, but it's exactly the kind of thing people will come up with if we turn them loose to explore public data.

I claim we all need these skills—skills that include filtering the data, understanding its dimensions, adjusting an analysis's complexity, attending to outliers, drilling down to explore details, and bringing outside knowledge and "data sense" to bear. And all this is somehow separate from (and arguably more important than) understanding inference.

COMPARING TUVA AND CODAP

The previous illustrations and procedures came from Tuva, and were easy for participants to make and do. CODAP is similarly easy, although the user interface is a bit different.

Overall, the two tools are quite similar, but differ in a number of important ways:

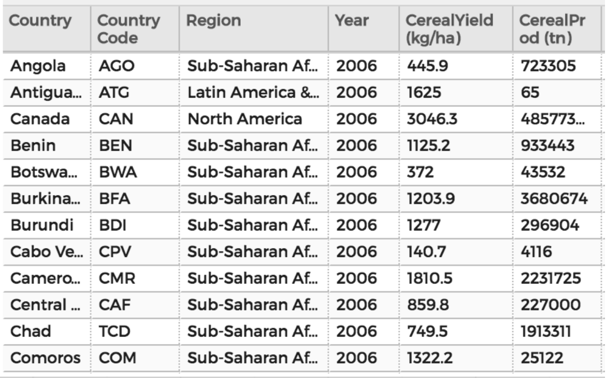
- Tuva is designed for easy access to curated data sets and "lessons"—series of screens with access to the live tool, where students can read instructions and answer questions. CODAP has no special data repository, though you can open files. CODAP is more geared towards getting data from "data interactives," which might generate data or get it from feeds.
- Tuva includes a wider variety of plot types including histograms, box plots, bar charts, pie charts, etc. CODAP, in contrast, currently makes all its plots with dots, though it allows adornments such as shading the IQR.
- Tuva has only one graph visible. CODAP allows any number of graphs, and supports synchronous selection among all views of the data.
- Tuva's data organization is flat, while CODAP's allows a hierarchical structure.

NON-FLAT DATA STRUCTURES IN CODAP

This last bit is unusual enough to warrant an explanation. When we looked at the individual country data above in Tuva, we had to open an entirely new data set. There was no way, for example, to start with the country data and aggregate it to see the region data.

In CODAP, you can create a new level of hierarchy—in this case, the regions—by dragging the defining attribute leftward in the table. You can also group the time-series values into their countries. You can think of this as grouping the cases using the values that get duplicated from row to row.

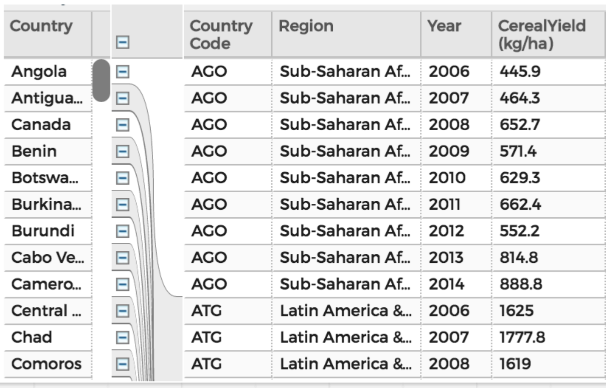
So if we begin with a flat table, we might see data for 2006 for every country, followed by data for 2007, then 2008, and so forth, as in Figure 5:



Country	Country Code	Region	Year	CerealYield (kg/ha)	CerealPr od (tn)
Angola	AGO	Sub-Saharan Af...	2006	445.9	723305
Antigua...	ATG	Latin America &...	2006	1625	65
Canada	CAN	North America	2006	3046.3	485773...
Benin	BEN	Sub-Saharan Af...	2006	1125.2	933443
Botswa...	BWA	Sub-Saharan Af...	2006	372	43532
Burkina...	BFA	Sub-Saharan Af...	2006	1203.9	3680674
Burundi	BDI	Sub-Saharan Af...	2006	1277	296904
Cabo Ve...	CPV	Sub-Saharan Af...	2006	140.7	4116
Camero...	CMR	Sub-Saharan Af...	2006	1810.5	2231725
Central ...	CAF	Sub-Saharan Af...	2006	859.8	227000
Chad	TCD	Sub-Saharan Af...	2006	749.5	1913311
Comoros	COM	Sub-Saharan Af...	2006	1322.2	25122

Figure 5. Table in CODAP showing crop yields by country and year

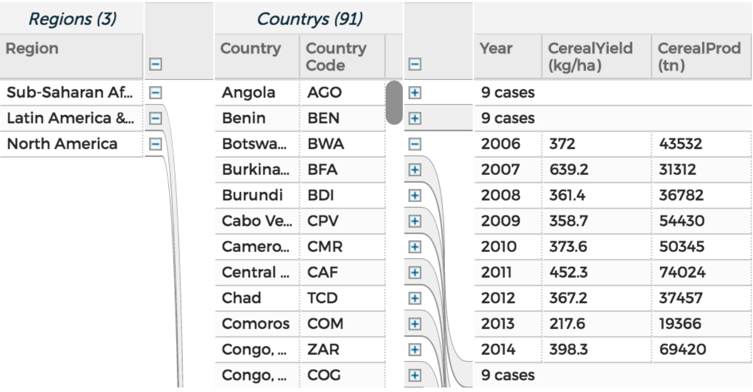
If we drag Country to the left, we see Figure 6



Country	Country Code	Region	Year	CerealYield (kg/ha)
Angola	AGO	Sub-Saharan Af...	2006	445.9
Antigua...	AGO	Sub-Saharan Af...	2007	464.3
Canada	AGO	Sub-Saharan Af...	2008	652.7
Benin	AGO	Sub-Saharan Af...	2009	571.4
Botswa...	AGO	Sub-Saharan Af...	2010	629.3
Burkina...	AGO	Sub-Saharan Af...	2011	662.4
Burundi	AGO	Sub-Saharan Af...	2012	552.2
Cabo Ve...	AGO	Sub-Saharan Af...	2013	814.8
Camero...	AGO	Sub-Saharan Af...	2014	888.8
Central ...	ATG	Latin America &...	2006	1625
Chad	ATG	Latin America &...	2007	1777.8
Comoros	ATG	Latin America &...	2008	1619

Figure 6. The same table, now grouped by country.

Now you can see that Angola's values for 2006–2014 are now grouped together, and you can collapse [–] them if you wish. Antigua's data follow Angola. We notice that country "code" really belongs with the country name; we can promote it as well. Then, Region really belongs farther left, in a new, higher level, as in Figure 7:



Regions (3)	Countrys (91)				
Region	Country	Country Code	Year	CerealYield (kg/ha)	CerealProd (tn)
Sub-Saharan Af...	Angola	AGO	9 cases		
Latin America &...	Benin	BEN	9 cases		
North America	Botswa...	BWA	2006	372	43532
	Burkina...	BFA	2007	639.2	31312
	Burundi	BDI	2008	361.4	36782
	Cabo Ve...	CPV	2009	358.7	54430
	Camero...	CMR	2010	373.6	50345
	Central ...	CAF	2011	452.3	74024
	Chad	TCD	2012	367.2	37457
	Comoros	COM	2013	217.6	19366
	Congo, ...	ZAR	2014	398.3	69420
	Congo, ...	COG	9 cases		

Figure 7. The same table, also grouped by Region.

Now we can make our analyses and graphs using attributes from *any level in the hierarchy*, and CODAP does the appropriate thing. We can also make new columns—again, at any level in the hierarchy—and either enter values or create aggregate formulas. So if you want the highest yield for each country, you would make a new column at the middle, "Countrys" level and write an appropriate formula.

Part of CODAP's research agenda is figuring out how useful this capability is. One can argue that students, left to their own devices, often organize data hierarchically using paper and pencil. Then we have to force our students to flatten their data in order to enter it into the analysis software. How interesting to have analysis software that "understands" and supports the human organization!

CONCLUSIONS

Working with tools for novices, and imagining data-literacy learning, raises important questions for discussion. Here are two to ponder:

First, what is the relative value of curated versus self-selected data sets? In the workshop we mostly used data I had chosen, so it was curated but purposely not sanitized. It is more powerful, however, to learn with data you have chosen yourself. Unfortunately, (a) it can be hard to find the data you're looking for; (b) it can be hard (though we showed how it can be done) to take arbitrary data of your own choosing and move it into such tools; and (c) that data may not teach what we instructors want to teach.

Second, should we use microdata or pre-aggregated data? Of course the answer is "both," but under what conditions? Microdata are often closer to the "story" and easier to relate to, but they require particular skills (Frischemeier, 2016) are harder to come by.

We hope that as a result of this workshop, participants became at least a little familiar with both systems as potential platforms for their own work, and as sources of ideas for their own development efforts.

REFERENCES

Desmedt, M. 2016. European Statistics and Eurostat's contribution to improving stat lit. Paper given at this conference. In J. Engel (Ed.) Proceedings, IASE 2016 Roundtable Berlin. Retrieved: http://iase-web.org/Conference_Proceedings.php.

Engel J., 2016. *Statistics education and monitoring progress towards civil rights*. In J. Engel (Ed.) Proceedings, IASE 2016 Roundtable Berlin. Retrieved: http://iase-web.org/Conference_Proceedings.php.

Finzer, W. 2016 and ongoing. Common Online Data Analysis Platform (CODAP). Emeryville, CA: Concord Consortium. (http://concord.org/codap)

Frischemeier, D., Bieler R., & Engel, J. (2016). *Competencies and dispositions for exploring micro data with digital tools*. In J. Engel (Ed.) Proceedings, IASE 2016 Roundtable Berlin. Retrieved: http://iase-web.org/Conference_Proceedings.php.

Hall, P.K. and Cleveland, L. (2016). IPUMS International: Promoting understanding of statistics about society with free online data from international censuses. In J. Engel (Ed.) Proceedings, IASE 2016 Roundtable Berlin. Retrieved: http://iase-web.org/Conference_Proceedings.php.

Lopez-Martin, M., Batanero, C., & Arteaga, P. (2016). Using United Nations Data in the Training of Teachers to Reach Statistics. In J. Engel (Ed.) Proceedings, IASE 2016 Roundtable Berlin. Retrieved: http://iase-web.org/Conference_Proceedings.php.

Parikh, H. (2016) and ongoing. *Tuva*. New York, NY: Tuva. (http://www.tuvalabs.com)

And once again, the link to step-by-step instructions: http://bit.ly/TEE2016Berlin.