# Data Management Briefing Memo

Andrew McCartney

## Introduction

Because of my long-held interests in global studies, international development, cross-cultural exchange, and education, I took this project as an opportunity to find the overlap in the Venn diagram of all of these and to explore a bit. Having a quantitative mindset in these fields is something of a rarity, and I admit that part of my relish in putting together these particular datasets is that the idea of quantifying cultural factors such as Hofstede (Hofstede, 1983; Hofstede & Minkov, 2013) has done will ruffle some methodological feathers. In particular, I wanted to see which educational features could be associated with these cultural factors, and because of the simultaneous interest in development, I brought in economic data for this as well.

As is perhaps not the best practice, I went into this exercise agnostic with respect to any particular hypotheses but rather to see which relationships I could find by exploratory data analysis. Some of the results were surprising, but most fall in line with what might be considered predictable for this kind of research. Unfortunately, the nature of the phenomena being studied, and by extension the data, preclude much more thorough research, as will be discussed herein.

## Were these files easy to use?

I’ll begin by describing the easier of the data sets: Hofstede’s work was easily available via his personal website (Hofstede & Minkov, 2013). Prior to the beginning of this project, I had done some minor cleaning of this file on my own, namely to add a variable (“is region”) to mark those observations which were not at the national level. Likewise, the dataset the Human Development Index required some excel work to make it strictly rectangular variable-observation data, but this was done with absurd ease.

Having done those both, I moved on to the data for the millennium development goals (World Health Organization, n.d.) and sustainable development goals (United Nations, n.d.). Such indicator data taken from the main UN data browser arrive in an extreme long-format, with the variable of interest given as “value” and then some meta-data about that variable such as whether it is a ratio, a percentage, etc. and what country and year it represents. I extracted 41 such datasets. Upon unzipping them, I renamed each file because the standard filename upon download is a timestamp given down to the millisecond in Greenwich Mean Time, which precluded any useful looping in stata. Rather, I gave each a single integer designation to make for easier loops later. Originally, I also renamed the “value” of each of these to the indicator listed by the UN, but I decided this was too easy and was begging to be looped through, so I renamed them all “value” again so I could use such a loop. I will detail this process below.

## Well Documented?

At no point in this process did I download documentation; however, on reflection documentation does exist for most of the UN data, but is linked in fine print below the main data browser on each page. Above this, I just noticed, is a link for downloading the entire Millennium Development Goal Dataset. However, its formatting is so frustrating I scarcely regret missing it until now.

Having said this, documentation for Hofstede’s Cultural Dimensions is limited, in part because it is partially used as a commercial product for businesses, having started out as a tool in industrial-organizational psychology. I did not obtain his book or test documentation for this project, and that I think is a serious flaw if I wish to continue to use these data going forward.

## Internally Consistent?

For the most part, data were internally consistent. Sometimes percentages were above 100%; this happens for example with school enrollments because they are calculated values based on attendance figures and population figures which sometimes overestimate the number of students. Likewise, Hofstede’s indices range from zero to 100, but occasionally have values above 100 for reasons unknown to me. His website offers datasets that force all values to be less than or equal to 100, but I didn’t like the idea of missing those data and assumed I could subtract the extra if I ever needed to.

## What issues did I confront?

The main issues I confronted stemmed from the nature of the UN data sets. At first, I renamed many of these variables by hand, and I consistently had to re-name each of the files individually, as well. I also think that, due to the way in which I compiled the data, someone following my .do files from their own raw data (not the ones on my flash drive) would have to manually drop additional fields, as well. One of the greater annoyances was the fact that the UN data do not utilize country codes, which meant that every time I switched from millennium to sustainable development goals or to the HDI or Hofstede I had to re-name about a dozen countries, e.g. “Korea, Republic of” for “Korea (Republic of).” When trying to loop through country names at the end of the data visualization process, I realized I still had two groups of Moldovas because of this problem, for example. And neither of them was Transnistria. Finally, the variable names I took from the UN indicator names are so long that stata begins to obscure what each variable is meant to be, which will surely frustrate any follower-researchers, though I’m not sure how to get Stata to allow for the full labels.

In general, I also have some issues with Stata not behaving in ways that I expect. Part of this is because I have grown used to using a language that “thinks” in vectors rather than looping. R, for example, is more efficient when applying something to all elements of a vector; languages like Julia and Stata are otherwise suited to loops. Stata’s interpretation of what a loop should look like is somewhat out of sync with other programming languages, so adjustment was required.

There were some syntactical issues that prevented full exploration; for example, at one point I loop through selected African countries; despite a few days’ worth of trying, I still haven’t quite figured out how to define a macro over strings with embedded spaces—this is also a problem for me at work, where I am re-writing my data management procedures into stata for posterity/clarity/consistency. I’ll need to solve it soon.

## How could these data be improved to make them more accessible, more accurate, more user-friendly for future projects?

My main suggestion would be to have a more interactive browser or API for these data at the UN’s website. I have the option to download single variables one-by-one, or to download the entirety of the MDGs all together? There’s got to be a better way to select which fields are desired (exoplanets.org has precisely this setup) and exclude all others. Unfortunately, the bigger problem with these data is that they’re just *missing*. Some indicators haven’t yet been collected, others never were collected or couldn’t have been. For example, Somalia literally has no functioning national school system, and estimating the primary completion rate for girls, e.g., is essentially impossible: hence, no Somalia data for that indicator, or most others either.

## What challenges did I encounter, and how did I solve them?

These were myriad, but often rewarding to figure out a solution, even if it is difficult to tell if it is the “correct” or most efficient solution. For a small example, when trying to run all graphs consecutively, graphs that worked individually would generate an error “missing )” when run as a whole do file. To solve this, I made each one of my graphs its own subroutine / do file, which of course generated another problem: the changing of directories for all of these meant that they couldn’t all run simultaneously, but *this I was able to solve* by changing directories at the end of every individual do file but the master.

Other things were minor but enormously relieving; for example, the y ~ x…. equations given in the plots I’ve used are generated automatically as subtitles within a user written package. To change the subtitle size to allow for faceting requires not that you write subtitle(size(small)) because that *prints* “size(small)” but instead to enter size as a second argument as “subtitle(,size(small)).” I tried this because it was the intuitive thing to do and had no expectation that it would work, but behold. I’m not sure if that says more about me “thinking in stata” than having the same intuition as the individual who wrote that package, but it worked and I was delighted.

## Highlight your two greatest data management achievements from doing this project.

As I mentioned in my presentation, I knew I would need to find a way to loop through a local to change the variable names from “value” to the actual variable. It turns out the way to do this is to use “parallel locals” which involve calling two simultaneous locals, one referencing the other, then nesting them within one’s for loop. Code for this is reproduced in Appendix A. I’m particularly fond of this example because I was able to use capture to widen 10 datasets within this loop without having to pull them from the block to deal with them as a separate unit. Ultimately, this block of code required a local for: Variable names, the number associated with the total words of *that* local, a forvalues macro within the loop, and a local within the loop to specify the correspondence between the names macro and the count macro.

Here’s one more achievement: looping through countries to make a large number of faceted graphs. As noted above, I would have preferred to use more African nations (or I could do Eastern Europe next time) but given the multi-word nature of many of their names, I restricted myself. Code for this is in Appendix B. This took an enormous amount of time to get right even with that modification (when looping with strings, when to use “” is not yet intuitive for me). I was able to loop through years well enough but that generated uninteresting data (one to many merge issues). I also wanted to call the graphs in the combine as a loop themselves, but stata did not allow for this. Even after I got the loop to work, getting the graph export function to work correctly took some googling. However, I am satisfied-ish with the result, and I’m intrigued by what else I could use it for.

Ultimately, I took this as an opportunity for skills practice with a dataset that is inherently interesting for me. I enjoyed tackling the problems, though I wish the data were more complete. As noted above, I’m now in the process of transferring these skills to less exciting but more impactful work on our grant project to make sure that a consistent flow of data usage is in place for the project if I ever leave (as it stands, I’m the only one who ever looks at our quantitative data) and the practice of this has raised similar problems as this project, so it’s been a delight to tackle them.

References

Cheung, H. Y., & Chan, A. W. (2008). Relationships amongst cultural dimensions, educational expenditure and class size of different nations. *International Journal of Educational Development*, *28*(6), 698-707.

Hofstede, G. (1983). National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization*, *13*(1-2), 46-74.

Hofstede, G., & Bond, M. H. (1984). Hofstede's culture dimensions: An independent validation using Rokeach's value survey. *Journal of cross-cultural psychology*, *15*(4), 417-433.

Hofstede, G., & Minkov, M. (2013). VSM 2013. *Values survey module*.

United Nations. (n.d.). *Sustainable Development Goals.* retrieved from http://www.un.org/sustainabledevelopment/sustainable-development-goals/

World Health Organization. (n.d.). *Millennium Development Goals (MDGs).* Retrieved from http://www.who.int/topics/millennium\_development\_goals/about/en/

### Appendix A

local sdgnames "low\_sec\_profic\_math low\_sec\_profic\_read prim\_profic\_math prim\_profic\_math pre\_prim\_particip any\_educ\_train\_12mo gender\_parity\_numeracy gender\_parity\_literacy econ\_parity\_literacy econ\_parity\_numeracy gend\_parity\_tch\_train\_pre\_prim gend\_parity\_tch\_train\_prim gend\_parity\_tch\_train\_up\_sec computers\_low\_sec electricity\_low\_sec drinking\_h20\_low\_sec internet\_low\_sec sss\_low\_sec computers\_prim electricity\_prim drinking\_h20\_prim internet\_prim sss\_prim computers\_up\_sec electricity\_up\_sec internet\_up\_sec sss\_up\_sec tch\_trn\_low\_sec tch\_trn\_pre\_prim tch\_trn\_prim tch\_trn\_up\_sec"

local m : word count `sdgnames'

forvalues p = 1(1)`m' {

use "UNdata\_Export\_SDG`p'\_m", clear

local b: word `p' of `sdgnames'

rename value `b'

capture reshape wide `b', i(countryorarea year) j(sex) string

save UNdata\_Export\_SDG`p'\_renamed, replace

}

// I'm pretty proud of this one I'm not gonna lie.

### Appendix B

// let's do this for just a few countries of interest:

local nation "Angola Benin Botswana Burundi Cameroon Chad Gambia Lesotho Liberia Malawi Rwanda Sudan Tunisia Uganda Zambia"

foreach q in `nation' {

twoway (scatter pr\_en\_girls year)(qfit pr\_en\_girls year) if countryorarea == "`q'", name(`q'\_scatter, replace) legend(off) title("`q'")

}

#delimit;

graph combine

Angola\_scatter Benin\_scatter Botswana\_scatter Burundi\_scatter

Cameroon\_scatter Chad\_scatter Gambia\_scatter Lesotho\_scatter

Liberia\_scatter Rwanda\_scatter

Tunisia\_scatter Uganda\_scatter Zambia\_scatter,

col(4)

title( "Primary Enrollment of Girls" )

subtitle( "For select African Nations, 1990-2014" )

scheme(s1mono)

;

graph export african\_nations\_girls.png, replace