MEASURE Evaluation

March 2017

Frequently Asked Questions about Geographic Information Systems

Tidy Data: The Key to Success with Spatial Data

Tips on Data Structure

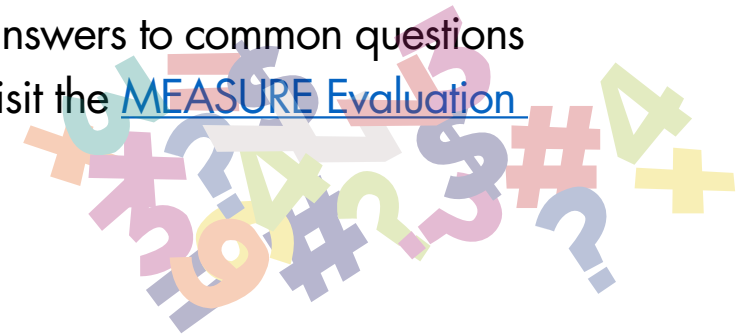
Preface

Most work with geographic information systems (GIS) revolves around data. Before a single map is made, a considerable amount of work is usually necessary to make sure the data is mappable.

In the early days of GIS, the software's handling of data was inflexible. Most programs would accept data in only a few narrowly specific file formats. For instance, they could not read data directly from spreadsheets. Modern GIS software can accommodate many more formats.

However, although GIS is more forgiving now with regard to different file formats, it still has strict requirements regarding the structure of data. No matter the file format, GIS software expects the data to conform to basic standards of tidy data.

This FAQ presents basic information on the concept of tidy data and how GIS rely on it. It is one in a series of FAQs on important topics that are relevant to GIS and spatial data. These FAQs are intended to provide brief answers to common questions and steer you to sources of more detailed information. Visit the MEASURE Evaluation website to read more about GIS.



What does "file structure" mean? How is it different from "file format"?

File format typically refers to the specific format a software program uses to save or open a file. For instance, Microsoft Excel's default file format is .xsl (or .xslx). In addition to containing the actual data (for example, the numbers in the spreadsheet or the words in a document), the file format also contains metadata information. A file format will store both the metadata and the actual content in a standard way that a program can read.

File structure refers to the way the data are stored in the file. This is typically relevant for numeric data such as would be found in a spreadsheet.

What does "tidy data" mean?

"Tidy data" refers to a standard structure for data. As the name suggests, tidy data are data that have been organized in an orderly structure that makes them easier to use.

Tidy data was first described in a paper by Hadley Wickham: "Tidy Data" (Wickham, 2014). In the paper, he presents three fundamental concepts of tidy data:

1) Each variable forms a column.
2) Each observation forms a row.
3) Each type of observational unit forms a table.

Class	
Mammal	**Number of feet**
Horse	4
Dog	2
Cat	4
Reptile	**Number of Feet**
Snake	0
Turtle	2
Bird	**Number of Feet**
Eagle	2
Ostrich	2

In this table, the left column has multiple data types–classes such as "Mammal or Reptile" as well as species such as "Horse or Dog."

There are also rows in the spreadsheet that are blank. This spreadsheet may be easy for a human being to read and interpret, but a GIS will not be able to use data structured like this.

Here is an example of what a table that conforms to tidy data standards should look like.

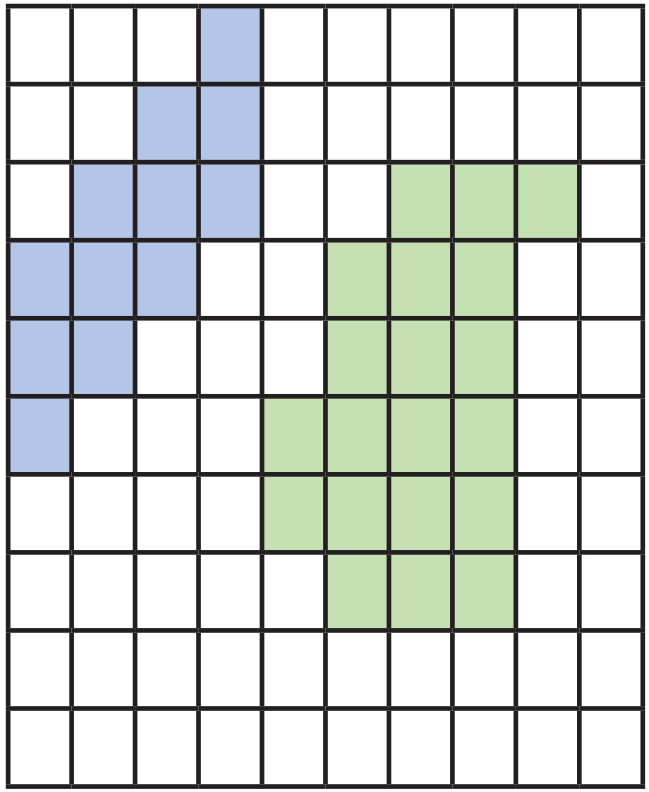
Animal	Number of feet	Class
Horse	4	Mammal
Eagle	2	Bird
Turtle	4	Reptile
Dog	4	Mammal
Snake	0	Reptile
Ostrich	2	Bird
Cat	4	Mammal

How does this relate to GIS?

A GIS stores data in two formats. One of these is raster. **When data is in raster format, the data values are stored as a matrix of data values that correspond to locations on the earth.**

0	0	0	1	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0
0	1	1	1	0	0	2	2	2	0
1	1	1	0	0	2	2	2	0	0
1	1	0	0	0	2	2	2	0	0
1	0	0	0	2	2	2	2	0	0
0	0	0	0	2	2	2	2	0	0
0	0	0	0	0	2	2	2	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

When raster data is displayed in a GIS, it will look like this.



Each value in the matrix corresponds to a data value at a geographic location. The structure of the data is a simple matrix of numbers. An example of raster data would be a satellite image. The image is made up of many cells and each cell contains a value that corresponds to the reflectance of what is on the ground at that location.

Because raster data adheres to a very concrete structure, the concept of tidy data does not apply to raster data.

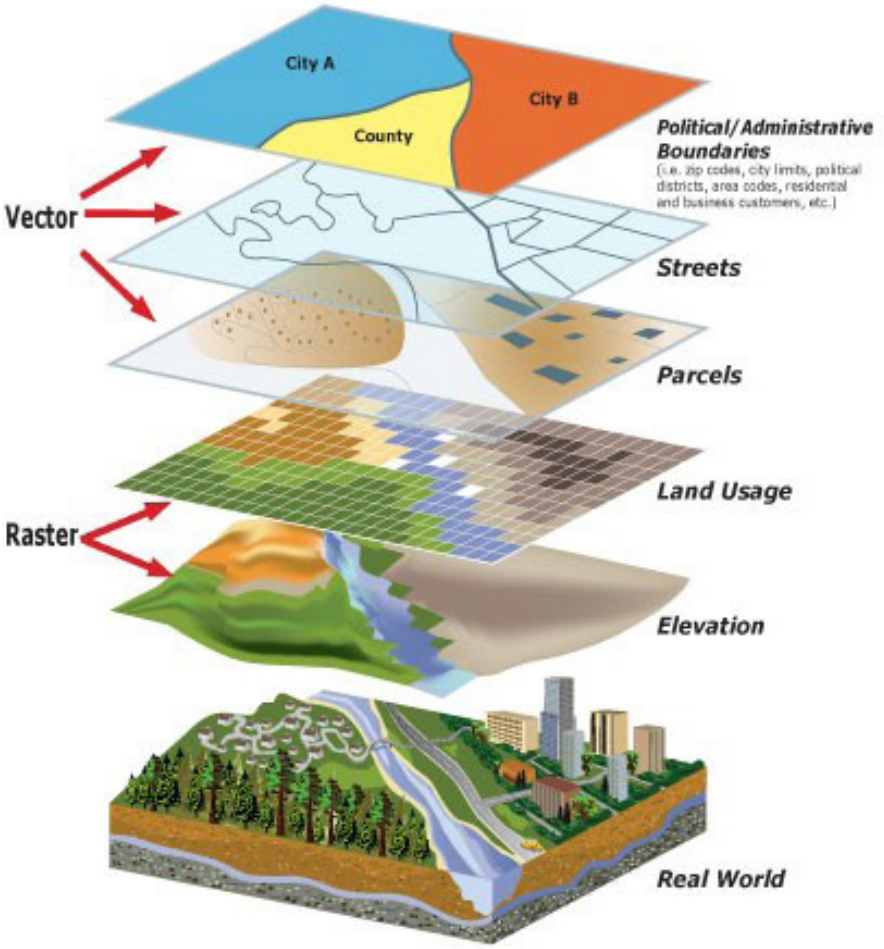
The other way that a GIS stores data is in vector format. When data is stored in vector format, the data exists in a data table where each record corresponds to a geographic location and the data values are stored in columns. When working with vector data, a GIS sees that each record is its own individual geographic entity to be mapped.

In other words, vector data must be tidy data.

Facility ID	Name	Latitude	Longitude	Number of staff
3K4R200	Eastern Health Clinic	-47.48516	61.69449	13
27LS611	Southern Health Clinic	-6.05422	19.66357	4
1N291B2	Western Health Clinic	-48.36875	109.76463	9

The table above is an example of how a vector data set would typically be structured. Each record corresponds to a health facility and each column corresponds to an attribute of the facility. (GIS Commons, n.d.)

A GIS can layer both raster and vector data together.



IV How can I make my data tidy?

The complexity of your data and the specific way it may be stored in a spreadsheet will determine the exact process for making data tidy. It may be necessary to significantly restructure the data to make them tidy. Unfortunately, this is typically a manual process that can require a considerable amount of time.

However, there are some resources that can make the process a little easier. A list of some of the available tools are available at the end of this FAQ.

V

Once it's tidy, will my GIS be able to use my data?

Having tidy data will make it much easier for a GIS to interpret your data. However, some data cleaning may still be necessary. For instance, the GIS software may have requirements about the way that variables are named, such as no spaces; can't start with numbers; words that are reserved and can't be used as a variable name. There may also be issues around whether the GIS considers a variable to be a text variable or a numeric variable.

Reviewing the documentation of the software you use will help identify the proper structure.

VI Is this only applicable to GIS?

No. Following basic tidy data protocols will make analysis with many other software programs easier to do. Data visualization programs, statistical analysis programs, or spreadsheets will all be able to read tidy data easily.

VII Where can I learn more?

To learn more about tidy data and well-formed data structures, you'll find that the original "Tidy Data" article (Wickham, 2014) provides a good introduction. Code libraries have also been created for the programming languages R and Python that can help make data tidy. A search of the web will find these.

Nicholas Hould has an overview of tools in Python programming language (Hould, 2016).

Stata provides tools; an overview of some of them can be found here: http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial (Carolina Population Center, n.d.)

Excel is not necessarily the best tool to change untidy data into tidy data, but there are some things it can do. Microsoft has a page describing how to clean data and offers some plugins that could be helpful: https://support.office.com/en-us/article/Top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19 (Microsoft, Inc., n.d.)

A good overview of some useful Excel functions can be found here: http://myexcelonline.com/blog/top-excel-data-cleansing-techniques/ (My Excel Online, n.d.)

VIII References

Carolina Population Center. (n.d.). *Introduction to Stata*. Retrieved from http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial

GIS Commons. (n.d.). *An introductory textbook on geographic information systems*. Retrieved from http://giscommons.org/

Hould, N. (2016). *Tidy data in Python*. Retrieved from http://www.jeannicholashould.com/tidy-data-in-python.html

Microsoft, Inc. (n.d.). *Top ten ways to clean your data*. Retrieved from Microsoft Office Support: https://support.office.com/en-us/article/Top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19

My Excel Online. (n.d.). *Top Excel data cleaning techniques*. Retrieved from http://myexcelonline.com/blog/top-excel-data-cleansing-techniques/

Wickham, H. (2017). Tidy data toolkit. Retrieved from ftp://cran.r-project.org/pub/R/web/packages/tidyr/vignettes/tidy-data.html

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10). Retrieved from http://www.jstatsoft.org/article/view/v059i10

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, North Carolina 27517
Phone: +1 919-445-9350 | Fax: +1 919-445-9353
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. SR-17-142

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation