WORKING PAPER

Assessing Spatial Data Quality

Using Five Data Anomalies

Speeding the Process for Master Facility Lists and Other Large Data Sets

January 2019

USAID
FROM THE AMERICAN PEOPLE

PEPFAR

MEASURE Evaluation

WORKING PAPER

Assessing Spatial Data Quality

Using Five Data Anomalies

Speeding the Process for Master Facility Lists and Other Large Data Sets

John Spencer, MA, MEASURE Evaluation

Becky Wilkes, MS, MEASURE Evaluation

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350
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. WP-19-227

USAID FROM THE AMERICAN PEOPLE

PEPFAR

MEASURE Evaluation

ACKNOWLEDGMENTS

The authors wish to thank Kristen Wares, Gina Safarty, and Ana Scholl, of the United States Agency for International Development (USAID), and Dr. Nathan Heard, of the United States Department of State, for their early input and help defining the general problems of assessing spatial data quality for master facility lists.

We would also like to thank the knowledge management team at MEASURE Evaluation—funded by USAID and the United States President's Emergency Plan for AIDS Relief and based at the University of North Carolina at Chapel Hill—for editorial, design, and production services. We also thank Cindy Young-Turner at ICF International for editing it.

CONTENTS

ABBREVIATIONS

GPS	global positioning system
MFL	master facility list
PEPFAR	United States President's Emergency Plan for AIDS Relief
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

With the increased ease of the collection of geographic data coordinates and the desire for accurate country master facility lists (MFLs) comes the need for tools and methods with which to rapidly assess the quality of large spatial data sets. Global health professionals who have had limited training in the use of geographic information systems may need guidance in assessing spatial data. Identifying data quality issues in data sets of this size is challenging, because of the complex relationship between the spatial components and the attributes of the data.

Informed by spatial data quality literature, this paper presents a framework for assessing common issues with spatial data and identifies five specific potential data anomalies that can be identified and further investigated to increase the quality of a spatial data set, such as an MFL. Focusing on these five anomalies will provide quantifiable results, which help in planning a practical, effective strategy for corrections. This approach yields not only a list of the locations that need to be corrected, but also feedback on what may be wrong with the data.

INTRODUCTION

Ease of Geographic Data Collection Means an Increase in Data for Monitoring and Evaluation

It is easier than ever before to collect geographic coordinates. Only a few years ago, obtaining a latitude/longitude coordinate of a location frequently required the use of a dedicated Global Positioning System (GPS) receiver and specialized software and cables to download the collected data points and use them in a spreadsheet or mapping program. The current state of the art is much different. Smartphone and tablet users can collect location data with little effort, and many applications are available to leverage those data and facilitate their use. This ease of collection has resulted in an increase in the creation of geocoded data sets. For global health professionals, the ease of collecting geographic coordinates means that it is possible to store and maintain databases with accurate locations of features that may have an impact on health. Health facilities, disease outbreaks, wells, schools, and markets can now easily be given geographic locations that are stored in a spreadsheet or database. These databases can then be used to support monitoring and evaluation and inform decision making.

Master Facility Lists and Other Spatial Data Sets

National MFLs are an example of a database that frequently stores geographic data. These lists are databases that typically contain the locations of a country's health facilities and information on those facilities. They are usually created and maintained by a national health ministry or combination of ministries. Because they contain a record for every health facility in the country, the resulting database can be quite large, with many thousands of records.

In addition to national MFLs, many organizations are building their own data sets of program locations or other sites of interest, such as facilities, schools, and wells. A nongovernmental organization in a moderately sized country could be operating in hundreds if not thousands of locations, and all of these geographic data can be compiled in data sets that can be quite large.

Components of Data Quality

Data quality involves multidimensional components, so assessing a data set with thousands of records is a challenging task. Spatial data sets of this size are increasingly common in global health, and some national MFLs have tens of thousands of data points. It is impractical to review every record individually in such large databases; such an effort would take hundreds, potentially thousands of hours. It is essential, however, that the data have minimal errors.

Identifying data quality issues in data sets of this size is also challenging because of the complex relationship between the spatial components and the attributes of the data. Determining whether a record is located in the correct spot and contains the correct attributes is more complex than it might seem at first glance. To illustrate, simply taking the data points and plotting them on a map will identify coordinates that are clearly incorrect, such as ones that appear outside of the country or in the middle of a lake. This approach, however, will miss

other errors, such as sites that should be in one district but show up in another. For such cases, the problem could be either that the coordinate is incorrect or that the field containing the district name is incorrect.

Literature on Spatial Data Quality

There is a considerable body of literature on data quality and spatial data quality in general, but most of it focuses on issues of spatial accuracy—not broader issues of data quality. Spatial accuracy concerns how accurately a coordinate corresponds to the true location on Earth. In other words, spatial accuracy involves measuring whether a location is correct within a tolerance of a specific distance (Ahmand, 2014).

Spatial accuracy is an important consideration of overall data quality, but it is not the only consideration. Accuracy of the other attributes in the data set and whether there is agreement between the spatial coordinate and the other attributes are also important (Hong & Huang, 2016).

An article on the creation of spatial data quality analysis tools by Devillers, Bedard, and Jeansoulin notes that as the volume of data increases and we lose our ability to efficiently analyze all the information, we will increasingly require "new analytical and visualization tools capable of providing humans with a logical summary of the uncertainty of information present in the system" (Devillers, et al. 2007).

In attempting to create just such a tool, the authors have focused on certain data anomalies that can be discovered and examined in large spatial data sets, to help assess data accuracy within both the spatial domain and the attribute domain, and also agreement between the spatial domain and the attribute domain. This agreement is also known as *logical consistency* and has been noted as an important aspect of data quality assessment by several authors (Girres & Touya 2010), (Devillers & Jeansoulin 2010).

DISCUSSION AND METHODS

A Framework for Efficiently Assessing Data Quality

A framework for assessing the data quality of spatial databases is needed to identify data quality issues as effectively and efficiently as possible. This framework can help guide the quality assessment process and ensure consistency. It should be clear, easy to apply, and repeatable.

Any framework for assessing data quality should consider the accuracy and precision of spatial data, the accuracy and thoroughness of attribute information, and—just as important—the practicality of assessing all the elements. These principles can be formalized as follows:

- The coordinates should have a degree of spatial accuracy appropriate to the intended use of the data.
- The coordinates should have spatial precision that is adequate for the intended use of the data.
- The spatial domain and the attribute domain should agree.
- The attribute domain should be complete and accurate.

Although there would likely be little argument about these principles, the challenge is how to apply them in a practical manner when assessing large spatial data sets.

Data Anomalies

This paper uses the above principles, which are important for assessing data quality, to discuss five specific anomalies in geocoded data. These anomalies can be used as a framework to inform a spatial data quality assessment effort in a practical manner. Data can be evaluated to identify the presence of any of these anomalies, and then the anomalous records can be investigated to determine whether a data quality issue needs to be addressed. Focusing on records most likely to have an error will greatly reduce the time and effort necessary to identify and resolve errors in the database, compared to a systematic review of every record. This method may not identify every record that has incorrect information, but it will systematically identify records that should be investigated to resolve the anomalies. Owners of spatial data sets can then employ other methods to identify records with other errors. They may decide to spot-check randomly selected records, or they may decide to allow errors to be found as the database is used and then employ whatever protocol may be in place to correct the errors.

Focusing on these five anomalies will provide quantifiable results, which help in planning an effective strategy for corrections. This approach yields not only a list of which locations need to be corrected but also specific feedback on what may be wrong with the data.

Data Domains

In geocoded data, there are two data domains: *spatial* information and *attribute* information. Data quality is a consideration for both.

The spatial information domain refers to the accuracy and precision of the actual geographic coordinates recorded—in other words, the data that place a site at a specific location on the Earth. Key questions for ensuring the quality of data in the spatial domain are the following: Do the coordinates accurately locate the

phenomenon on the Earth? Are they in an accepted format? Do they have a sufficient number of significant digits to provide the precision needed?

The attribute information domain refers to the data associated with the coordinate, such as name, address, or other characteristics of the location. Key questions for ensuring the quality of data in the attribute domain are the following: Is the correct value recorded for attributes of interest (i.e., site name, number of beds, and unique ID)? Is a consistent schema employed?

RESULTS

Five Anomalies Defined

Attempting to determine whether a record contains errors either in the spatial or attribute domain requires a systematic approach to reviewing variables across every record. One approach would be to check every record individually, validating each location and the values in each variable. This approach, which is commonly employed, becomes unwieldly with any but the smallest data set.

A more efficient approach is to reduce the number of records and focus only on those most likely to have data quality issues. This can be done by identifying anomalies in the data that may indicate data quality issues and prioritizing the investigation of those records. Scripts can be created in a geographic information system or in a programming language such as R or Python to identify records with these anomalies. Excel functions can also be employed, although Excel is not as robust a tool in terms of being able to efficiently identify anomalies.

We propose focusing on the following five anomalies, which, if present, can indicate possible data quality issues:

- Missing coordinates
- Obvious flaw in a coordinate, or lack of precision
- Duplicate coordinates for distinct places
- Duplicate key attributes
- Coordinate not where it would be expected

These five anomalies cover both the spatial and attribute domains. In our experience, these anomalies identify the most common data quality issues with discrete location data. Identifying records with these anomalies is only one step in assessing data quality. Doing so will not necessarily uncover every error in the database, but it will disclose the clear issues.

The five anomalies are described in more detail, as follows.

1. Missing coordinates

If the data do not contain coordinates, then it is not possible to locate the site. Identifying sites that lack coordinates can be done using a simple query that looks for missing values in the coordinate field.

2. Obvious flaw in a coordinate, or lack of precision

Sometimes data quality issues can be detected by looking at the coordinates. Typos can be identified if the coordinate is very different from other coordinates that are supposed to be nearby. The coordinate could be clearly out of range (latitude values beyond 89 or -89; longitude values beyond 179 or -179), or it could be noticeably different because of a typographic error. For example, the longitude values for all the other sites in the database may start with -79, but the value for one site is -97.

Another possible flaw in the coordinate is an insufficient number of significant digits. The number of significant digits included with a coordinate will greatly affect its precision. It is important to use the correct number of significant digits, to ensure that the database is as precise as required. There is no one-size-fits-all solution for the correct number of significant digits; the proper number will depend on the precision required.

For example, finding a specific corner of a property boundary might require sub-centimeter precision. That wouldn't be necessary to locate a large building such as a hospital, however, that level of accuracy is not necessary. Most GPS units, smartphones, and tablets collect a point with six decimal places, a level of precision much greater than is actually needed in the case of a health facility location.

This table shows the increasing effect of including varying numbers of significant digits of precision (Wikipedia, n.d.).

NOTE: This is at the equator and is applicable to Lat and Long. As you move further away from the equator those values change.

Table: Precision in geographic coordinates

Coordinate	Precision	Area that can be represented
32.0	111 km	Country or large region
32.01	11.1 km	Large city or district
32.01	1.1 km	Town or village
32.001	111.3 m	Neighborhood
32.0001	11.1 m	Individual street
32.00001	1.1 m	Individual trees
32.000001	111.3 mm	Individual humans

In terms of data quality, records in which the coordinate has fewer significant digits than required for the desired precision would likely be considered problematic.

3. Duplicate coordinates for distinct places

A record that has the same geographic coordinates as another record may indicate a data quality issue, but it is also possible that the location contains two items of interest. For example, a large medical center building may have multiple healthcare practices, and the coordinates recorded for those practices could be the same. These instances do not necessarily compromise the quality of the data set.

Instances of duplicate coordinates for sites that actually are located in two different places would be problematic.

4. Duplicate key attributes

Every record in a data set needs to be uniquely differentiated from other records. This can be done by many different methods. A unique ID can serve as a differentiator, as can a name, address, or phone number. The components of a data set that serve as differentiators are known as the signature domain (Health Facility Assessment TWG, 2017). Duplications in the signature domain can be an indication of data quality issues, but not always; multiple health facilities could have the same name but be in different locations. For example, multiple facilities may be called "King Francis Health Center," with each located in a different district.

5. Coordinate not where it would be expected

One of the checks to test data quality is to confirm that the coordinate is in the expected location. If a coordinate is supposed to be in a specific district, is that where it appears? If a coordinate is not in the expected location, the coordinate could be incorrect, or the attribute that lists the district name could be incorrect.

Cases in which the coordinate is in an unexpected location can be identified with the process used to identify the second anomaly (obvious flaw in the coordinate). A special effort is needed to identify those records that are not captured in that process, however. For example, a spatial join (a GIS operation that will match point features in one file to area features in another) can be used to determine whether the coordinate is in the district where it is expected to be.

CONCLUSION

A challenge with any data set is ensuring the quality of the data. With large data sets that contain hundreds or thousands of geographic coordinates, validating the spatial and attribute components of the data can quickly become an overwhelming task. Here we have presented five data anomalies that can serve as a framework for identifying records with data quality issues. Although additional data quality assessment will likely be required, this framework offers a starting point for examining a large data set efficiently. If these errors are identified early, the data set can then be systematically corrected and made stronger, more useful, and more trustworthy.

REFERENCES

Wikipedia. (n.d.). Decimal degrees. Accessed January 3, 2017. Retrieved from https://en.wikipedia.org/wiki/Decimal_degrees

Health Facility Assessment Technical Working Group. (n.d.). The signature domain and geographic coordinates: A standardized approach for uniquely identifying a health facility. Chapel Hill, NC, USA: MEASURE Evaluation, University of North Carolina. Retrieved from http://www.measureevaluation.org/resources/publications/wp-07-91

Ahmand, T. (2014). Spatial data quality. New Delhi, India: Indian Agricultural Statistics Research Institute.

Devillers, R., Bedard, Y., Jeansoulin, R., & Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science,* 21 (3): 261–282.

Devillers, R., & Jeansoulin, R. (2010). Fundamentals of spatial data quality. John Wiley & Sons, ProQuest Ebook Central.

Girres, J., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS,* 14: 435–459. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9671.2010.01203.x

Hong, J. H., & Huang, M. L. (2016). Function workflow design for a geographic information system: A data quality perspective. In S. Wenzhong, B. Wu, & A. Stein (Eds.), *Uncertainty Modeling and Quality Control for Spatial Data.* Boca Raton, FL, USA: CRC Press, Taylor and Francis Group.

WORKING PAPER

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350
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. WP-19-227