Geocoding Afrobarometer Rounds 1-6: Methodology & Data Quality

Ariel BenYishay
Renee Rotberg
Jessica Wells
Zhonghui Lv
Seth Goodman
Lidia Kovacevic
Dan Runfola

AidData

*Institute for the
Theory and Practice
of International Relations*

The College of William and Mary

abenyishay@aiddata.org
drunfola@aiddata.org

AidData.org

Preview

This document presents the methodology used to geocode the Afrobarometer (www.afrobarometer.org) database, consisting of public attitude surveys on democracy, governance, economic conditions, and related issues in more than 35 countries in Africa. This document is designed to give potential users of the data a transparent view into the methodology followed, as well as post-hoc assessments of the quality of the data produced.

Data Production

This geocoding follows a double-blind methodology AidData developed for geo-referencing development projects (Strandow et al. 2011); it was applied to geocode the location of Afrobarometer Enumeration Areas (EAs). Leveraging a team of trained geocoders, coders review project documentation and assign latitude and longitude coordinates for each EA location. For each EA, the coders also include standardized codes describing the type of location (e.g., administrative zone, school, road). Two enumerators use geo databases such as Geonames, Google Maps, and OpenStreetMap to find coordinates, while they review statoids, encyclopedias, Wikipedia, and government websites to confirm location hierarchy and location type when assigning latitude and longitude coordinates. If the two enumerators disagree, the project is moved into an arbitration round. This approach captures geographic information at several levels — coordinate, city, and administrative divisions — for each location.

AIDDATA
A Research Lab at William & Mary

Table of Contents

Acknowledgements

This work was made possible through the vital collaboration of Afrobarometer and financial support from USAID, the Hewlett Foundation and the College of William and Mary. We would also like to thank the many students at William and Mary who contributed to this data, including Allison Bowers, Kaitlan Shaub, Kate Connors, Ava Chafin, Lidia Kovacevic, Emily Mudd, Grace Grimsley, Niña Francesca Ledonio, Ryan Haarstick, Trevor Hampton, Xinyi Wang, Samuel Patecell, Marilyn Ma, Zoë Connor, Ania Blaszkiewicz, Liz Hall, Cara Vetsch, Feng Li, Ifechi Ugwu, Jennifer Steinhilber, Madison Free, Vishaal Mittal, Matthew Horrigan, Rose Olwell, Joanna Tan, Mackenzie Morrow Foster, Samantha Wooley, Emy Urban, Daniel Gurley, William Sheahan, Mitchell Dunlap, Ellie Bentley, Congyu Wang, Edward Chung, Cyrus Tidwell, Richard Tischner, Marylin Ampadu, Sadie Peloquin, Emily Holtzman, Alexandra Quinn. This work was performed in part using computational facilities at the College of William and Mary which were provided with the assistance of the National Science Foundation, the Virginia Port Authority, Virginia's Commonwealth Technology Research Fund and the Office of Naval Research.

About AidData

AidData is a research and innovation lab located at the College of William & Mary that seeks to make development finance more transparent, accountable, and effective. Users can track over $40 trillion in funding for development including remittances, foreign direct investment, aid, and most recently US private foundation flows all on a publicly accessible data portal on AidData.org. AidDta's work is made possible through funding from and partnerships with USAID, the World Bank, the Asian Development Bank, the African Development Bank, the Islamic Development Bank, the Open Aid Partnership, DFATD, the Hewlett Foundation, the Gates Foundation, Humanity United, and 20+ finance and planning ministries in Asia, Africa, and Latin America.

License for Use

Recommended Citation

BenYishay, A., Rotberg, R., Wells, J., Lv, Z., Goodman, S., Kovacevic, L., Runfola, D. 2017. *Geocoding Afrobarometer Rounds 1-6: Methodology & Data Quality.* AidData. Available online at http://geo.aiddata.org.

Summary & Motivation

Academic researchers, policy analysts and other practitioners are increasingly turning to geospatial data to enable new research; following this demand, many organizations such as AidData, Terrapop, CIESIN, and DHS have begun collecting geospatial data or adding geospatial information to existing products. This document summarizes the addition of geospatial information to the Afrobarometer (www.afrobarometer.org) survey rounds, consisting of public attitude surveys on democracy, governance, economic conditions, and related issues in more than 35 countries in Africa. This document is designed to give potential users of the data a transparent view into the methodology followed, as well as post-hoc assessments of the quality of the data produced.

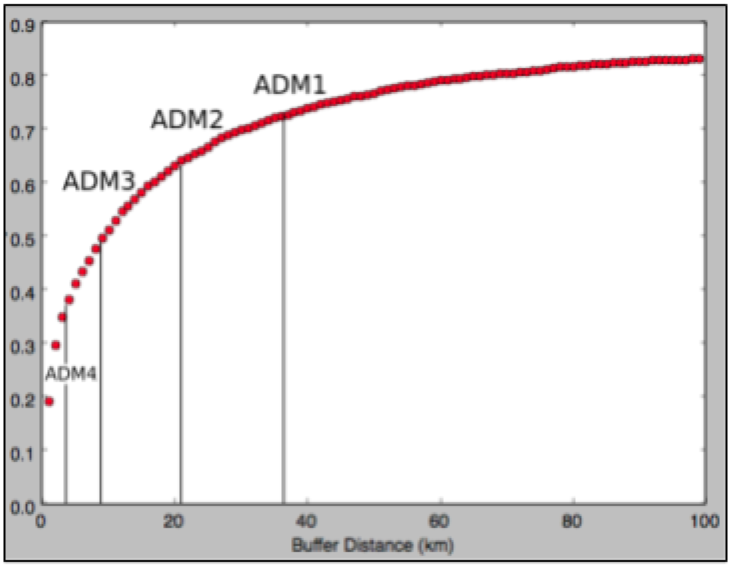
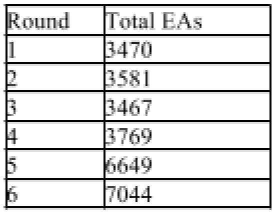


Figure 1. Comparison of Auto and Human Coding. The X-axis represents the size of a circular buffer extending from every human-coded location. The Y-axis shows the percentage of auto-coded locations that fall within the buffer.

In addition to geocoding these locations, we also carried out comparisons of our human geocoding approach to automated coding approaches (specifically, those employed by Knutsen et al. 2016). The comparison was conducted using Round 5 of Afrobarometer, which was fully auto and human coded across Africa. Findings suggest that autocoding may be sufficient for relatively coarse-scale analyses, but has considerable limitations for detailed sub-national analyses. These results are summarized in Figure 1, with vertical lines of commonly used analytic units included as points of reference — i.e., an analysis seeking to use ADM-2 could optimistically expect around 60% accuracy in geocoded locations, as contrasted to human coded approaches. Given that this accuracy may not be high enough for all sub-national analysis, the geocoding process summarized in this document provides substantial improvement in locational precision.

Geocoding Overview



Round	Total EAs
1	3470
2	3581
3	3467
4	3769
5	6649
6	7044

Table 1. EAs coded by round.

AidData sought to geocode "Enumeration Areas" for each Afrobarometer survey location (typically, populated places). A summary of the locations identified by round is provided in Table 1. This geocoding follows a double-blind methodology AidData developed for geo-referencing development projects (Strandow et al. 2011). Leveraging a team of trained geocoders, the geocoding methodology relies on a double-blind coding system, where two experts employ a defined hierarchy of geographic terms and independently assign latitude and longitude coordinates and standardized place names to each enumeration area. The two enumerators use geo databases such as Geonames, Google Maps, and OpenStreetMap to find coordinates, while they review statoids, encyclopedias, Wikipedia, and government websites to confirm location hierarchy and location type when assigning latitude and longitude coordinates. If the two enumerators disagree, the project is moved into an arbitration round where a geocoding project manager reconciles the codes to assign a master set of geocodes for all of the locations described in the available project documents. This approach captures geographic information at several levels — coordinate, city, and administrative divisions — for each location, thereby allowing the data to be visualized and analyzed in different ways depending upon the geographic unit of interest.

After enumerators identify the locations of Enumeration Areas (EAs), AidData performs many procedures to ensure data quality, including de-duplication of projects and locations, correcting logical inconsistencies, finding and correcting field and data type mismatches, correcting and aligning project locations within country and administrative boundaries, and validating place names and correcting gazetteer inconsistencies.

Geocoding Procedures Unique to Afrobarometer

Different enumeration areas were coded as exact or approximate depending on a variety of criteria. Exact geocodes were identified when the location an enumerator coded had exact and known administrative boundaries. Therefore, most populated places, ADMs, and Independent Political Entities (IPEs) are coded as exact.[1]

AidData used a hierarchy of place names provided by Afrobarometer, but additional project documentation to validate these locations was not available. As such, other than due diligence such as checking against third-party sources, relatively little information was available to validate the correct locations.

In some cases, individual EAs straddled multiple locations (sampling of respondents was carried out across two uniquely-named locations). In such cases, both locations were coded to the same EA. Thus, some EAs in the data have two sets of location-specific values.

Quality Assurance & Data Quality Assessment Procedures for Afrobarometer

In the case of Afrobarometer, two key steps of quality assurance are undertaken. First, data is deduplicated to ensure that each Afrobarometer EA is represented only once in the final data product. Second, each spatial field is checked for consistency — i.e., it is ensured that each EA location fell into the country in which the survey was conducted. In the case of Afrobarometer, the three-letter country code associated with each EA ID was used to confirm the spatial accuracy of each location.

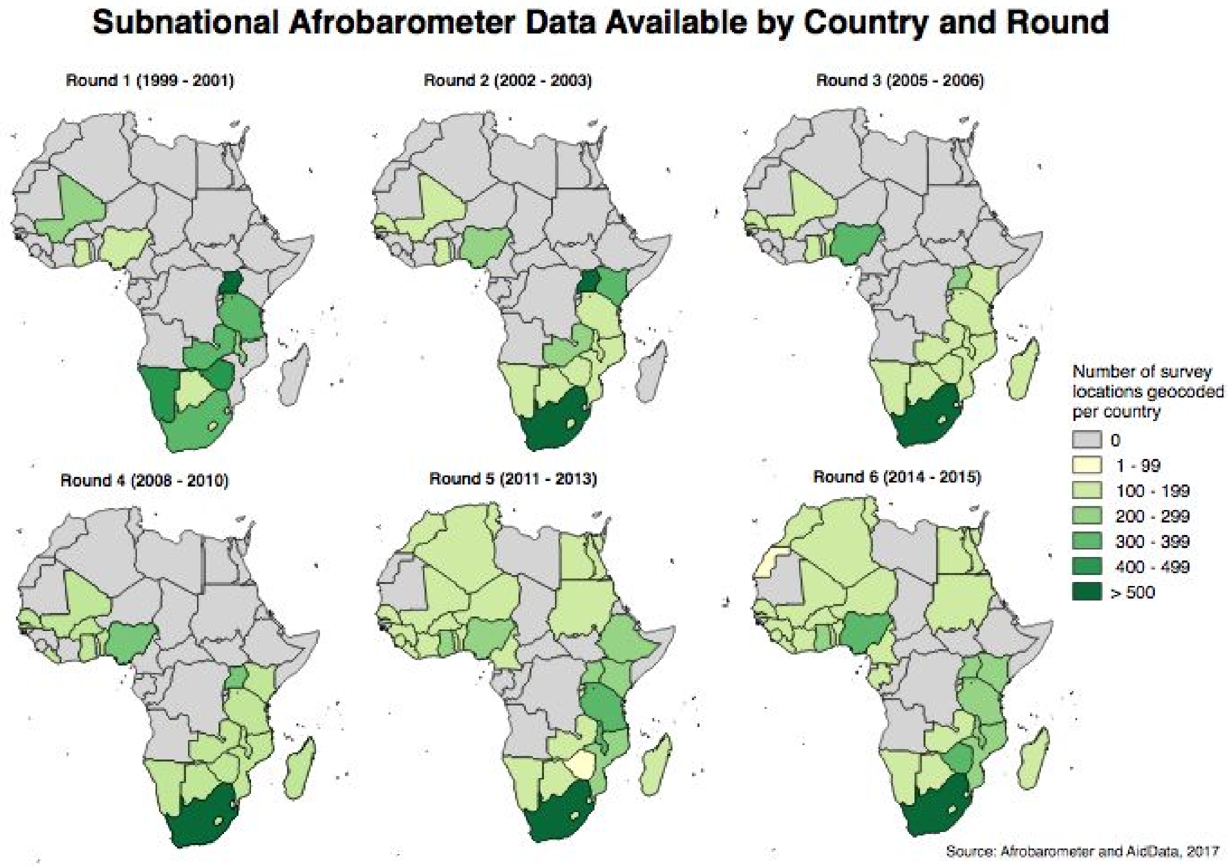
Because of the limited information available on the locations of Afrobarometer surveys, an additional step of data quality assessment was undertaken to provide users with information on the expected cases in which this data could — or should — be used. Quality assessment was performed using Round 5 data of Afrobarometer. Sixty-one projects, or approximately 1% of the total dataset, were selected at random, and three levels of quality were assessed through a "geocoding plus" methodology, in which an expert coder used all possible resources — including primary, secondary, and other sources — to code a small subset of the data at a very high time cost. The three levels of quality assessed included:

1. If the locations were factually accurate, to the best of the coder's knowledge.
2. If the locations were of the same level of granularity provided in the source Afrobarometer data.

[1] Approximate geocodes are summarized as when a location did not have exact administrative boundaries. For example, most parks, mountain ranges, and other topographical features will be coded as approximate. While part of AidData's standard geocoding methodology, these were rare in the Afrobarometer data.

3. If it was possible to find a higher level of granularity than what was recorded, given additional resources.

Of the 61 projects randomly selected, all locations were found to be factually accurate given the available information. However, in 13% of cases (9 projects), more granular information was found based on the use of additional sources of information. Across all assessed projects, 30% (18 projects) were geocoded to the exact place name provided by Afrobarometer; remaining projects were "rolled up" to coarser units of observation (i.e., administrative districts) due to a lack of information on the exact location of place names.


Subnational Afrobarometer Data Available by Country and Round

Round 1 (1999 - 2001)

Round 2 (2002 - 2003)

Round 3 (2005 - 2006)

Round 4 (2008 - 2010)

Round 5 (2011 - 2013)

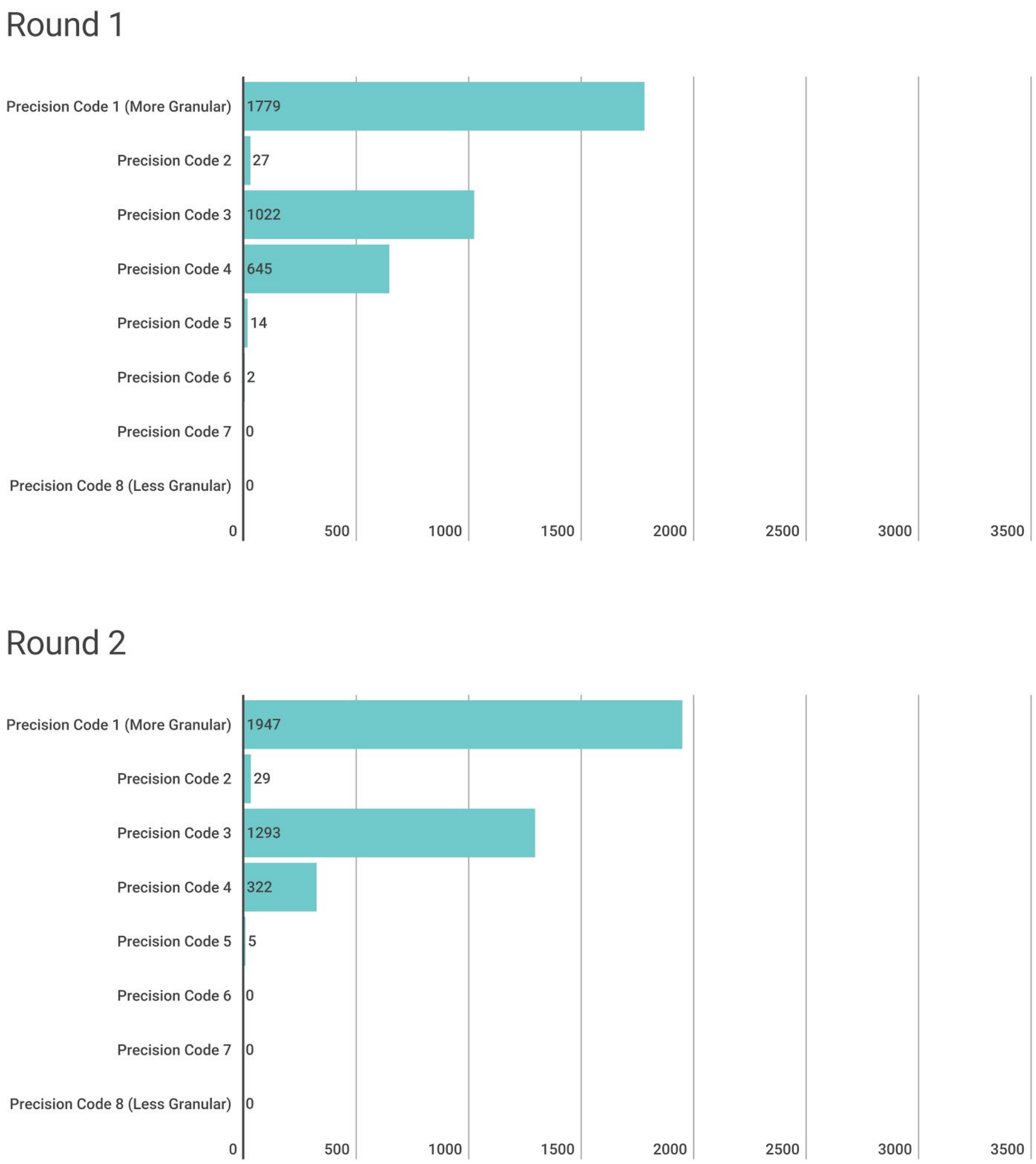
Round 6 (2014 - 2015)

Source: Afrobarometer and AidData, 2017

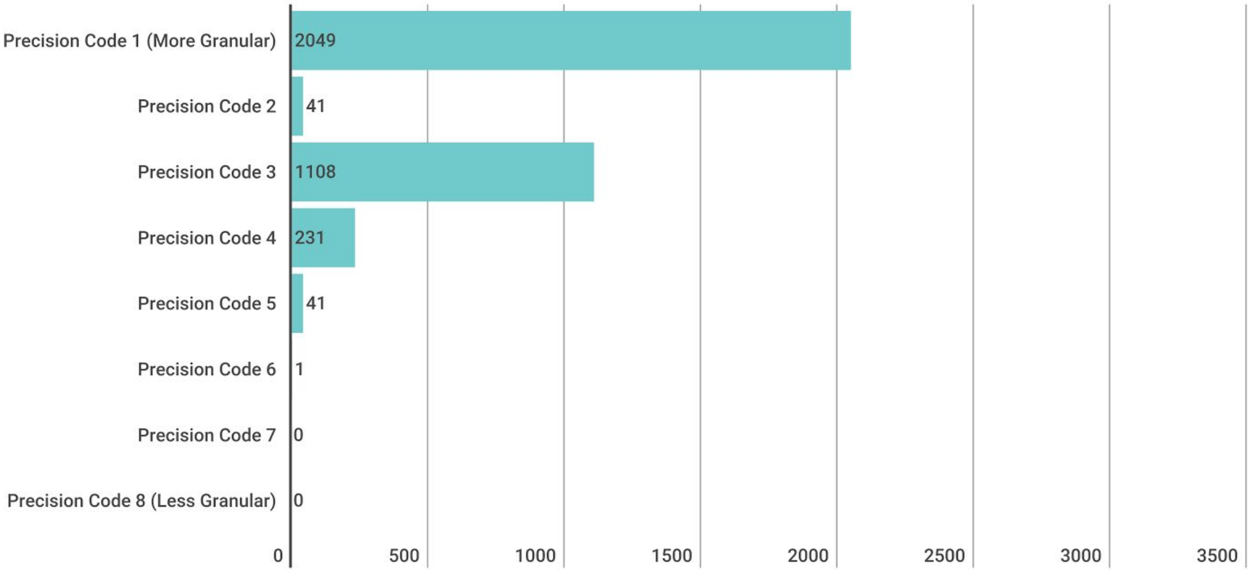
Figure 2. Geospatial distribution of geocoded locations.

Broadly, the quality and quantity of data geocoded was relatively consistent over time, with a similar percentage of points located in each country and with similar levels of geographic precision. Figure 2 illustrates the number of locations geocoded by country during each wave of the survey. Figure 3 illustrates the relative geographic precision of these locations in each round. As this figure shows, during all waves Precision Code 1 (the most granular available) is the most frequently coded location.

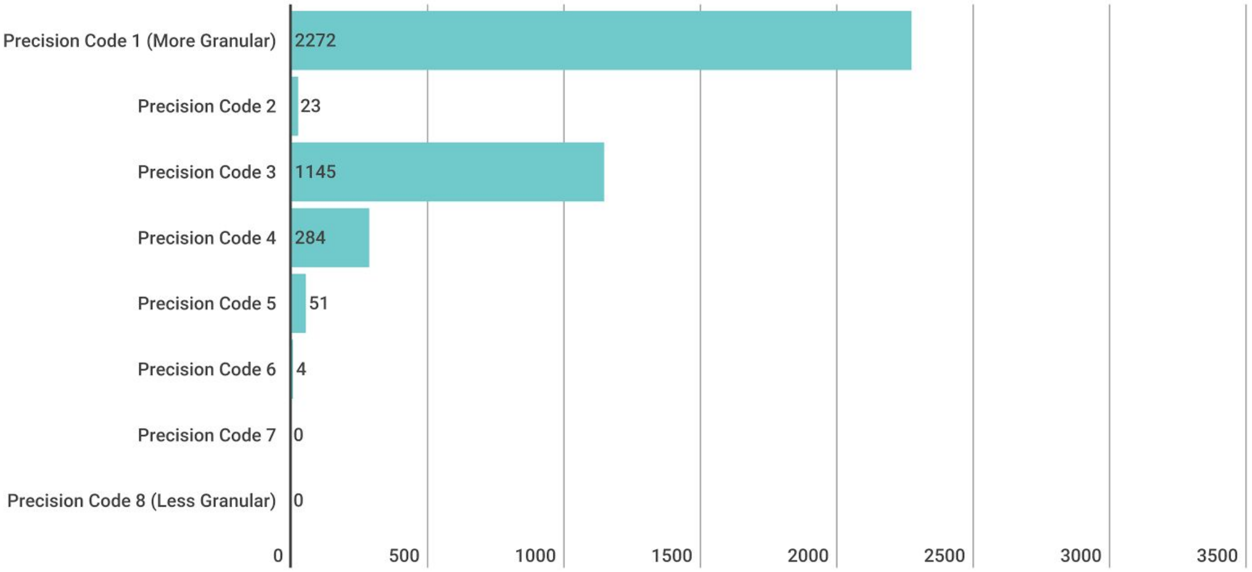
Figure 3. Geocoded locations by precision code



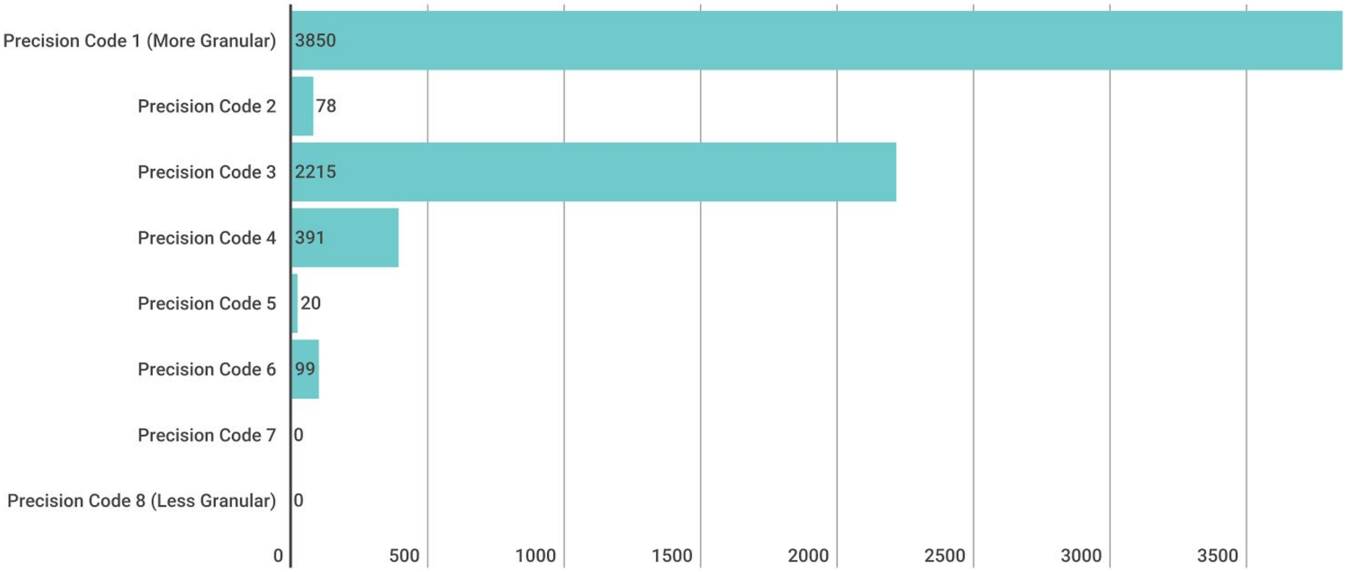
Round 3



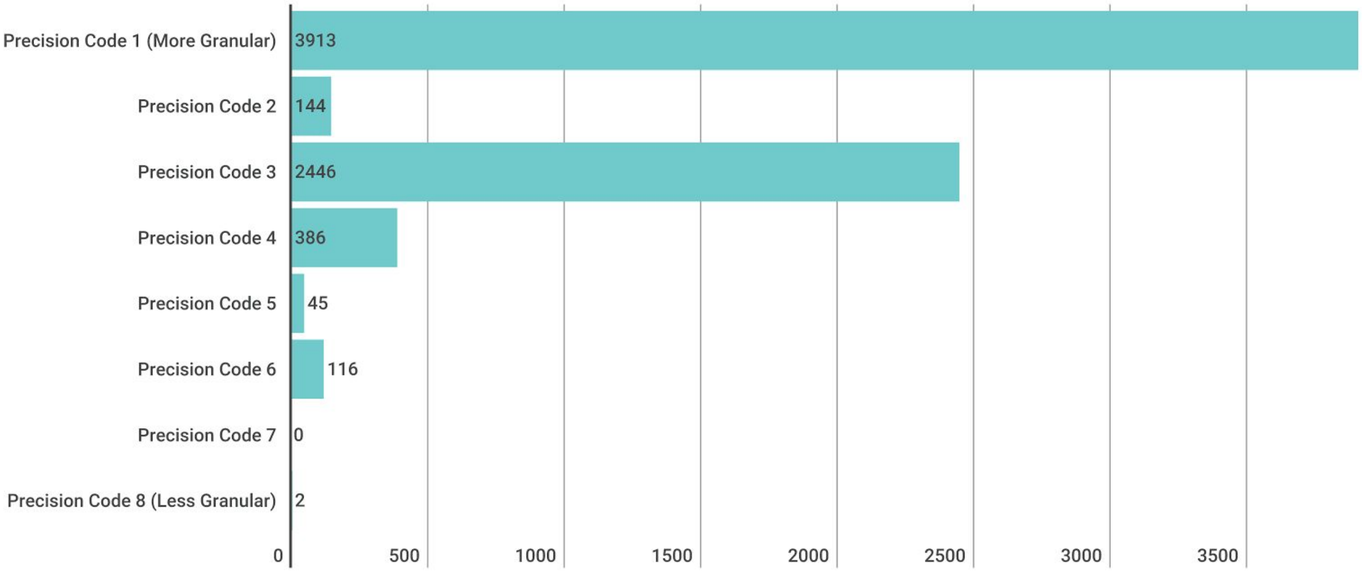
Round 4



Round 5



Round 6



References

Knutsen, C. H., Kotsadam, A., Olsen, E. H. and Wig, T. (2017), Mining and Local Corruption in Africa. American Journal of Political Science, 61: 320–334. doi:10.1111/ajps.12268

Strandow, Daniel, Michael Findley, Daniel Nielson, and Josh Powell. 2011. The UCDP-AidData Codebook on Geo-referencing Foreign Aid. Version 1.1. Uppsala Conflict Data Program. Uppsala, Sweden: Uppsala University.

Appendix

AidData Geocode Location Variables	
geoname_id	Location identity ID from http://www.geonames.org/
precision_code	AidData precision code (represents levels of location granularity, with smaller values generally being more granular)
place_name	Location name
latitude	Latitude
longitude	Longitude
location_type_code	Location type (see http://www.geonames.org/export/codes.html)
geoname_adm_code	Geography Code
geoname_adm_name	Geography Name
location_class	IATI location class defines whether the location refers to an administrative division, a populated place, a structure, or topographic feature. (see http://iatistandard.org/202/codelists/GeographicLocationClass/)
Geographic exactness	IATI exactness defines whether the location represents the precise project activity location, or approximate. (See http://iatistandard.org/202/codelists/GeographicExactness/)