Getting Started with GeoQuery

A quick-start guide to the download and use of spatial data for international development

geo.aiddata.org

GeoQuery Quick Start Handbook v. 1.01, December 2017
WWW.GEOQUERY.ORG | GEO.AIDDATA.ORG

THIS WORK IS AUTHORED BY DAN RUNFOLA, SETH GOODMAN, AND ZHONGHUI LV, AS WELL AS MANY COLLABORATORS WITHIN THE AIDDATA RESEARCH LAB AT THE COLLEGE OF WILLIAM AND MARY. GEOQUERY IS MADE POSSIBLE BY THE SUPPORT OF USAID, KFW, HUMANITY UNITED, THE WORLD BANK, THE GLOBAL ENVIRONMENTAL FACILITY, THE MACARTHUR FOUNDATION, AND THE COLLEGE OF WILLIAM AND MARY. 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

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. AidData'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.

As a research institution, we ask that users of the GeoQuery framework cite:
Goodman, S., BenYishay, A., Runfola, D., 2017. Overview of the geo Framework. AidData. Available online at http://geo.aiddata.org/. DOI: 10.13140/RG.2.2.28363.59686

Contents

1. What is GeoQuery?

Extracting value from existing spatial information can be extremely challenging for non-expert users. Programs such as ArcGIS and Q can take years to master, datasets can be large and unwieldy, and spatial weighting and corrections can be difficult to understand and implement. Historically, the human and computational costs associated with obtaining relevant spatial data - e.g. the intensity of nighttime lights, the amount of tree cover, or the distance from a major road - at meaningful units of analysis, such as census tracts or village administrative boundaries, has prevented spatial analyses in a wide range of fields.

Some recent initiatives have simplified access to the rapidly growing array of rich spatial data for non-experts. Of particular note are the efforts of PRIO grid, TerraPop, growUP, and a number of NASA products. These tools engage slightly different user groups, ranging from academic researchers to practitioners. GeoQuery adds to this ecosystem by providing a scaleable, highly parallelized computational framework designed to enable non-experts to quickly extract massive amounts of spatial information at fully customizable geographic units. Where previous tools have relied heavily on preprocessing, revolved heavily around a single data source, or pre-determined the geographic boundaries a user can utilize, GeoQuery allows all of these factors to be dynamic. Through GeoQuery, we provide a flexible, expert-curated solution which enables easier access to a wider set of spatial information than has been available to date.

This document provides information on how to quickly run analysis for international development using spatial data accessed via GeoQuery. It is aimed at users with little technical background in the use of spatial data, and seeks to introduce both the challenges and solutions GeoQuery provides.

2. Quick Start Guide

2.1 Downloading Data

The primary purpose of GeoQuery is to provide spatial data in a "spreadsheet" format - where each row represents a geographic boundary and each column represents the value of a spatial dataset in that boundary - as well as a PDF describing what each variable means. An example of the extract you'll be creating in this quick start guide can be seen in figure 2.4.

The steps to generate the example dataset are as follows:

1. Go to geo.aiddata.org and select "Get Data".
2. On the following screen, you can select the geographic boundaries you are interested in aggregating data to. For this example, type in "Ghana".
3. Select the 2nd level administrative district option (Figure 2.1).
4. Select World Bank Geocoded Aid Data on the left of the data selection screen (see figure 2.2), and then select filters on the right so as to include aid from 1998 to 2000, and any number of sectors which are related to child mortality.[1] Then click the "Add to Request" button.
5. Select Population (GPW V3, UN adjusted), choose the year 2000, and add to request
6. Select Child Mortality in Africa, choose the years 1990 and 2000, and add to request.
7. Click Submit Request in the upper-right, then Review Request.
8. Enter the email address you would like the results sent to on the right, and click submit (see Figure 2.3).

[1] To exactly replicate the data in this quick start guide, select the end years 1998, 1999 and 2000, and sectors including all education, energy generation, government and civil society, health sectors, other social infrastructure, and water supply. In practice, GeoQuery enables practitioners to build filters as specific (a single project) or as coarse (an entire sector or portfolio) as they want for their use case. The data and implicit theory of change presented here are meant for illustrative purposes only.

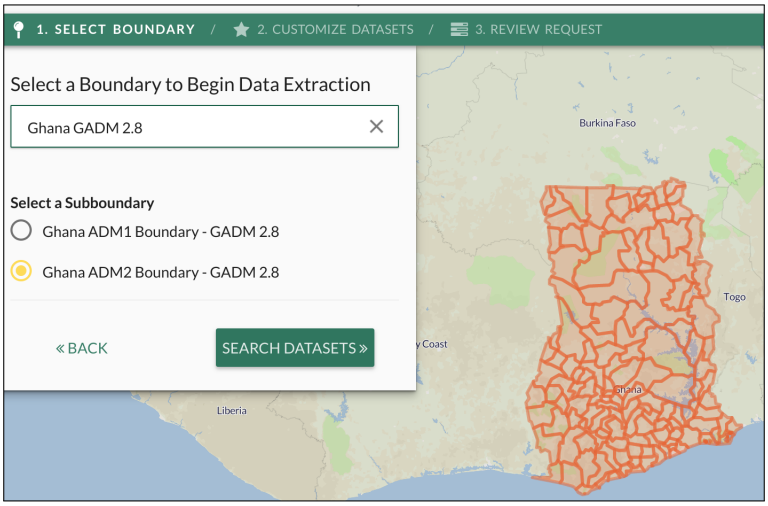

Figure 2.1: Screen for the selection of boundaries. Each boundary presented on the map will be a row in the data that is generated for download.

With the above steps completed, your request will have triggered a job on the SciClone high performance computing cluster at the College of William and Mary. All jobs are processed dynamically, so depending on the complexity of your request you will likely receive a completion email within 5 minutes to approximately 4 hours for extremely large or complex jobs.[2]

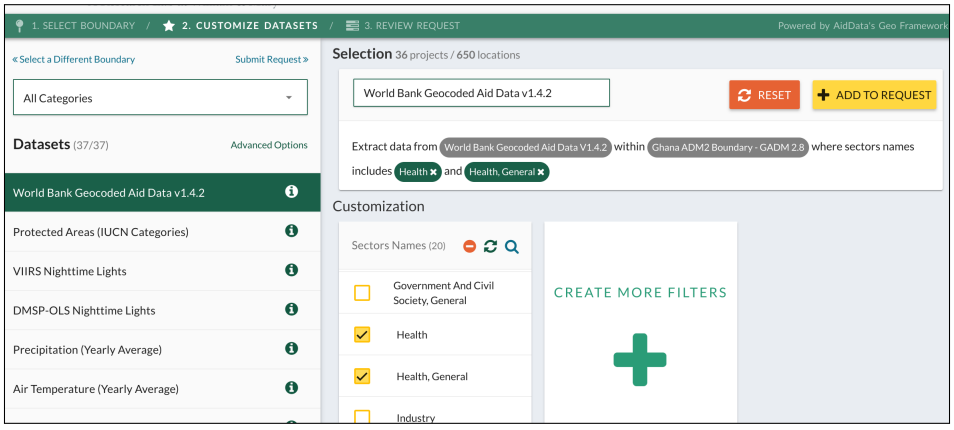

Figure 2.2: Screen for the selection of datasets. You can add as many data sets for download as you wish, where each data set will be a new column in the final spreadsheet you receive. In the case of Aid Data, additional filters such as year by clicking the "+" on the right pane.

When your job is complete, you will receive an email directing you to the permanent download page for that data. You can also always retrieve your history of data requests and downloads by clicking "Past Requests" at the upper-right of GeoQuery and entering your email address.

[2] In some cases jobs may take up to 24 hours (or longer) depending on the number of users on the system, i.e. if many users concurrently put in extremely complex jobs. If you have put in a request for a job and it has been longer than 24 hours, you can reach out to the GeoQuery support team at geo@aiddata.wm.edu.

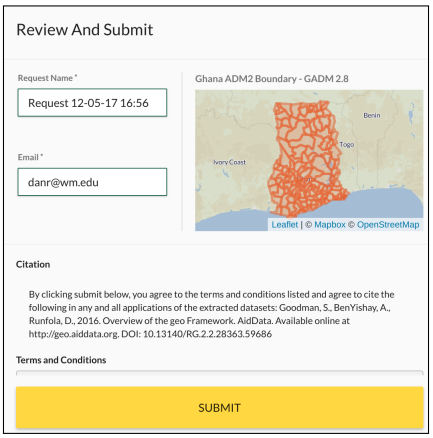

Figure 2.3: Screen to confirm and submit your request. Once you click submit, your job will be submitted and a confirmation email sent.

2.2 Understanding GeoQuery Data

Once your request is completed, you can download three different files: a PDF containing descriptions of your data, a CSV containing your data itself, and a zip file containing everything generated during your request. If you download the CSV and open it in a common program such a Libre Office or Excel, you will see an output very similar to that seen in figure 2.4. Each row in this file is a geographic boundary, and each column is a summary of geographic information.

	A	B	C	D
1	asdf_id	worldbank_geocodedresearchrelease_level1_v1_4_2.c2523c5.sum	africa_child_mortality.1990.mean	africa_child_mortality.2000.mean
2	0	360612.9074	26.44280973	25.7828
3	1	826928.9015	25.54296774	24.91577
4	2	392431.6959	16.09930979	20.21558
5	3	430901.8636	21.47161479	12.4963
6	4	707563.5138	19.56386736	8.665572

Figure 2.4: Example of a GeoQuery extract. Each row represents a geographic boundary, and each column represents data in that boundary. As an example, the first row in this figure is from North Adansi, Ashanti, Ghana. The second value (360612) is the sum of World Bank aid in North Adansi which matches the users request, and the third column is the estimated number of under-5 deaths per 1000 child-years within North Adansi during the 1990s. Detailed information on each variable is found in the metadata PDF.

In the PDF automatically prepared for you, you will find a description of exactly what each column means, how it was generated, and how to cite the underlying information. To identify what a column name means, the quickest way to find it is to open your pdf and search through it (i.e., ctrl or command + f) for the name of the column. Units and other key information are provided, as seen in figure 2.5.

Dataset 2 - Child Mortality in Africa

Title	Child Mortality in Africa
Name	africa_child_mortality
Version	1
Column Names	Format: "africa_child_mortality.<temporal>.<method>" for all combinations of <temporal> and <method> which can be found in the "Temporal Selection" and "Extract Types Selected" fields below (2 columns total)
Temporal Selection	2000, 1990
Extract Types Selected	mean (average under-5 mortality for each unit of analysis)
Description	Under-5 mortality (estimated deaths per 1,000 child-years) for the 1980s, 1990s, and 2000s. Based on geo-spatial interpolation methods using data from the Demographic and Health Survey across 28 Sub-Saharan countries.
Details	Rasterization of point based data.
Bounding Box	[[[-17.5, 25.0], [-17.5, -30.6], [47.8, -30.6], [47.8, 25.0], [-17.5, 25.0]]]
Date Added	2017-05-08
Date Updated	2017-09-26
Source Name	Stanford Data Portal
Source Link	http://sheftneal9.wixsite.com/fse-data/paper-1

Figure 2.5: Example from the metadata PDF generated by GeoQuery. The explicit definition, citation information, and a variety of other characteristics of each dataset are reported for all data selected by the user.

In addition to the column name interpretations, this PDF also provides information on the boundary files being used to generate your results. In the current version of GeoQuery (December 2017), these boundaries are largely retrieved from www.gadm.org. [3] A few exceptions - for example, a global grid accessible by searching for "Grid" - are currently available.[4]

2.2.1 Aid Data

The AidData information provided in GeoQuery is generated leveraging the geocoded information created by www.aiddata.org. Using this information, which is geocoded to points, the SciClone infrastructure automatically identifies the appropriate geographic boundary to which each aid project was attributed. For example, if a grant to improve health education was given to an administrative body in a district, the aid is allocated across that entire district rather than to a single arbitrary point within the district. If appropriate for your analysis, you can choose to disable this functionality by restricting aid to only "Precision Code 1", or only aid that is known to exact points.

Because of underlying data limitations (the exact end or start date of all projects are rarely known), selecting a range for years is not feasible at this time. However, we do offer the option to select only projects for which given end or start years are known. As a product of this, if you choose to filter by end or start year, you will only get projects which started or ended on the year(s) you choose; if you choose no start or end year, all projects will be included.

[3] Future versions of GeoQuery will rely on a new, fully open licensed boundary dataset; more information on these new boundaries will be available Spring 2018.

[4] The option to upload new boundary files is frequently requested, but not supported by our current infrastructure. Please contact us (geo@aiddata.wm.edu) if you have a boundary you need data for; we can generally accommodate requests within a few business days.

2.3 Exploring Data

Unlike most spatial data tools, GeoQuery is designed for users that seek to use spatial data in traditional software packages (i.e., Excel, STATA, R, LibreOffice, Google Sheets), not specialized mapping software. Thus, most graphics can be created using common spreadsheet techniques - i.e., figure 2.6 is created by dividing the World Bank aid selected in the example abstract (1998 - 2000 aid that might be relevant for preventing under-5 mortality) by population, to measure per-capita aid for each district within Uganda.

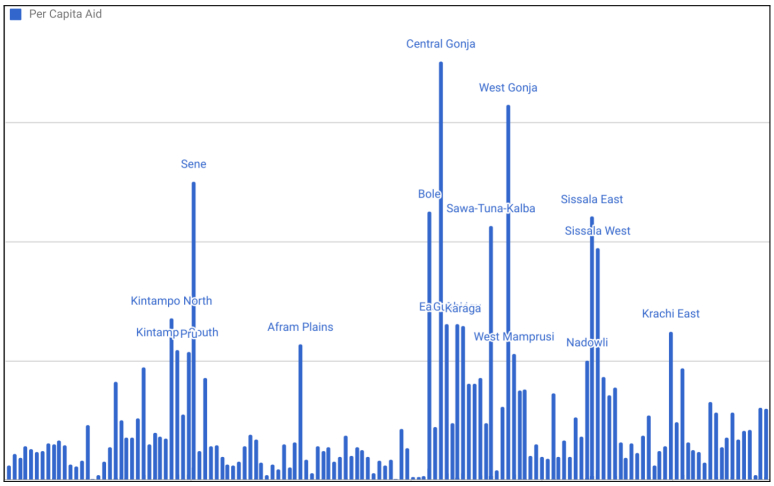

Figure 2.6: Example of descriptive visualization of per capita aid for each second level administrative zone of Ghana, with large outliers highlighted.

NOTE If you are accustomed to using GIS software platforms such as QGIS, you can join data downloaded from GeoQuery to boundary shapefiles retrieved from www.gadm.org for visualization tasks.

Another way to leverage the data in GeoQuery is in targeting analyses. Figure 2.7 provides an example in which the rate of child mortality within each district from 1990 - 2000 is compared to the World Bank aid we know ended in 1998, 1999 or 2000. While a coarse measure, this figure illustrates that World Bank projects in the 90s appear to have been allocated to areas that also had higher levels of child mortality. In practice, individual researchers must choose the types of aid they seek to analyze based on their specific research questions.

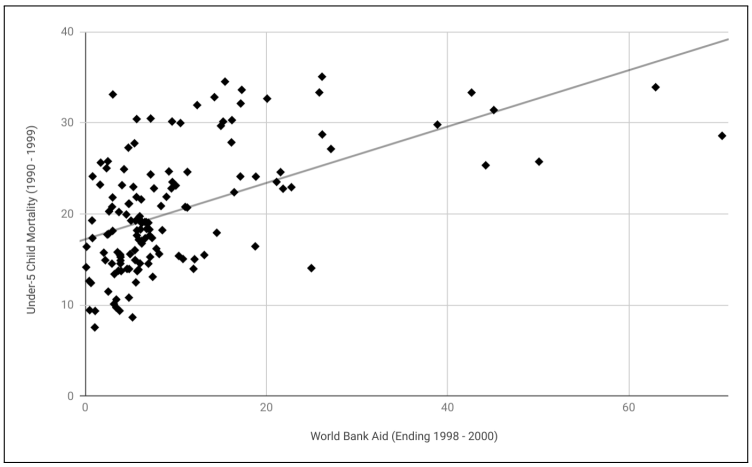

Figure 2.7: Example of a targeting analysis using GeoQuery. Here, we examine if World Bank aid in the 90s tended to be allocated to areas with higher rates of child mortality, as indicated by the upward trending pattern.

3. GeoBoundaries: Technical Detail

The GeoBoundaries dataset exists to provide an openly sourced spatial dataset of global administrative divisions. The first of its kind, this dataset will be updated periodically ensuring the accuracy and integrity of data provided. This dataset is geared towards academics, practitioners, and policymakers alike, allowing for analyses on a macro to a micro level to be precise. The dataset was compiled from existing openly licensed raw datasets, with individual sources and licenses attributed to every country and level. The alpha version of GeoBoundaries currently includes 581 shapefiles reaching 196 independent territories from the national boundary level to the neighborhood level in some cases. This dataset exists in both shapefile and geojson formats and has been quality checked to reflect the most accurate open data on administrative boundaries in the world.

3.1 Technical Details

The GeoBoundaries dataset is compiled by searching for and requesting open spatial data on administrative boundaries. The sources used for GeoBoundaries include data from the United Nations Office for the Coordination of Humanitarian Affairs on the Humanitarian Data Exchange[1], crowd-sourced boundaries from OpenStreetMap[2], and data collected directly from government geographic information offices and local geographic groups around the world. GeoBoundaries only uses data from sources where fully open source licenses can be obtained. The typical search chain checks official sources (e.g., governments), followed by other reputable collection sources (e.g., NGOs), data aggregation sources (e.g., Humanitarian Data Exchange), and if no other information is available, "raw" sources such as OpenStreetMaps.

When gathering data for GeoBoundaries, the GeoBoundaries team searches for and collects boundary data by country and administrative level. All data is downloaded with all associated files, any available metadata, and relevant notes detailing search and collection process retained. Following this, the data are checked, cleaned, and standardized.

[1] https://data.humdata.org/

[2] https://www.openstreetmap.org

3.1.1 Checking Process

Data was cross referenced with open sources available that contained information on the country's administrative divisions, known as reference information. This meant utilizing Wikipedia, Statoids, and other official information on country government websites. Reference information on Wikipedia was usually just as correct as the reference information on government websites. In some cases, we could not verify all rows included in the raw data, due to issues such as raw data being in another language or the inability to find accurate reference information on the boundaries of a country. This happened usually in cases of finer administrative levels, such as ADM 2, ADM 3, and finer. If raw data was missing information about a certain boundary level, i.e. missing an entire district or province, we rectified the missing information in the cleaning process.

3.1.2 Cleaning Process

The cleaning process was used to rectify any errors found in raw data. In these cases, the geometry and attribute tables of the raw data were changed to reflect an accurate boundary level. This was not possible in all cases due to the lack of reference information being available; however, any outdated and inaccurate boundaries will be updated as such information becomes available. The metadata table included in the download of the GeoBoundaries dataset contains a column that denotes whether changes were made to the geometry or attribute table outside of the standardizing process. There are two key processes that were performed in order to clean raw data: removing water territories and dissolving for ADM 0 levels.

Removing Water Territories

Many country boundaries encompass water features such as lakes and rivers, and include parts of ocean in their claimed territory, known as territorial waters.[3] The inclusion of these features can skew quantitative analyses. For example, the territorial extent of Uganda encompasses part of Lake Victoria. Many analyses at AidData seek to understand the relationship of aid on a subnational level. If Uganda's claim on Lake Victoria were included, the analysis would show an inaccurate amount of aid going to Lake Victoria, which is most likely incorrect. By removing the actual extent of the water territories, and solely focusing on the land territories, all analyses derived from the GeoBoundaries dataset will be more accurate and precise.

Some administrative layers in the GeoBoundaries dataset have excluded these water territories based on raw data that excludes them as well.

Not all boundaries in the GeoBoundaries dataset have removed water territories from their extent. In cases where one raw data layer in a country excludes water territories, and one raw data layer in a country does not, the boundaries have been clipped so that the total extent of the country's boundaries (in all levels) are the same.

Dissolving for National Boundaries

All national boundary shapes, also known as ADM 0's, were derived from a higher administrative level, either ADM 1 or ADM 2. The raw data layer of the higher administrative level was dissolved completely to derive the extent of the boundary without internal divisions, or the ADM 0 level. All source data attributed to the ADM 0 layer comes from the raw data it was dissolved from, as noted in the metadata table. All instances of dissolve not resulting in an ADM 0 layer are noted separately in the metadata table.

[3] https://en.wikipedia.org/wiki/Territorial_waters

Projection Information

All boundaries are projected to WGS 1984 (EPSG 4326) to ensure standardization across boundaries.

3.1.3 Standardizing Process

Once the data was checked and cleaned, the attribute table and file name was standardized to reflect a processed dataset. All file names are saved as:

"[ISO code]_ADM[#]"

in a zipped shapefile. For example, the ADM 1 level of Armenia is saved as:

"ARM_ADM1.zip"

This format is standard across all processed data layers, countries, and levels. Within the standardized attribute table, there are four required fields:

- ID
- Shape
- Name
- Level
- ISO Code*

The International Organization for Standardization (ISO)[4] generates codes for the names of countries, boundary divisions, and special geographic areas of interest. Each country in the world has a unique ISO code, and each ADM 1 and some ADM 2 levels have specific ISO codes for their boundary divisions. ISO Codes are used in the GeoBoundaries dataset because they represent a standard unique identifier across the dataset. ISO codes were included for all ADM 0 levels, most ADM 1 levels, and a few ADM 2 levels, and were last updated in 2013. For some countries, ISO codes are outdated, meaning a country has changed their boundary divisions since the International Organization for Standardization reviewed their divisions. All ISO codes in the ADM 0 levels reflect the three letter country code given by ISO. All ISO codes in the ADM 1 or ADM 2 levels reflect the standard codes from 3166-2[5], and represent the subnational boundaries of a country. These codes are created from the two letter ISO country code, with a two or three letter or number unique identifier for each division. In the case of Kosovo, ISO has not recognized Kosovo as an independent territory; however, the European Union uses the code "XKX" to denote Kosovo. For more information on ISO codes, please visit https://www.iso.org/standards.html.

[4] https://www.iso.org/home.html
[5] https://en.wikipedia.org/wiki/ISO_3166-2