Computers and Geosciences 122 (2019) 103–112

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

GeoQuery: Integrating HPC systems and public web-based geospatial data tools☆

Seth Goodman[a,b], Ariel BenYishay[a,c], Zhonghui Lv[a], Daniel Runfola[a,b,*]

[a] Global Research Institute, AidData, The College of William and Mary, USA
[b] Department of Applied Science, The College of William and Mary, USA
[c] Department of Economics, The College of William and Mary, USA

ARTICLE INFO

Keywords:
Geographical information science & systems
Parallel and high-performance computing
Algorithms
Data processing
Spatial statistics
Computational methods

ABSTRACT

Interdisciplinary use of geospatial data requires the integration of data from a breadth of sources, and frequently involves the harmonization of different methods of sampling, measurement, and technical data types. These integrative efforts are often inhibited by fundamental geocomputational challenges, including a lack of memory efficient or parallel processing approaches to traditional methods such as zonal statistics. GeoQuery (geoquery.org) is a dynamic web application which utilizes a High Performance Computing cluster and novel parallel geospatial data processing methods to overcome these challenges. Through an online interface, GeoQuery users can request geospatial data - which spans categories including geophysical, environmental and social measurements - to be aggregated to user-selected units of analysis (e.g., subnational administrative boundaries). Once a request has been processed, users are provided with permanent links to access their customized data and documentation. Datasets made available through GeoQuery are reviewed, prepared, and provisioned by geospatial data specialists, with processing routines tailored for each dataset. The code used and steps taken while preparing datasets and processing user requests are publicly available, ensuring transparency and replicability of all data and processes. By mediating the complexities of working with geospatial data, GeoQuery reduces the barriers to entry and the related costs of incorporating geospatial data into research across disciplines. This paper presents the technology and methods used by GeoQuery to process and manage geospatial data and user requests.

1. Introduction

Geospatial data is expanding in terms of quantity, quality, scope, and accessibility (Miller and Goodchild, 2015). Researchers have access to geospatial data in different formats from a wide variety of sources, including data generated by satellites (USGS, 2017; ESA, 2017; NOAA Earth Observation Group, 2017), geoparsing and coding of news articles (Leetaru and Schrodt, 2013), GPS enabled consumer devices (Zandbergen and Barbeau, 2011), census records (Minnesota Population Center, 2015) and many more. Given that the process of collecting and preparing data for research can be the most time consuming portions of a project even for researchers with relevant experience and skills (Kugler et al., 2015; Munson, 2012), tools that reduce the time needed for these steps can be of significant use to researchers.

This paper presents GeoQuery, a solution to many of the challenges of integrating and accessing geospatial data, which combines the computational power of high performance computing (HPC) and the usability of a simple web application. Sections 1.1 and 1.2 provide an overview of the types of geospatial data being generated and concomitant need for the tool presented here. Section 2 explores the details of the data management and processing infrastructure employed in an HPC cluster environment, the data types available through GeoQuery and related preparation steps, data processing methods, and the design of the web application. Section 3 provides a discussion and conclusion, including usage within research and academic communities, lessons learned since the development and launch of GeoQuery, and avenues for future work.

☆ D.R. and S.G. conceptualized the project, selected and refined relevant algorithms, and collaboratively wrote this piece. A.B. and Z.L. assisted in algorithmic development, data preparation and testing. S.G. wrote program code for core GeoQuery functionality; Z.L. assisted on many preprocessing routines.

* Corresponding author. Global Research Institute, AidData, The College of William and Mary, USA.
E-mail address: danr@wm.edu (D. Runfola).

https://doi.org/10.1016/j.cageo.2018.10.009
Received 5 June 2018; Received in revised form 17 October 2018; Accepted 24 October 2018
Available online 02 November 2018
0098-3004/ Published by Elsevier Ltd.

1.1. Variety and growth of geospatial data

Examples of geospatial data providers are many and varied, as are the types of data and applications available. Satellite based sensors are capable of measuring a range of physical properties such as soil moisture, vegetation, elevation, cloud structures, and ocean pigment concentrations. Some satellite products - including nighttime lights, forest cover, and vegetation metrics - have historic data available for over 25 years and are being produced at increasingly finer temporal and geospatial scales (Henderson et al., 2012; Bruederle and Hodler, 2017; Bundervoet, 2015; Hansen et al., 2013). Geospatial data generated by integrating remote imagery with other expertise and datasets have been produced for a range of additional covariates such as temperature (Matsuura and Willmott, 2015a), precipitation (Matsuura and Willmott, 2015b), slope (Jarvis et al., 2008; NASA, 2000), water bodies (Wessel and Smith, 1996) and more (Goodman et al., 2016).

The growing availability and use of geospatial data is not only limited to the geosciences. Within the international development community, geospatially referenced data on aid has increased significantly in recent years (USAID, 2015). Demographic and Health Surveys (DHS), Living Standards Measurement Study (LSMS) surveys, and Afrobarometer surveys have now been geocoded to the level of enumeration areas, providing information on the socioeconomic characteristics of previously “invisible” populations (Burgert et al., 2013; Grosh and Glewwe, 1995; BenYishay et al., 2017b). Multiple sources now regularly produce observations of social and violent conflict events, including the Uppsala Conflict Data Program's Georeferenced Events Dataset (Sundberg and Melander, 2013), the Armed Conflict Location and Event Database (Raleigh et al., 2010), the Integrated Crisis Early Warning System (Boschee et al., 2018), and the Social Conflict Analysis Database (Salehyan et al., 2012). Geospatial information on population (CIESIN et al., 2015; CIESIN, 2017), child mortality (Burke et al., 2016), travel time to cities (Nelson, 2008), natural resource deposits (Sara and Maggioni, 2014; Gilmore et al., 2005; Lujala, 2009; Buhaug and Lujala, 2005; Lujala et al., 2007), and protected areas (IUCN and UNEP-WCMC, 2016) are also increasingly accessible.

1.2. Challenges and community progress to date

While the challenges of geospatial data access have been acknowledge for decades, as geospatial data has increased in quantity and resolution so has the scope of computation and skillsets required (Goodchild, 2008, 2009). A simple illustration of this challenge is the range of file formats and data structures across data sources. Raster data can utilize file standards including GeoTiff, HDF4/5, netCDF, ASCII, BIL, MAP, and dozens of others (GDAL/OGR contributors, 2018); further, data may be provided as a global mosaic or tiled to arbitrary sizes or as raw scenes captured by satellites. Vector data formats include shapefiles, plain text formats such as GeoJSON or CSV, geospatial extensions for databases such as SpatialLite and PostGIS, and even geospatial PDF documents containing georeferenced information (GDAL/OGR contributors, 2018). Individual files may be compressed or archived using numerous approaches (Granneman, 2006), and the conventions used for organizing and naming directories and files are rarely consistent across organizations or projects (c.f. GLCF, 2017; Elvidge et al., 2017; NOAA Earth Observation Group, 2017; ESA, 2017).

Further exacerbating the challenge of using geospatial data is the fact that the growing number of users of geospatial data are as diverse as the sources, topics, and formats of geospatial data. While some disciplines and sub-fields have a long history of incorporating geospatial data into analyses, many emergent users do not (i.e., Rajabifard and Williamson, 2001; Corrado and Fingleton, 2012; Kim et al., 2016; Gregory et al., 2015). This lack of experience can lead to a duplication of effort across, or even within, disciplines (Malvárez et al., 2015; GSDI, 2012).

Many groups within both the geosciences and social sciences are seeking to provide easier access to geospatial data. Many tools and applications exist within the broader geospatial data ecosystem, which engage a variety of user groups with different purposes and limitations based on the primary goals of the application and the needs of their audiences. For example, AppEARS (Application for Extracting and Exploring Analysis Ready Samples) by NASA/USGS (lpdaac.usgs.gov) offers custom point and area samples for data available through NASA Earthdata (earthdata.nasa.gov), with restrictions on the complexity of individual requests. The Peace Research Institute Oslo (PRIO) has created PRIO-GRID, which offers data on armed conflicts, socio-economic conditions, ethnic groups, physical attributes, climatic conditions and more, aggregated to a 0.5 × 0.5 decimal degrees grid (Tollefsen et al., 2012). IPUMS (Integrated Public Use Microdata Series), part of the Minnesota Population Center at the University of Minnesota, maintains IPUMS Terra. This tool offers extensive microdata along with data on agriculture, climate, and land cover that can be accessed or visualized in different formats (Nawrotzki et al., 2016). ETH Zurich developed the Geographic Research On War, Unified Platform (GROWup) which provides data related to settlement patterns of politically active ethnic groups around the world from 1946 to 2013 (Girardin et al., 2015). Other tools focus on data discovery, such as DataONE (dataone.org) or the IRI Climate Data Library (iridl.ldeo.columbia.edu). Giovanni (giovanni.gsfc.nasa.gov) provides visualization and analysis tools for geophysical data. The scope and audience of tools can vary greatly; Google Earth Engine (earthengine.google.com) is a general platform with vast capabilities for using and accessing geospatial data.

GeoQuery adds to this ecosystem by providing a scalable, parallelized computational framework designed to enable non-experts to quickly aggregate geospatial information from arbitrary datasets to geographic boundaries. The collection of data in GeoQuery aims to serve users across disciplines, from the social science to the natural/geophysical sciences. By reducing the overhead of processing datasets at large scales and offering an extensive selection of curated multidisciplinary datasets along with permanent replication links for requests, GeoQuery fills a critical gap for researchers seeking fast and easy access to a wide range of geospatial data.

2. Methods

GeoQuery was envisioned as a web application that would allow users with little to no training in geospatial data retrieval to easily find, and aggregate disparate geospatial data sources into a single CSV-based file. Because GeoQuery was intended to serve a broad set of disciplines and potential applications, it incorporates a diverse collection of geographic boundaries defining units of analysis and measurement datasets that can be aggregated to the boundaries. In order to ensure that GeoQuery would be sustainable, a standardized data model was defined that allows fully automated processes to manage all aspects of data processing and user requests.

2.1. Data model

GeoQuery separates datasets into two primary categories: (1) geospatial data in vector format (points, lines, or polygons)[1] defining geographic boundaries of units of analysis (e.g., watersheds or administrative boundaries), and (2) geospatial data in raster format representing measurements to be aggregated to units of analysis (e.g., land cover or nighttime lights). The processing architecture presented here is designed to take these two types of data as an input, and produce tabular summary statistics for each arbitrary unit of analysis- a procedure generally referred to as “zonal statistics” in GIS software platforms

[1] GeoQuery currently only contains boundaries based on polygons, but the data model and methods discussed in this paper support point and line features as well.

(ESRI, 2016). Zonal statistics involves identifying measurement data (i.e., pixels from raster data) relevant to a given boundary feature[2] and aggregating measurement values using a specified aggregation method (e.g., mean, min, max).

2.1.1. Boundary datasets

All boundary data are stored in GeoJSON format (Bray, 2014) and contain any source attributes such as the name of individual units of analysis as well as a unique identifier added during the GeoQuery ingestion process. GeoQuery's primary source of vector data is GeoBoundaries (Seitz et al., 2018). GeoBoundaries is easily accessible for public usage and provides administrative zone information for nearly all countries at the ADM0, ADM1, and ADM2 level, with reduced coverage at finer levels.[3] Additionally, arbitrary grid products that enable global-scope analyses have been generated, including a 0.5 × 0.5 decimal degree global grid.

2.1.2. Measurement datasets

Sources of measurement data are many and varied: from satellite-based measurements of vegetation to features describing river networks and other water bodies. In every case, preprocessing for GeoQuery involves ensuring the data are in raster format with pixel values that allow for meaningful aggregation to arbitrary units of analysis. In some cases, such as with CIESIN's GPW population datasets (CIESIN, 2017) the raw data are accessed manually through a portal with a login, consist of few relatively small files, and are provided as a global rasters that do not require additional processing. However, other datasets can be more difficult to download and require extensive processing before being ready to be ingested into GeoQuery, and potentially require working in consultation with the data provider. One example is the Normalized Difference Vegetation Index (NDVI) provided by NASA's Long Term Data Record (LTDR) (NASA, 2017a,b). Daily data files in HDF format from multiple sensors across the span of the dataset must be downloaded from an FTP server, cleaned to prevent erroneous values resulting from aggregation, aggregated to common time steps (i.e., years or seasons), then output in GeoTiff format. Other datasets, such as VIIRS (Elvidge et al., 2017) are provided as raster tiles, each file covering only a portion of the earth, which must be mosaiced into a global raster (in addition to any other processing).

Data generated from sources other than satellites can pose additional challenges. Survey data from Afrobarometer (BenYishay et al., 2017b) is provided with place-names alone; these place-names must first be translated into geographic coordinates, and then rasterized[4] in such a way so as to preserve ordinal, continuous, and discrete types of survey responses. Conflict events, frequently provided with coordinates, must be rasterized to appropriate formats and resolutions. Each of these data sources requires customized pipelines before datasets are ingested into GeoQuery. As an example, raw information (e.g., PDF documents) on the geographic locations to which international aid has been allocated undergoes a geolocation process which identifies relevant features (administrative zones, roads, cities, forests, etc.) associated with the aid's disbursement. These resulting vector features are then rasterized and the associated aid value is evenly split over the resulting pixels. The result of this rasterization and aid distribution for each geolocated feature is summed to produce a final aid surface product.[5]

A further complication encountered during preprocessing is the spatial and temporal scope and resolution of some datasets, and the amount of computational time required to prepare them. In order to prepare datasets in a reasonable amount of time (hours vs days) these preprocessing routines are often parallelized. In cases where the raw data are in vector format the datasets go through a rasterization process mentioned in the previous examples. While these rasterization and preprocessing steps do impose certain restrictions on how datasets must be formatted, the resulting standardization enables GeoQuery to function using automated processes.

All processing steps and code used are made publicly available[6] so that users have access to every step taken to transform the raw source data into the version used in GeoQuery. The collection of measurement data available through GeoQuery is curated based on the quality, frequency, geospatial and temporal coverage of datasets. Datasets include a range of outcome measures, intervention data, and covariate information. An overview of measurement data incorporated into GeoQuery is shown in Table 1.

2.1.3. Ingesting datasets into GeoQuery

For all datasets in GeoQuery, a metadata record is constructed and validated prior to ingestion. This metadata stores the type of input data (boundary, measurement), along with key fields such as a citation and source details, a description of the dataset (and relevant data units when applicable), keywords used for searches, temporal information, and other details. Before being accepted into GeoQuery, the dataset and metadata are run through an automated validation process which ensures essential fields are included in the correct format, and generates additional metadata describing the geospatial coverage of the dataset, a record of files included in the dataset, and system information (date, versions of scripts, etc.). Once validation has completed, the metadata is added into GeoQuery's primary MongoDB database collection (MongoDB, 2018) which stores the metadata for datasets in GeoQuery, and the datasets are stored as GeoJSON and GeoTIFF files for boundary and measurement data respectively.

Upon being added to GeoQuery, automated processes will detect new datasets and run indexing procedures to associate the new boundary or measurement dataset with existing measurement or boundary datasets, respectively, that have overlapping geospatial coverage. This indexing procedure enables subsequent processes in GeoQuery to rapidly identify boundary and measurement datasets that are related using fast and simple standard queries instead of more complex geospatial queries. Once indexes have been generated, another automated task runs to determine all possible combinations of boundary and measurement datasets that could be requested by a user. Each potential combination is added as a task for automated zonal statistic routines to run. The following section will examine the methods used to run zonal statistics in GeoQuery.

2.2. Zonal statistics methods

One of the core challenges overcome by GeoQuery is the bulk aggregations of arbitrary rasters to arbitrary boundaries. This challenge is largely due to limitations of existing tools and platforms which do not have the capability to parallelize tasks or handle large aggregations (e.g., 30m estimates of forest cover aggregated to country of Russia). Such tasks can either take days to run or fail due to memory limitations.

Although implementations of zonal statistics tools are commonly found in a variety of existing GIS tools (Adamczyk and Tiede, 2017; ESRI, 2016; Rueda et al., 2005; Bunting et al., 2014), GeoQuery has

[2] The term “boundary feature” or simply “boundary” or “feature” refers to a single unit of analysis, whereas “boundary dataset” or “boundaries” refers to the complete dataset or all units of analysis.

[3] “ADM” notation refers to administrative levels, where ADM0 is the country level and ADM1 is the next finest administrative level, and so on.

[4] Rasterization of a simple vector feature defining a boundary creates a binary raster at an arbitrary resolution where values of 1 indicated pixels within the boundary. Rasterization of a vector feature representing measurement data will produce a raster with values based on a specified attribute of the vector feature. For additional details, see: https://www.gdal.org/gdal_rasterize.html.

[5] Implementation of this methodology can be found at https://github.com/aiddata/geo-hpc.

[6] https://github.com/aiddata/geo-datasets.

Table 1
Measurement data sources, June 2018.

Theme	Dataset Name	Source
Environmental/Geophysical	*World Database on Protected Areas*	IUCN and UNEP-WCMC (2016)
	Precipitation	Matsuura and Willmott (2015b)
	Air Temperature	Matsuura and Willmott (2015a)
	Ground Slope	NASA (2000)
	Physical Elevation	NASA (2000)
	On-Shore Petroleum Locations	Tollefsen et al. (2012)
	MODIS Land Cover V5.1	MODIS (2013)
	NDVI (LTDR)	NASA (2017a,b)
	Gemstone Deposits	Lujala (2009)
	ESA Land Cover (2.0.7)	ESA (2009)
	Drug Cultivation Sites	Buhaug and Lujala (2005)
	Gold Deposits	Lujala (2009)
	Ozone Concentration	Bouma et al. (1997)
	PM2.5 Concentration	Bouma et al. (1997)
International Aid	Global Environment Facility	GEF-IEO (2017)
	World Bank	AidData (2017)
	18 Country-specific datasets	Various
Socio-economic Datasets	Conflict Deaths	Sundberg and Melander (2013)
	Conflict Events	Raleigh et al. (2010)
	Population (V3, V4)	CIESIN (2000)
	Nighttime Lights (DMSP)	NOAA Earth Observation Group (2017)
	Nighttime Lights (VIIRS)	Elvidge et al. (2017)
	Travel time to Major Cities	Nelson (2008)
	Child Mortality in Africa	Burke et al. (2016)
	Trust in Country President (Africa)	BenYishay et al. (2017b)
Distance-based Metrics	Distance to Lootable Gold Deposits	Lujala (2009)
	Distance to Gemstone Deposits	Lujala (2009)
	Distance to Drug Cultivation Sites	Buhaug and Lujala (2005)
	Distance to Diamond Deposits	Gilmore et al. (2005)
	Distance to Coast	Wessel and Smith (1996)
	Distance to Water	Wessel and Smith (1996)
	Distance to Roads	CIESIN and ITOS (2013)
	Distance to Country Borders	Hijmans (2015a)

modified the Rasterstats Python package (Perry, 2016) to improve the flexibility and efficiency of the zonal statistics process in an HPC environment, and add additional functionality. These additions address three core aspects of zonal statistics which can be modified to improve computational efficiency and address specific usage concerns: (1) splitting individual boundary features into smaller pieces to manage memory usage and enable parallelization, (2) utilizing pixel coverage weights based on overlap of features with individual raster pixels to improve accuracy when using a feature which is small relative to resolution of raster data, and (3) incorporating weights based on latitude to accurately account for pixel area when calculating statistics which are area dependent.

2.2.1. Feature splitting

A simple but necessary step in generating statistics for any boundary requires determining which measurements (pixels) are associated with the boundary feature. In the case of raster-based measurements, this is equivalent to identifying which raster pixels intersect with a given boundary. This is accomplished by rasterizing the boundary vector - i.e., constructing a binary grid at the same geospatial resolution as the measurement data raster, in which each grid cell represented by a one indicates a pixel that intersects the boundary. This rasterized feature can then be applied as a mask to the measurement data to select only the relevant pixels. The selected raster pixels can then be passed to a statistical function for aggregation.

While a straightforward procedure, in order to enable arbitrary combinations of boundary and measurement data GeoQuery must be able to conduct this process irrespective of the scope or resolution of the input data. When a boundary covering a large area is combined with fine resolution measurement data, this process can result in a large amount of data (number of pixels) being read into memory. In HPC environments mitigating this is particularly important due to the shared-memory infrastructure of individual nodes in the cluster. For example, if a 16-core node (running 16 tasks in parallel) has 64 gigabytes of memory, the total memory being used by the 16 cores cannot exceed 64 gigabytes. If any core exceeds the total memory available on the node, not only will that core's task fail, but it may cause all tasks running on the node to fail as well.

To avoid exceeding the memory limits of a node, large boundary features are split into smaller pieces, and the zonal statistics process is run on each one individually before aggregating the results. This “feature splitting” approach is done following a procedure which seeks to:

1 Avoid splitting measurements in the underlying datasets (i.e., split along pixels edges rather than in middle of a pixel).
2 Split so as to guarantee that the total amount of memory being used by the set of cores on a node cannot cumulatively exceed the memory available on the node.
3 Preserve the geometry of the original boundary feature and the accuracy of the final, aggregated value.

To accomplish this, the maximum size of a feature that can be processed (given a set amount of memory) is defined in terms of the number of pixels covered by the feature. Since the number of pixels covered by a given feature will vary based on the resolution of the underlying measurement data, this must be assessed within the zonal statistics process, and subsequent feature splits applied dynamically. This pixel limit can be adjusted based on the available resources of a system and can be determined using simple scaling tests designed to estimate the number of pixels which suit the desired memory allocation per core.[7]

[7] Given a margin of error to account for data type and other factors, GeoQuery assumes approximately 250,000 pixels per 4 GB of memory when using the algorithms detailed in this paper.

2.2.2. Coverage weighting

The pixel size of measurement data can be highly variable across different measurement datasets (e.g., 30 m pixels vs 5 km pixels). In the case of small boundary features (small relative to the size of measurement data raster pixels), the area of individual units can be of a similar scale to the pixel size of the measurement data. This can result in pixels along the edges of a boundary, which only partially overlap with the boundary, constituting a large portion of the pixels used to calculate zonal statistics for the boundary. Pixels which overlap with multiple boundary features can present additional issues, such as in the case of aggregating population count. Including the population associated with a given pixel in the sum for multiple features would overestimate the true total population.

Because of cases like these, when generating zonal statistics for boundary features it can be useful to know the coverage or intersection of each pixel for the boundary feature being analyzed. These coverage estimates, or coverage weights, can then be used to improve the calculation of zonal statistics (Bunting et al., 2014; Hijmans, 2015b). GeoQuery utilizes coverage weights for all zonal statistics methods which could be impacted by pixel coverage (i.e., mean and sum but not min or max).

To illustrate the utility of coverage weighting in zonal statistics, consider the hypothetical measurement data seen in Fig. 1A, representing estimates of precipitation within each pixel.

Given this raster, and an arbitrary geographic boundary (red outline in Fig. 1B), a common use-case would be to apply zonal statistics to calculate the average precipitation (i.e., the mean value of pixels associated with the feature). Zonal statistics implementations generally select pixels based on whether the centroid of a pixel is covered by the boundary. Some tools such as Rasterstats or Starspan (Rueda et al., 2005) can also select pixels based on whether they intersect the boundary.

Ignoring the percent coverage shown in Fig. 1B and using the centroid based approach mentioned above, the two left pixels of the raster would be averaged and the two right pixels would be ignored, and the estimation of precipitation within the red boundary would follow Equations (1) and (2):

$$raster\ x\ coverage\ weights = weighted\ data$$

$$\begin{bmatrix} 10 & 30 \\ 40 & 20 \end{bmatrix} \times \begin{bmatrix} 1 & - \\ 1 & - \end{bmatrix} = \begin{bmatrix} 10 & - \\ 40 & - \end{bmatrix} \tag{1}$$

$$Mean = \frac{\sum weighted\_data}{\sum coverage\_weights} = \frac{10 + 40}{1 + 1} = 25 \tag{2}$$

As a comparison, incorporating the percent coverage of the boundary in the equation yields:

$$\text{raster x coverage weights} = \text{weighted data}$$

$$\begin{bmatrix} 10 & 30 \\ 40 & 20 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0.25 \\ 1 & 0.25 \end{bmatrix} = \begin{bmatrix} 5 & 7.5 \\ 40 & 5 \end{bmatrix} \tag{3}$$

$$Mean = \frac{\sum weighted\_data}{\sum coverage\_weights} = \frac{5 + 7.5 + 40 + 5}{0.5 + 0.25 + 1 + 0.25} = 28.75 \tag{4}$$

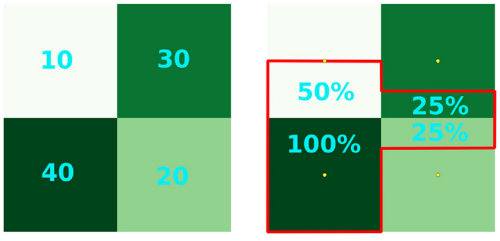

Fig. 1. A) Pixel values (left), and B) Boundary coverage of pixels (right).

Without pixel weights, zonal statistics results in an estimated mean of 25, while the coverage weighted method results in an estimated mean of 28.75 - approximately a 13% difference.

To mediate the resource demand of this procedure, GeoQuery implements coverage weighting by leveraging the ability to arbitrarily adjust the resolution at which boundary features are rasterized during zonal statistics. Instead of rasterizing at the same resolution as the measurement data, GeoQuery introduces a scaling factor to rasterize the boundary at a finer resolution (e.g., for a scaling factor of 10 rasterization would produce a 10 × 10 grid of binary values instead of a single binary value). This finer resolution rasterization is then aggregated back to the resolution of the measurement data to produce a percent coverage estimate[8] (e.g., the 100 binary values in the 10 × 10 grid are summed to produce a coverage weight percentage, x/100, at the measurement data raster resolution). Estimating coverage using this method avoids the increased computational complexity of determining the exact overlap (i.e., calculating intersection of each pixel's bounding box with the boundary geometry). The trade-off between accuracy and computational complexity results in imperfect coverage estimates, as seen in Fig. 2. In this example the coverage of the original top-right and bottom-right cells from above, which are actually 25%, are estimated as 30% with a scaling factor of 10.

2.2.3. Latitude weighting

Because the datasets leveraged in GeoQuery are collected over geographic space and represented using a 2D plane (i.e., a satellite image/raster data), the challenges of using geospatial projections to account for the three dimensional nature of the earth apply (Battersby et al., 2017; Madden et al., 2009). In particular, when aggregating data from an area represented by raster pixels, it is important to consider the physical area represented by each pixel, and how these areas may vary across pixels in a dataset. Understanding the area of a pixel requires knowing the projection information, or coordinate reference system (CRS), used for the data. Most widely used and publicly available global datasets utilize a geographic projection, CRS EPSG:4326 - commonly referred to as WGS84,[9] which uses latitude and longitude coordinates on the WGS84 reference ellipsoid. This CRS is used as a standard for all datasets in GeoQuery[10]. Using WGS84, the area represented by pixels in a raster dataset is dependent on latitude. In WGS84, as pixel observations approach the poles, lines of longitude converge and pixel area decreases (Fig. 3).

To accurately account for area when performing zonal statistics using WGS84 datasets, there are two potential methods: reprojecting the datasets to an equal area projection (Madden et al., 2009), or weighting pixels based on latitude (Kugler et al., 2015). Reprojecting a raster dataset involves resampling the underlying data in order to rebuild the surface using the new projection. This process can introduce changes into the data depending on the data type, resampling method, and raster resolution (Kugler et al., 2015; Nawrotzki et al., 2016). The second method, used in GeoQuery, involves (a) ensuring boundary datasets use WGS84 (a process that can be done with perfect accuracy given the vector format of boundary data (Kugler et al., 2015)), and (b) weighting pixels based on latitude during the zonal statistics stage of processing. Weights for the underlying raster data are generated for each row of pixels (representing measurements at some latitude), utilizing the Haversine distance formula (Van Brummelen, 2012) to

[8] The scaling factor of ten used by GeoQuery results in minimal added computational/memory costs. Larger scaling factors can provide greater accuracy but are slower and require more memory.

[9] "WGS84" can potentially be used ambiguously to refer to, for example, the WGS84 datum. In this paper "WGS84" will always refer specifically to the EPSG:4326 CRS.

[10] To date, all raster data used in GeoQuery has been made available in a geographic projection by the data providers and has not required additional reprojection.

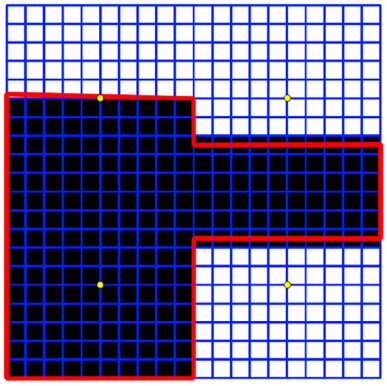

Fig. 2. In GeoQuery, the vector boundary (represented by the red boundary) is rasterized at a spatial resolution much finer than the measurement data we seek to aggregate (finer geospatial resolution illustrated by blue grid). This rasterized boundary is then aggregated back to the resolution of the measurement data (Fig. 1) to create a coverage estimate. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

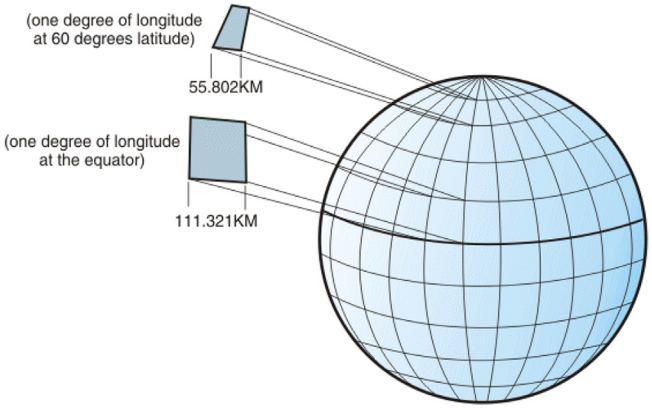

Fig. 3. Lines of longitude converging at poles (IBM Product Publications, 2018).

account for variable size of raster pixels as distance from the equator increases (Fig. 3). This approach incurs minimal additional computational costs during the zonal statistics process and requires no additional preprocessing or management of the data outside of the zonal statistics process.[11]

An illustration of this approach follows. Using the same example data from Fig. 1A, we assume that the raster has a hypothetical resolution of 40 decimal degrees,[12] and the top left corner is located at a latitude of 80° (using WGS84). Using the latitude at the center of each row of pixels (60° for the top row of raster, 20° for the bottom row of this example raster) the ratio of spherical distances calculated using the Haversine formula were 0.5 for the top row and 0.766 for the bottom row. The calculations in equations (5) and (6) take into account the underlying values of the raster data, the overlap between the boundary and raster pixels, and the weighting to account for the relative areas of each pixel.

raster x coverage weights x latitude correction = final data

$$\begin{bmatrix} 10 & 30 \\ 40 & 20 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0.25 \\ 1 & 0.25 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0.5 \\ 0.766 & 0.766 \end{bmatrix} = \begin{bmatrix} 2.5 & 3.75 \\ 30.64 & 3.83 \end{bmatrix} \quad (5)$$

$$Mean = \frac{\sum final\_data}{\sum (coverage\_weights * latitude\_correction)} = \frac{2.5 + 3.75 + 30.64 + 3.83}{0.5 * 0.5 + 0.25 * 0.5 + 1 * 0.766 + 0.25 * 0.766} = 30.56 \quad (6)$$

The resulting estimate of mean precipitation for the example boundary, 30.56, is approximately 18% different from the original estimate calculated in Section 2.2.2 and approximately 6% from the estimate incorporating only pixel weighting.

2.3. Web portal and request processing

Users access GeoQuery's data and functionality via a web portal, through which they are able to submit requests for particular measurement data aggregated to particular boundaries.[13] The user interface and experience of the portal was designed to allow users to quickly find and select the data needed for research by guiding them through a simple set of data pages that are intuitive to use. To submit a request, users first select a set of boundaries to use as units of analysis, which are displayed in a map view (Fig. 4), then select measurement datasets to be aggregated to the selected units.

The selection of measurement datasets (Fig. 5) allows users to set temporal filters and the desired aggregation methods such as sum, mean, maximum, minimum, or categorical values. All available statistical aggregation options are made available for each individual dataset unless the result of aggregation using the method would not be interpretable (e.g., a summation method is not available for a dataset containing categorical data, such as land cover).

Once a user has completed the request process, a record is created and entered into a database containing their selected units of analysis, measurement datasets, and processing options. Automated jobs running on the HPC cluster identify new requests, check for cached data, and manage preparing results. Once data for a user's request has been processed, custom documentation is generated and the user is sent an email linking them to a unique page for their request which serves as a permanent archive for their data and documentation.

3. Results and discussion

Since being launched in the second quarter of 2017, GeoQuery has run zonal statistics operations for over two million boundary-measurement data combinations.[14] As of August 2018, over 1750 users have submitted nearly 7000 data requests. Fig. 6 shows GeoQuery's usage since launch.

GeoQuery has been used in a range of projects and research. In addition to its use by local researchers, GeoQuery has been utilized by organizations including the World Bank, the Global Environment Facility, USAID, the Millenium Challenge Corporation, and the MacArthur Foundation. One of the largest groups of GeoQuery users has been dozens of universities and other educational institutions, including undergraduates, graduate students, and faculty. Multiple articles have been published using data from GeoQuery, including BenYishay et al. (2017a), Zhao et al. (2017a), Runfola et al. (2017b), Zhao et al. (2017b), and Marty et al. (2017), along with reports by the World Bank Group (World Bank Group, 2017), the Independent Evaluation Group of the World Bank (Runfola et al., 2017a), the Overseas Development Institute (ODI, 2017), and the Expert Group for Aid Studies (Isaksson, 2017).

[11] Distance between lines of latitude remains constant, so a measurement of longitudinal distance can be used to weight each pixel rather than the actual pixel area (i.e., latitude distance x longitude distance).

[12] Pixel size has been exaggerated to illustrate the effect of latitude on pixel area.

[13] User requests submitted through the GeoQuery web portal are flagged by GeoQuery as high priority tasks that will run before generic tasks generated by the automated processes described in Section 2.1.3.

[14] Where the boundary data is defined as all features for a given administrative level in a country (e.g., Afghanistan ADM1), and the measurement data is a single raster layer (e.g., VIIRS Nighttime Lights in 2014). The count of zonal statistic operations based on the individual features in each boundary data layer would be drastically higher.

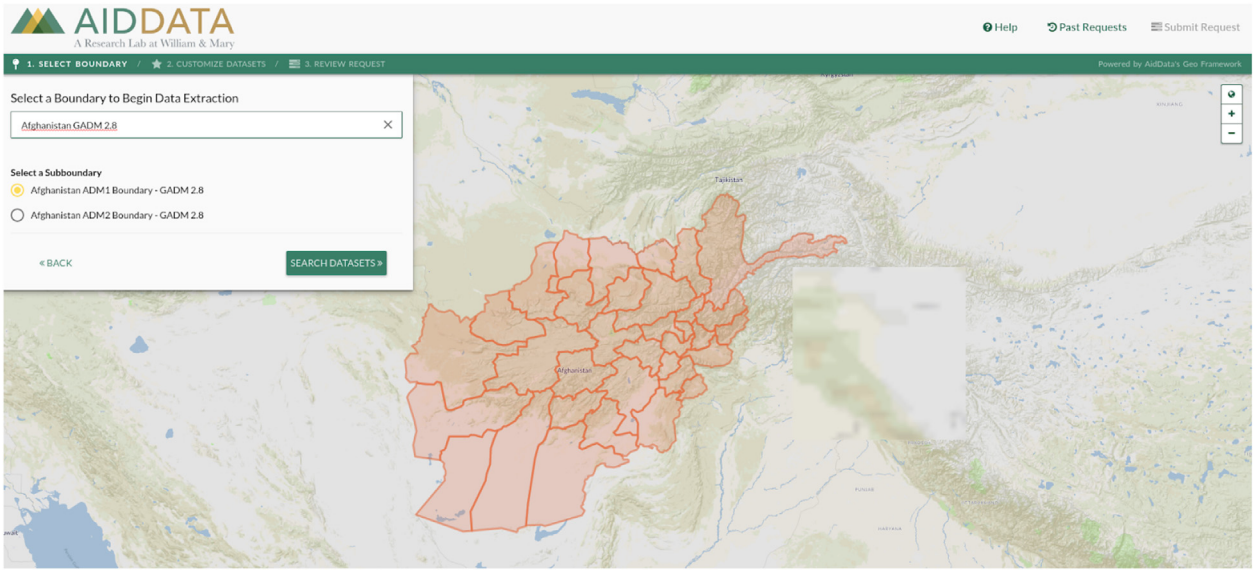

Fig. 4. GeoQuery boundary selection.

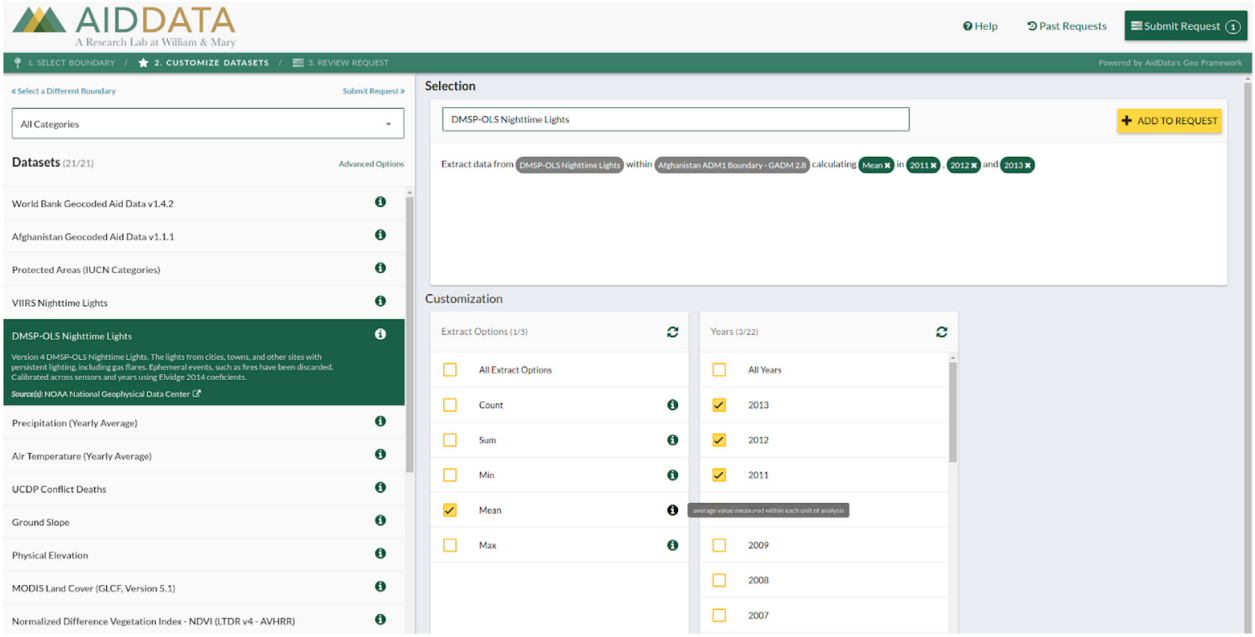

Fig. 5. GeoQuery measurement data selection.

To date, GeoQuery has seen the most uptake within the social sciences - particularly among research communities focused on global development and/or the intersection between society and nature - but still aims to serve a broader audience. The current, growing, collection of data available within GeoQuery is intended to support a wide range of research across disciplines, both by lowering the barrier to entry for researchers new to geospatial data, but also presenting a fundamentally technological improvement which will enable anyone to benefit from large scale data processing. A recent use case for GeoQuery as a tool within the geosciences is a geospatial impact evaluation performed for the Global Environment Facility (GEF-IEO, 2016, 2017) in which data from GeoQuery (NDVI, temperature, and precipitation) was used alongside supplemental data on tree cover (Hansen et al., 2013) and carbon stocks (Saatchi et al., 2011) to model and predict carbon sequestration around land degradation projects.

For any project using GeoQuery, it is important to consider the limitations and implications associated with the data, methods, and processing decisions used in GeoQuery. In any project using geospatial data there are numerous potential decisions to make: data to include, pre-processing methods, zonal statistics options, methodological considerations, and others. Given the large collection of datasets in GeoQuery, being used by a wide range of users in unknown applications, these types of decisions are made based on what will serve the broadest user base possible and impose the least limitations. Unfortunately, these decisions may not be right for everyone; the power of outsourcing large scale data processing to GeoQuery inherently comes at the cost of control over what data are available, and how the data are prepared and processed. While GeoQuery will continue to add additional datasets and options where possible and practical, users should consider these factors when using GeoQuery.

Important geospatial concepts which have not been discussed in this paper include the Modifiable Areal Unit Problem (MAUP) and ecological fallacy (Battersby et al., 2017; Openshaw, 1984, 1979). The MAUP and ecological fallacy both deal with the aggregation of geospatial data

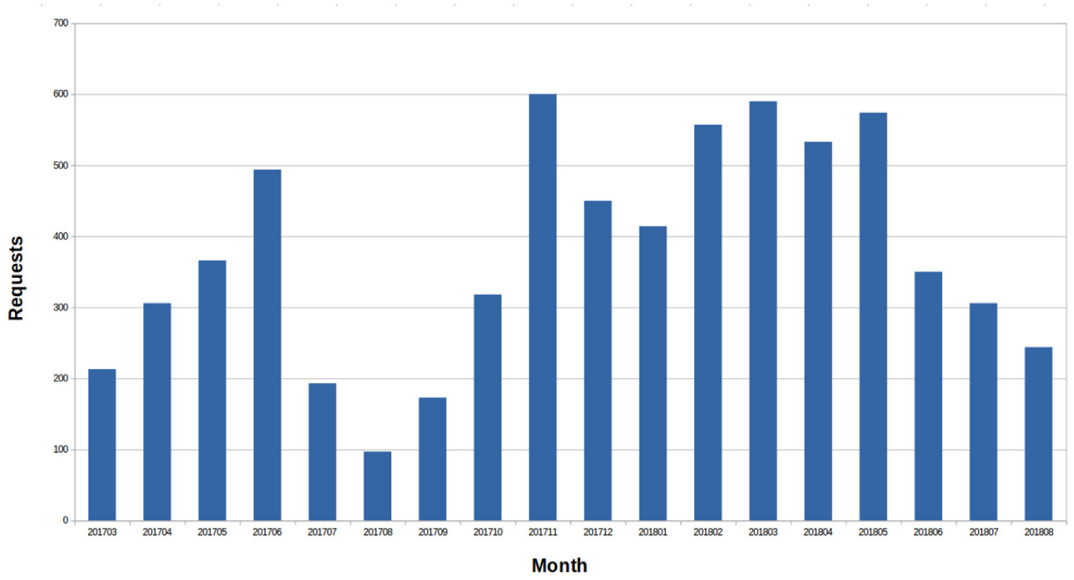

Fig. 6. GeoQuery usage by month.

and the interpretation of the results. While a discussion of these concepts, their implications, and methods for best handling them are beyond the scope of this paper, they are critical concepts users should review and understand when incorporating geospatial data into their work. GeoQuery does not attempt to solve these problems, but does provides documentation and supplemental information when possible to aid users in making informed decisions.

An important subset of what has been learned deals with the needs of GeoQuery's users. Since launch, the GeoQuery team has been in contact with nearly one hundred of GeoQuery's users in order to answer questions of processing and datasets, provide advice on how to best use data from GeoQuery in research, diagnose potential bugs during the initial beta, and listen to suggestions for new datasets and features. A key development early on which resulted from discussions with users involved improving the documentation provided with requests to help users better understand the format and content of their request results, and also providing basic tutorials on how to incorporate their request data into research using common statistical tools.

One of the most common requests to date has been the ability for users to upload custom boundary data. Dynamically incorporating user boundaries into GeoQuery presents a variety of technical challenges, but is being considered for future development. Discussions revolving around boundary data also helped identify the need for a fully open source set of global administrative boundary data and led to the creation of GeoBoundaries. There are also plans to incorporate other boundary data, including variable resolution global grids and protected areas, in the future. Other areas of future work include new measurement datasets, additional aggregation methods (e.g., fragmentation statistics), the potential for API access, and a visualization tool.

3.1. Conclusion

Geospatial data is an expansive source of information that is useful for a broad spectrum of researchers across disciplines. Researchers attempting to incorporate geospatial data into their work are faced with a multitude of potential sources and formats of data. Identifying, managing, and leveraging the right data for a particular application can be difficult without sufficient training and experience. GeoQuery presents a new solution to the technical elements of this challenge designed around a high performance computing environment using a flexible data framework, along with a web interface. By implementing a parallelized and automated set of processes, GeoQuery is able to handle user requests rapidly, averaging a few minutes or less. All requests come with customized documentation and a permanent page archiving data for future use, sharing, and replication. Additionally, GeoQuery's codebase and processing methodologies are open source. By reducing the barriers to finding and accessing geospatial data, GeoQuery aims to empower a broad range of data users across disciplines to produce new and meaningful research and insights.

Acknowledgements

This work was made possible by the support of USAID, KFW, Humanity United, the World Bank, the Global Environment 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.

We thank the many users of GeoQuery that have provided invaluable feedback. We would also like to thank Rachel Oberman, Lauren Hobbs, John Napoli, Leigh Seitz and other members of the GeoQuery lab who have contributed to GeoQuery.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2018.10.009.

References

Adamczyk, J., Tiede, D., 2017. Zonalmetrics - a python toolbox for zonal landscape structure analysis. Comput. Geosci. 99, 91–99. URL. http://www.sciencedirect.com/science/article/pii/S0098300416306586.

AidData, 2017. Subnationally geocoded world bank ibrd-ida projects. URL. http://aiddata.org/datasets.

Battersby, S.E., daan Strebe, D., Finn, M.P., 2017. Shapes on a plane: evaluating the impact of projection distortion on spatial binning. Cartogr. Geogr. Inf. Sci. 44 (5), 410–421. URL. https://doi.org/10.1080/15230406.2016.1180263.

BenYishay, A., Heuser, S., Runfola, D., Trichler, R., 2017a. Indigenous land rights and deforestation: evidence from the Brazilian Amazon. J. Environ. Econ. Manag. 86, 29–47. ISSN 0095-0696. https://doi.org/10.1016/j.jeem.2017.07.008. http://www.sciencedirect.com/science/article/pii/S0095069617304813.

BenYishay, A., Rotberg, R., Wells, J., Lv, Z., Goodman, S., Kovacevic, L., Runfola, D., 2017b. Geocoding afrobarometer rounds 1 - 6: methodology & data quality. URL. http://aiddata.org/data/geocoded-afrobarometer-data-v1.

Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., Ward, M., 2018. Icews

coded event data. URL. https://doi.org/10.7910/DVN/28075.
Bouma, M.J., Kovats, R.S., Goubet, S.A., Cox, J.S.H., Haines, A., 1997. Global assessment of El Niño's disaster burden. The Lancet. URL. http://www.sciencedirect.com/science/article/pii/S0140673697045091.
Bray, T., Mar. 2014. The JavaScript object notation (JSON) data interchange format. RFC 7159. URL. https://rfc-editor.org/rfc/rfc7159.txt.
Bruederle, A., Hodler, R., 2017. Nighttime lights as a proxy for human development at the local level. URL. https://ideas.repec.org/p/ces/ceswps/_6555.html.
Buhaug, H., Lujala, P., 2005. Accounting for scale: measuring geography in quantitative studies of civil war. Polit. Geogr. 24, 399–418.
Bundervoet, Tom, Maiyo, L.S.A., 2015. Bright lights, big cities: measuring national and subnational economic growth in africa from outer space, with an application to Kenya and Rwanda. URL. https://openknowledge.worldbank.org/handle/10986/22883.
Bunting, P., Clewley, D., Lucas, R.M., Gillingham, S., 2014. The remote sensing and gis software library (rsgislib). Comput. Geosci. 62, 216–226. URL. http://www.sciencedirect.com/science/article/pii/S0098300413002288.
Burgert, C.R., Colston, J., Roy, T., Zachary, B., 2013. Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys.
Burke, M., Heft-Neal, S., Bendavid, E., 2016. Sources of variation in under-5 mortality across sub-saharan africa: a spatial analysis. The Lancet Global Health 4 (12).
CIESIN, 2000. Gridded Population of the World (GPW). Tech. Rep. SEDAC URL. http://sedac.ciesin.columbia.edu/data/collection/gpw-v3.
CIESIN, 2017. Gridded Population of the World, Version 4 (Gpwv4): Population Count, Revision 10. Accessed on March 2016. URL. https://doi.org/10.7927/H4PG1PPM.
CIESIN, FAO, U., CIAT, 2015. Gridded population of the world, version 3 (gpwv3): population count grid. Accessed on March 2016. URL. https://doi.org/10.7927/H4639MPP.
CIESIN, ITOS, 20180322, 2013. Global roads open access data set, version 1 (groadsv1). URL. https://doi.org/10.7927/H4VD6WCT.
5Corrado, L., Fingleton, B., 2012. Where is the economics in spatial econometrics?*. J. Reg. Sci. 52 (2), 210–239. URL. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9787.2011.00726.x.
Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. Viirs night-time lights. URL. https://doi.org/10.1080/01431161.2017.1342050.
ESA, 2009. GlobCover. URL. http://due.esrin.esa.int/globcover/.
ESA, 2017. Esa land cover. URL. http://maps.elie.ucl.ac.be/CCI/viewer/index.php.
ESRI, 2016. How zonal statistics works. URL. http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/h-how-zonal-statistics-works.htm.
GDAL/OGR contributors, 2018. The GDAL/OGR Geospatial Data Abstraction Software Library. URL. http://gdal.org.
GEF-IEO, 2016. Value for Money Analysis for the Land Degradation Projects of the Gef. independent Evaluation Office of the Global Environment Facility, Washington DC URL. https://www.thegef.org/sites/default/files/council-meetingdocuments/EN_GEF.ME_C.51.Inf_02_VFM\%20Analysis\%20for\%20LD\%20Projects\%20o\%20GEF.pdf.
GEF-IEO, 2017. 2017 Ieo Brief: Land Degradation Focal Area Study. independent Evaluation Office of the Global Environment Facility, Washington DC URL. http://www.gefieo.org/sites/default/files/ieo/signposts/files/land-degradation-2017-brief.pdf.
Gilmore, E., Gleditsch, N.P., Lujala, P., Rød, J.K., 2005. Conflict diamonds: a new dataset. Conflict Manag. Peace Sci. 22 (3), 257–292.
Girardin, L., Hunziker, P., Cederman, L.-E., Vogt, M., 2015. Growup - geographical research on war, unified platform. URL. http://growup.ethz.ch/.
GLCF, 2017. Glcf landcover download. Global land cover facility. URL. http://glcf.umd.edu/data/landsat/.
Goodchild, M.F., 2008. Geographic information science: the grand challenges. In: The Handbook of Geographic Information Science. Blackwell, Malden, MA, pp. 596–608.
Goodchild, M.F., 2009. Geographic information systems and science: today and tomorrow. Procedia Earth and Planetary Science 1 (1), 1037–1043.
Goodman, S., BenYishay, A., Runfola, D., 2016. Overview of the GeoFramework. Tech. Rep. William and Mary, Williamsburg, VA URL. www.geoquery.org.
Granneman, S., 10, 2006. Archiving and compression. URL. https://www.linuxjournal.com/article/9370.
3Gregory, I., Donaldson, C., Murrieta-Flores, P., Rayson, P., 2015. Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. Int. J. Humanit. Arts Comput. 9 (1), 1–14. URL. http://www.euppublishing.com/doi/10.3366/ijhac.2015.0135.
Grosh, M.E., Glewwe, P., 1995. A Guide to Living Standards Measurement Study Surveys and Their Data Sets (English). Living Standards Measurement Study (Lsms) Working Paper No. Lsm 120. URL. http://documents.worldbank.org/curated/en/270551468764720584/A-guide-to-living-standards-measurement-study-surveys-and-their-data-sets.
GSDI, 2012. Spatial Data Infrastructure Cookbook. Viewed on 9 April 2018. URL. http://gsdiassociation.org/index.php/publications/sdi-cookbooks.html.
Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853. URL. http://science.sciencemag.org/content/342/6160/850.
Henderson, J.V., Storeygard, A., Weil, D.N., April 2012. Measuring economic growth from outer space. Am. Econ. Rev. 102 (2), 994–1028. URL. http://www.aeaweb.org/articles?id=10.1257/aer.102.2.994.
Hijmans, R., 2015a. Gadm Database of Global Administrative Areas - Version 2.8. URL. http://gadm.org/.
Hijmans, R., 2015b. Raster: Geographic Data Analysis and Modeling. R Package Version 2.5-2. URL. http://CRAN.R-project.org/package=raster.
IBM Product Publications, March 2018. Ibm knowledge center- geographic coordinate system. URL. https://www.ibm.com/support/knowledgecenter/en/SSEPEK_11.0.0/spatl/src/tpc/spatl_csb3022a.html, Accessed date: 21 March 2018.
Isaksson, A.-S., 2017. Geospatial Analysis of Aid: a New Approach to Aid Evaluation, Expertgruppen För Biståndsanalys. EBA URL. http://eba.se/wp-content/uploads/2017/10/Geospatial-Isaksson-webb.pdf.
IUCN, UNEP-WCMC, 2016. The world database on protected areas (wdpa). URL. www.protectedplanet.net.
Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the Globe Version 4. URL. http://srtm.csi.cgiar.org.
Kim, D., Sarker, M., Vyas, P., Feb, 2016. Role of spatial tools in public health policy-making of Bangladesh: opportunities and challenges. J. Health Popul. Nutr. 35 (1), 8. URL. https://doi.org/10.1186/s41043-016-0045-1.
Kugler, T.A., Van Riper, D.C., Manson, S.M., Haynes, D.A., Donato, J., Stinebaugh, K., 2015. Terra populus: workflows for integrating and harmonizing geospatial population and environmental data. J. Map Geogr. Libr. 11 (2), 180–206. URL. http://www.terrapop.org.
Leetaru, K., Schrodt, P.A., 2013. Gdelt: global data on events, location, and tone, 1979–2012. ISA Annual Convention 2 (4).
Lujala, P., 2009. Deadly combat over natural resources: gems, petroleum, drugs, and the severity of armed civil conflict. J. Conflict Resolut. 51 (1), 50–71.
Lujala, P., Rød, J.K., Thieme, N., 2007. Fighting over oil: introducing a new dataset. Conflict Manag. Peace Sci. 24 (3), 239–256.
Madden, M., for Photogrammetry, A.S., Sensing, R., 2009. Manual of Geographic Information Systems. American Society for Photogrammetry and Remote Sensing, Bethesda, Md (includes bibliographical references and index).
Malvárez, G.C., Pintado, E.G., Navas, F., Giordano, A., Oct, 2015. Spatial data and its importance for the implementation of unep map iczm protocol for the mediterranean. J. Coast Conserv. 19 (5), 633–641. URL. https://doi.org/10.1007/s11852-015-0372-1.
Marty, R., Dolan, C.B., Leu, M., Runfola, D., 2017. Taking the health aid debate to the subnational level: the impact and allocation of foreign health aid in Malawi. BMJ Global Health 2 (1) URL. http://gh.bmj.com/content/2/1/e000129.
Matsuura, K., Willmott, C.J., 2015a. Terrestrial Air Temperature: 1900-2014 Gridded Monthly Time Series. URL. http://climate.geog.udel.edu/~climate/html_pages/download.html#T2014.
Matsuura, K., Willmott, C.J., 2015b. Terrestrial Precipitation: 1900-2014 Gridded Monthly Time Series. URL. http://climate.geog.udel.edu/~climate/html_pages/download.html#P2014.
Miller, H.J., Goodchild, M.F., Aug 2015. Data-driven geography. Geojournal 80 (4), 449–461. URL. https://doi.org/10.1007/s10708-014-9602-6.
Minnesota Population Center, 2015. Integrated Public Use Microdata Series, International: Version 6.4.
MODIS, 2013. MODIS land cover. URL. https://lpdaac.usgs.gov/products/modis_products_table.
MongoDB, 2018. Mongodb. URL. http://mongodb.com.
Munson, M.A., May 2012. A study on the importance of and time spent on different modeling steps. SIGKDD Explor. Newsl. 13 (2), 65–71. URL. http://doi.acm.org/10.1145/2207243.2207253.
NASA, 2000. Shuttle-radar topography mission. URL. http://www2.jpl.nasa.gov/srtm/.
NASA, 2017a. Long term data record. URL. https://ltdr.modaps.eosdis.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi?fileName=products.
NASA, 2017b. LTDR (Land Long Term Data Record) Home. Tech. Rep. NASA URL. https://ltdr.modaps.eosdis.nasa.gov/cgi-bin/ltdr/ltdrPage.cgi.
Nawrotzki, R.J., Schlak, A.M., Kugler, T.A., 2016. Climate, migration, and the local food security context: introducing terra populus. Popul. Environ. 38 (2), 164–184.
Nelson, A., 2008. Estimated Travel Time to the Nearest City of 50,000 or More People in Year 2000. URL. http://forobs.jrc.ec.europa.eu/products/gam/.
NOAA Earth Observation Group, 2017. Dmsp-ols nighttime lights. URL. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.
ODI, 2017. Aid Allocation within Countries: Does it Go to Areas Left behind? overseas Development Institute URL. https://www.odi.org/sites/odi.org.uk/files/resource-documents/11658.pdf.
Openshaw, S., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Spatistica applications in the spatial sciences 127–144 URL. https://ci.nii.ac.jp/naid/10009667572/en/.
Openshaw, S., 1984. The modifiable areal unit problem. Concepts and Techniques in Modern Geography. URL. https://ci.nii.ac.jp/naid/10024464407/en/.
Perry, M., 2016. Rasterstats: Summarize Geospatial Raster Datasets Based on Vector Geometries. python Package Version 0.10.3. URL. https://pypi.python.org/pypi/rasterstats/0.10.3.
Rajabifard, A., Williamson, I.P., 2001. Spatial data infrastructures: concept, sdi hierarchy and future directions. URL. http://hdl.handle.net/11343/33897.
Raleigh, C., Linke, A., Hegre, H., Karlsen, J., 2010. Introducing acled: an armed conflict location and event dataset: special data feature. J. Peace Res. 47 (5), 651–660. URL. https://doi.org/10.1177/0022343310378914.
Rueda, C., Greenberg, J., Ustin, S., 2005. Starspan: a Tool for Fast Selective Pixel Extraction from Remotely Sensed Data.
Runfola, D.M., A, B., G, B., J, N., J, T., 2017a. Environmental Impact of World Bank Projects: a Case Study of Integrating Value for Money Analysis into Impact Evaluations. What Works? Value for Money in Impact Evaluations. Independent Evaluation Group, the World Bank.
Runfola, D., BenYishay, A., Tanner, J., Buchanan, G., Nagol, J., Leu, M., Goodman, S., Trichler, R., Marty, R., 2017b. A top-down approach to estimating spatially heterogeneous impacts of development aid on vegetative carbon sequestration.

Sustainability 9 (3) URL. http://www.mdpi.com/2071-1050/9/3/409.
Saatchi, S.S., Harris, N.L., Brown, S., Lefsky, M., Mitchard, E.T.A., Salas, W., Zutta, B.R., Buermann, W., Lewis, S.L., Hagen, S., Petrova, S., White, L., Silman, M., Morel, A., 2011. Benchmark map of forest carbon stocks in tropical regions across three continents. Proc. Natl. Acad. Sci. Unit. States Am. 108 (24), 9899–9904. URL. http://www.pnas.org/content/108/24/9899.
Salehyan, I., Hendrix, C.S., Hamner, J., Case, C., Linebarger, C., Stull, E., Williams, J., 2012. Social conflict in africa: a new database. Int. Interact. 38 (4), 503–511. URL. https://doi.org/10.1080/03050629.2012.697426.
Sara, B., Maggioni, M.A., 2014. Blood diamonds, dirty gold and spatial spill-overs: measuring conflict dynamics in western africa. Peace Econ. Peace Sci. Publ. Pol. 20 (4), 551–564.
Seitz, L., Lv, Z., Goodman, S., Runfola, D., 2018. Geoquery user's guide - chapter 3: geoboundaries – a global, redistributable map of administrative zones. URL. http://www.geoboundaries.org.
Sundberg, R., Melander, E., 2013. Introducing the ucdp georeferenced event dataset. J. Peace Res. 50 (4), 523–532. URL. https://doi.org/10.1177/0022343313484347.
Tollefsen, A.F., Strand, H., Buhaug, H., 2012. Prio-grid: a unified spatial data structure. J. Peace Res. 49 (2), 363–374. URL. grid.prio.org.
USAID, 2015. Higher Education Solutions Network - Bi-annual Report (Fy2015). aidData Center for Development Policy, Washington DC URL. http://pdf.usaid.gov/pdf_docs/PA00KSXN.pdf.
USGS, 2017. Usgs earthexplorer. URL. https://earthexplorer.usgs.gov/.
Van Brummelen, G., 2012. Heavenly Mathematics: the Forgotten Art of Spherical Trigonometry. Princeton University Press.
Wessel, P., Smith, W.H.F., 1996. A global, self-consistent, hierarchical, high-resolution shoreline database. J. Geophys. Res.: Solid Earth 101 (B4), 8741–8743. URL. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JB00104.
World Bank Group, 2017. Macroeconomics & Fiscal Management. myanmar Economic Monitor. URL. http://www.academia.edu/35139152/MYANMAR_ECONOMIC_FORECAST_ANALYSIS_2017-2020.
Zandbergen, Barbeau, S.J., 2011. Positional accuracy of assisted gps data from high-sensitivity gps-enabled mobile phones. J. Navig. 64, 381–399.
Zhao, J., Kemper, P., Runfola, D., 2017a. Quantifying heterogeneous causal treatment effects in World Bank development finance projects. In: Data Mining and Knowledge Discovery (ECML PKDD), URL. http://ecmlpkdd2017.ijs.si/papers/paperID507.pdf.
Zhao, J., Kemper, P., Runfola, D., 2017b. Simulation study in quantifying heterogeneous causal effects. URL. http://informs-sim.org/wsc17papers/.