Global Gridded Electricity: a high spatial resolution dataset of world electricity consumption

Stephen Jarvis[1*]

January 11, 2018

1. Energy & Resources Group, University of California at Berkeley
*Correspondence and requests for materials: jarviss@berkeley.edu

Abstract

Recent years have seen a profusion of new spatial datasets mapping global socioeconomic data at fine spatial scales. The Global Gridded Electricity (GGE) dataset is the first such dataset to map global demand for electricity. This new resource will assist with research studying the critical role that geography plays in debates on development, urbanization, electrification, renewables integration and climate change. The input electricity consumption data was collected from a diverse range of national and subnational sources. The predictive layers used in the analysis include satellite data (nightlights, land cover and elevation), weather station data (temperature and degree days), volunteered geographic data (electricity infrastructure) and socioeconomic data (population, electricity access, economic output and employment). These two are combined using a "Random Forest" regression approach to create a set of 30-arc-second global grids that map electricity consumption at scales of approximately one square kilometer. Data for 2000, 2005, 2010 and 2015 can now be downloaded from the GGE data repository. The dataset and the methodology used to create it are described here.

Background & Summary

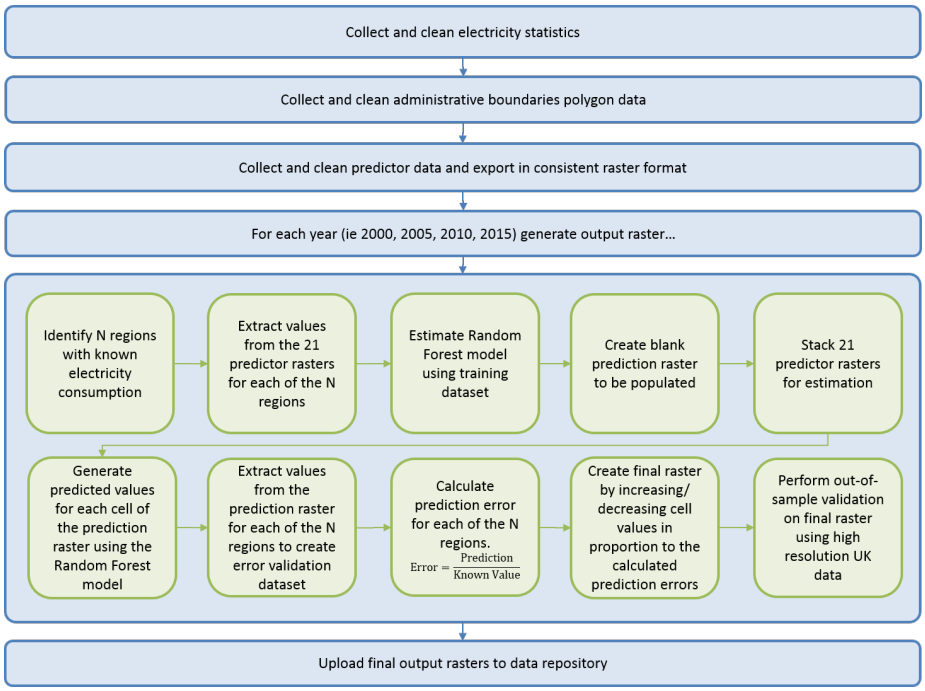
Electricity consumption varies dramatically across the globe, from dense urban centers to remote rural regions. In 2015 an average square kilometer of Washington DC required 130 times more electricity than an average square kilometer of Washington state [1]. Economic growth and electrification are also responsible for continued rapid growth in electricity demand around the world, particularly in developing countries. The twenty years between 1995 and 2015 saw world electricity demand grow by 68%, driven in large part by a 184% increase in India and a 363% increase in China [2]. The geography of where this electricity demand is located plays a central role in a number of critical academic and policy

debates spanning issues as diverse as development, energy access, urbanization, renewables integration, distributed generation and climate change.

Despite the important role that geography plays in the modern economy and the power grid that supports it, detailed spatial data on electricity consumption is limited. National data on electricity consumption is relatively comprehensive for recent decades, but has poor spatial resolution. Detailed subnational data for electricity consumption is increasingly available and the spatial resolution for these data can sometimes be very good, with countries like the UK publishing electricity consumption statistics for over four hundred subnational regions. However, there are currently only a few developed countries providing this kind of information. The patchwork of available data means there is a real gap in our capacity to analyze the geographic distribution of electricity consumption. This makes conducting analysis at the scale of towns or cities challenging, particularly in a developing country context.

The GGE dataset aims to remedy the limited nature of available data on electricity demand. The dataset comprises a time series of global electricity consumption data with a 30-arc-second spatial resolution (approximately one square kilometer). It was created by first compiling all available national and subnational data on electricity consumption. Additional data was then collected on potential predictor variables that are correlated with electricity consumption and also available at finer spatial resolution. The predictor variables used included satellite data (nightlights, land cover and elevation), weather station data (temperature and degree days), volunteered geographic data (power grid infrastructure) and socioeconomic data (population, electricity access, sectoral economic output and employment). The dataset was produced using a dasymmetric redistribution approach based on the methods developed by Stevens et. al. (2015) and recently employed in the WorldPop dataset [49, 51]. In each year for which the dataset was produced, a cross-sectional training dataset was created, with log electricity consumption as the dependent variable and the set of predictor covariates as the independent variables. A flexible non-parametric predictive model was then estimated using "Random Forests" regression. This model was used to predict electricity consumption at a spatial resolution of approximately one square kilometer using the predictor variables. Some final proportional adjustments were made to these predictions to ensure the original "known" national totals could be recovered from the final dataset. Out-of-sample prediction performance was then used to validate the quality of the dataset. This first release of the dataset includes a set of GeoTIFF layers for 2000, 2005, 2010 and 2015.

Figure 1: Dataset Production Diagram



Methods

Data collection and processing

Electricity consumption data

Data on national annual electricity consumption was collected for all countries for the period 2000 to 2015. The primary source of national-level data was the United Nations Energy Statistics Database [3]. This was supplemented by additional national data from the International Energy Agency's Energy Balance Database [4].

Equivalent data was then collected from individual country institutions for as many subnational regions as possibles, such as states, counties, provinces etc. This subnational data came from a range of institutions, including national statistics offices, government ministries and departments, regulatory agencies, power grid companies and power system operators. Subnational data was collected for Australia, Argentina, Austria, Brazil, Canada, Chile, China, France, Germany, India, Italy, Mexico, the Netherlands, Norway, Poland, Spain, South Africa, South Korea, Sweden, the United Kingdom and the United States [5–27].

These data were collated together and converted to kilowatt-hours (kwh) for consistency. Often there were small discrepencies between the national values

reported by the UN and the summed values from the subnational data reported by various national institutions. These were reconciled by treating the national data from the UN as the "true" value and then re-scaling any subnational values such that they correctly summed to the UN national totals. The magnitude of any adjustments made during this process were small. It should also be noted that in principle the national electricity data used here should capture all consumption of electricity, irrespective of its source [52]. In practice collecting all data relating to consumption of off-grid, self-generated or distributed electricity is challenging, and the guidance for the subnational data is less comprehensive. Even so, where there are omissions it is likely that these are a very small share of total electricity use.

This data collection exercise resulted in observations of electricity consumption for 591 distinct regions of the globe in 2000, 882 in 2005, 906 in 2010 and 925 in 2015. These observations ranged from entire countries (e.g. Bolivia) to small subnational regions (e.g. Livorno Province in Tuscany, Italy).

Economic data

Data on the following economic variables was collected: total population, gross domestic product (GDP), total employment and the proportion of the population with access to electricity. For GDP and employment additional information was collected on how these were divided amongst three broad sectors: agriculture, industry and services. As with the electricity consumption data, the data collection process for these various economic statistics involved starting with national data. This was then combined with as much spatially disaggregated subnational data as possible. The same procedure of rescaling to match the national totals was employed here. For almost all countries subnational data was available at an equivalent or more disaggregated level for these indicators than was the case for electricity consumption.

The World Bank Development Indicators were the key source of national data [28]. These were supplemented by additional national data collected directly from the United Nations, particularly on population and GDP where the UN statistics included additonal coverage for a number of smaller countries [29, 30]. Subnational data for European Union countries was taken from Eurostat's Regional Statistics Database [31]. Subnational data for US states and counties was taken from the Bureau of Economic Analysis (BEA) [32]. Subnational data for other major economies was taken from the Regional Databases of the Organisation for Economic Cooperation and Development (OECD) [33, 34]. Whilst the OECD database already includes observations for subnational regions in European countries and the US, the data available directly from Eurostat and the BEA was more comprehensive and covered a finer disaggregation of subnational regions (e.g. US counties). Where data on GDP was unavailable, data on gross value added (GVA) or earnings was collected instead and re-scaled to match the GDP series. The GVA and earnings data were particularly valuable for calculating the sectoral shares of GDP because statistics on economic output by industry are usually reported as GVA rather than GDP. Once they had been

collected all series were converted to consistent units. For total GDP this was nominal US dollars. For total population or total employment this was numbers of people or employees. For the sectoral breakdowns of GDP and employment and the population with electricity access, these were calculated as percentage shares.

An important part of the construction of the economic data was the need to impute missing values in order to have a consistent set of data that could be exported in raster format. Without this, the predictive model would produce a missing value for any cell that did not have complete data. Imputation of missing values in the economic data was primarily achieved through a combination of 1) linearly interpolating across time within a given series, 2) using the data from the more complete series (e.g. population) to extrapolate the less complete series (e.g. employment) across time and 3) using data from the more complete aggregated administrative regions (e.g. states) to extrapolate to less aggregated administrative regions (e.g. counties). Any missing data after these steps were taken generally covered a very small portion of the data on sectoral shares for a number of small island nations (e.g. the Falkland Islands). To remedy this the remaining missing data was filled in with a simple global average value. This was done using the population-weighted average of the national values for each series using the non-missing data. Importantly, the data imputation described here (particularly this last step imputing with a global average value) is unlikely to significantly bias the production of the final dataset due to the dasymmetric approach that is employed. Full details on the approach taken can be found in the accompanying code. For 2015 this harmonization and imputation exercise resulted in a cross-section of just over 5500 distinct regions of the globe with observations of population, GDP, employment, sectoral shares of GDP and employment, and share of the population with access to electricity. Again these ranged from entire countries to small subnational regions.

Finally, in order to conduct the prediction process all the predictor data needed to be converted into a consistent 30 arc-second raster format. To do this for these economic predictors, the data for a given region was converted to per capita values and then assigned to the cells that fell within their corresponding administrative boundary polygon (see description of administrative boundaries below). These regional tiles of cells were then combined together (i.e. "mosaiced") to form a complete per capita global 30-arc-second raster image for each of the economic predictor variables. This approach was preferred to rasterizing the polygons directly as it proved to be computationally more efficient and did not fail in the event of overlaps from slightly mismatched polygons. In cases where there were overlapping cells, these were reconciled by taking the mean. The resulting per capita rasters were then multiplied by a corresponding population density raster image (described below) to get the desired final set of economic density rasters. For example, the density of economic output for each 30-arc-second cell, i, was calculated as:

$$\begin{pmatrix} GDP \\ percapita \end{pmatrix} \times \begin{pmatrix} Population \\ density \end{pmatrix} = \frac{GDP_i}{POP_i} \times \frac{POP_i}{AREA_i} = \frac{GDP_i}{AREA_i} = \begin{pmatrix} GDP \\ density \end{pmatrix}$$

This approach of combining per capita values with a population density raster is consistent with the methodology used to create the economic raster data in the GEcon dataset [49]. Effectively it amounts to allocating the economic data (in this case GDP) within a given region in proportion to population.

Population data

Population data was taken from the Gridded Population of the World (GPW) dataset published by the Socioeconomic Data and Applications Center (SEDAC), which produces the GPW dataset in collaboration with the US National Aeronautics and Space Administration (NASA) [35]. This dataset provides five-yearly global 30 arc-second raster files for population that are based on national censuses. Data for 2000, 2005, 2010 and 2015 were taken from GPW v4. The native format of these data are annual 30-arc-second GeoTIFF images.

Nightlights data

The nighttime lights data were taken from the US National Oceanic and Atmospheric Administration (NOAA). Satellite data on nighttime lights has already been used widely in research as a predictor of economic activity, including electricity consumption. The link to electricity consumption is both direct (e.g. street lights powered by electricity are a key source of nighttime lights) and indirect (e.g. sources of light at night are generally concentrated in urban areas where electricity consumption is high).

The data series used for 2000, 2005 and 2010 is the DMSP nightlights data which provides annual global 30-arc-second GeoTIFFs for 1992 to 2013 [36]. A well-known drawback of the DMSP data is that the upper level of its brightness scale is topcoded, creating saturation in urban centers. This paper deals with this by using the radiance calibrated version of these data which corrects for the saturation effect.

For 2015 the more recent VIIRS nightlights data was used [37]. The native format of these data are annual 15-arc-second GeoTIFF images, and so these were resampled to the desired 30-arc-second resolution.

Finally the extent of the nighttime lights data does not go above a latitude of 75°N, whilst the output dataset is created up to a latitude of 85°N. As such the nighttime lights data is assumed to have values of zero brightness in this region of missing data. This seems reasonable given the sparseness of any settlements at these high latitudes.

Land Cover data

The land cover data were taken from the Climate Change Initiative (CCI) at the European Space Agency (ESA) [38]. Land cover data has been used in previous research as predictors of population distributions. The rationale is that people tend to live in urban areas or near cropland rather than in remote regions covered by desert, ice or dense forest. The CCI land cover dataset

classifies land cover into thirty detailed categories. These were summarized into eight broad categories in line with other research using these data [47]. These were urban, cropland, wetland, shrubland, grassland, forest, bare/desert and water/snow/ice. The native format of these data is a multi-band 10-arc-second GeoTIFF image, with cells being categorical. Annual data was extracted from the relevant band and resampled to a 30-arc-second resolution. Processing these data resulted in eight distinct images where each cell value was the proportion of that cell covered by a given classification type. Finally the water/ice/snow layers were also used to create a separate water mask layer for each analysis year. The mask was defined by cells with 100% water coverage.

Elevation data

The elevation data are from the Global 30 Arc-Second Elevation (GTOPO30) dataset created by the US Geological Survey (USGS) [39]. Elevation data have been used in previous research as predictors of population distributions. The rationale is that people tend to live in low-lying areas rather than in mountainous regions. The native format of these data is a 30-arc-second GeoTIFF image. The same elevation file was used for all years and values refer to meters of elevation above or below sea level.

Temperature data

The temperature data are the Monthly Land + Ocean Average Temperature global grid data from Berkeley Earth [40]. Temperature data has been used in countless studies to explain variation in electricity consumption. The relationship between temperature measures and electricity consumption is again both direct (e.g. hot temperatures increasing demand for air conditioning and cold temperatures increasing demand for electric heating) and indirect (e.g. parts of the world with extreme high or low temperatures tend to be more sparsely inhabited and less developed). To capture within-year variation in the data around the annual average, cooling degree days (CDDs) and heating degree days (HDDs) were calculated. Degree days were calculated relative to a "bliss point" temperature of 18°C. Because these are monthly data, degree days were calculated by assuming the deviation between average monthly temperatures and 18°C was constant on all days of the month.

Usually degree days are calculated using daily data, and so there is a risk in using monthly averages that degree days would be incorrectly estimated. To check this an equivalent exercise was conducted using the experimental daily dataset recently published by Berkeley Earth. The agreement was good with no obvious signs of systematic bias and so the monthly data appears suitable for the present application. The daily dataset is not used here because it is still classed as experimental and because it only provides data for land areas. The relatively coarse resolution of the data means having land-only observations creates challenges extracting values for small coastal or island regions.

The native format of these data is a 1-degree NetCDF file. Whilst this is a

lower resolution than the other spatial datasets used, changes in temperature across space are relatively gradual at these spatial scales. Annual data was extracted from the relevant band and resampled to a 30-arc-second resolution.

Grid Infrastructure data

The grid infrastructure data are from Open Street Map (OSM) [41]. Clearly an important predictor of electricity consumption in a given area will be the presence of the necessary power grid infrastructure. OSM contains spatial information on a wide range of power grid infrastructure, including substations, power lines and generation stations. Here the choice was made to focus on substations. This was done for two reasons. First, substations tend to capture points where voltages are being transformed, usually for the purpose of local distribution to end consumers. Being near to a substation is likely to be a powerful indicator of whether areas are connected to the grid, and so can help differentiate between on- and off-grid areas, particularly in rural settings. This is potentially less true for other types of grid infrastructure such as power stations or high voltage transmission lines, which are further from end consumers in the electricity supply chain. Second, the coverage of the OSM data on substations appears to be much more comprehensive than the data on power lines, particularly lower voltage distribution lines.

To download the data any objects tagged as substations were identified and saved as geoJSON files. These were then read in as points and polygons. The vast majority were polygons, but in reality these polygons merely showed the footprint of the actual substation (i.e. the boundary created by the walls or fences surrounding the electrical equipment). The polygons were therefore converted to points for consistency by using the coordinates of the polygon centroid. This gave a dataset of latitude and longitude coordinates for just over 254,000 substations. These points were then converted to a global 30 arc-second raster such that the value in each cell was the distance in kilometers to the nearest substation. Because of computational constraints the distances were calculated for a 15-arc-minute raster and then resampled to a 30-arc-second resolution. Whilst this does mean a loss of precision, visual inspection of the resulting dataset indicates that the resulting image still performs well at identifying regions that are significant distances ($>$25km) from grid infrastructure.

The choice was made to use the distance to the nearest substation because it is particularly robust to the variable coverage of the OSM data. Because the data are created by volunteer contributors some regions appear to have much more complete coverage of power grid infrastructure. This issue is even more pronounced for the power lines data, particularly with regards to lower voltage distribution lines. Focusing on the distance to the nearest reported substation increases the likelihood of consistently producing reliable estimates, even in areas where there is apparent underreporting. This is because there is often substantial clustering of substations, and so only a small subset of the universe of grid substations needs to be reported before the nearest distance approaches the "true" value.

Finally, it was not possible to create a time series of these rasters for each of the analysis years. This is because OSM was only started in 2004 and user contributions are continually being updated as coverage improves. As such the same 2017 snapshot of observations of grid infrastructure had to be used as a predictor in all versions of the dataset. Whilst this is not ideal, it is not unreasonable to think that today's grid infrastructure can still provide valuable information on the past. This is particularly true in more developed areas where access to electricity has already reached saturation.

Administrative Boundaries

Shapefiles of administrative boundaries were taken from the Global Administrative Areas database (GADM) [42]. These were supplemented by additional shapefiles of European regional areas from Eurostat, UK local authorities from the Office of National Statistics, Canadian Census Divisions from Statistics Canada, Australian Statistical Areas from the Australian Bureau of Statistics and French regions from France's Etalab [43–47]. In using these data the choice was made to convert each polygon in each shapefile into an eqivalent raster object of 30-arc-second cells. This allowed for significant improvements in processing times.

Predictive model estimation

In order to estimate a predictive model a training dataset had to be created. To do this the electricity consumption data was combined with the corresponding data for the predictor variables. To do this the administrative boundaries of each region were used to identify the cells in a global 30-arc-second grid that fell within that region. The values in the identified cells were then extracted and averaged to get average values for each region. When averaging for a given region, the values of each cell had to be weighted by the area of the cells. This is because the curvature of the Earth means that rasters projected in degree space have high or low latitude cells that represent smaller km^2 areas than mid latitude cells. This weighting by area was done by assuming a spherical globe such that the area of a given cell was the area of a cell at the equator multiplied by the cosine of the cell's latitude (in degrees), where a 30-arc-second cell at the equator was assumed to be 0.928km by 0.921km. This approach was validated by comparing the values extracted from the population and economic variable rasters with the data that was originally collected from the World Bank, UN, Eurostat, BEA and OECD. These checks indicated the extraction approach had negligible impacts on accuracy whilst substantially improving processing times. In 2015 the result was a cross-sectional training dataset with 925 observations, where each observation was a region where electricity consumption was observed matched with its respective values for each of the predictors.

A Random Forest regression model was estimated using the cross-sectional training dataset for years 2000, 2005, 2010 and 2015 [49]. The dependent variable was the log of electricity consumption per square kilometer. The log was

Table 1: Description of Predictors

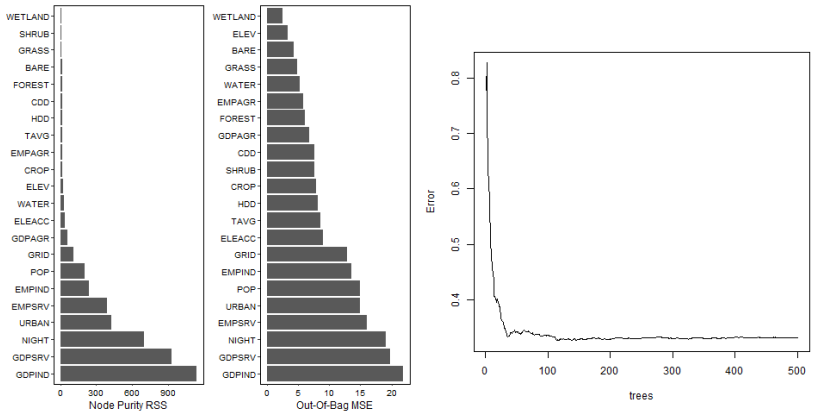
Predictor Name	Abbreviation	Units	Source
Agricultural GDP	GDPAGR	USD$/km^2	[28–35, 42–47]
Industrial GDP	GDPIND	USD$/km^2	[28–35, 42–47]
Services GDP	GDPSRV	USD$/km^2	[28–35, 42–47]
Agricultural employment	EMPAGR	employees/km^2	[28–35, 42–47]
Industrial employment	EMPIND	employees/km^2	[28–35, 42–47]
Services employment	EMPSRV	employees/km^2	[28–35, 42–47]
Electricity access	ELEACC	%	[28, 42]
Population	POP	people/km^2	[35]
Nighttime brightness	NIGHT	unitless	[36, 37]
Urban area	URBAN	%	[38]
Cropland area	CROP	%	[38]
Wetland area	WETLAND	%	[38]
Shrubland area	SHRUB	%	[38]
Grassland area	GRASS	%	[38]
Forest area	FOREST	%	[38]
Bare/desert area	BARE	%	[38]
Water/ice/snow area	WATER	%	[38]
Elevation	ELEV	m	[39]
Average temperature	TAVG	°C	[40]
Cooling degree days	CDD	°C-days	[40]
Heating degree days	HDD	°C-days	[40]
Distance to substation	GRID	km	[41]

used as this provided a better fit and is consistent with the approach taken by Lloyd et. al. (2017) [51]. The independent variables were the 21 predictor variables shown in Table 1. Figure 2 shows some key model summary statistics. Figure 2a plots two measures of the importance of each of the predictor variables. As might be expected, the most important variables include economic activity associated with industrial and service sectors as well as the variables for urban land cover and night-time lights. Figure 2b illustrates how the model's Out-Of-Bag error declines and converges rapidly as the number of trees increases.

Dataset production

To produce the dataset the complete set of predictor rasters for each year were stacked and the model predictions were generated for each 30-arc-second cell. The Random Forest model generates an ensemble of prediction values for each cell and so various approaches were tested when summarizing the outputs of the model to a single estimate for each cell (e.g. mean, median etc.). The median was found to perform the best when predictions were compared against the "known" values in the original training dataset (see Figure 4). Furthermore, because the log of electricity consumption was used as the dependent variable the final values needed to be transformed back to get the desired unlogged value. Fortunately another desirable property of opting for the median is that this got

Figure 2: Model Summary Statistics



(a) Predictor Importance (2015) (b) Error Convergence (2015)

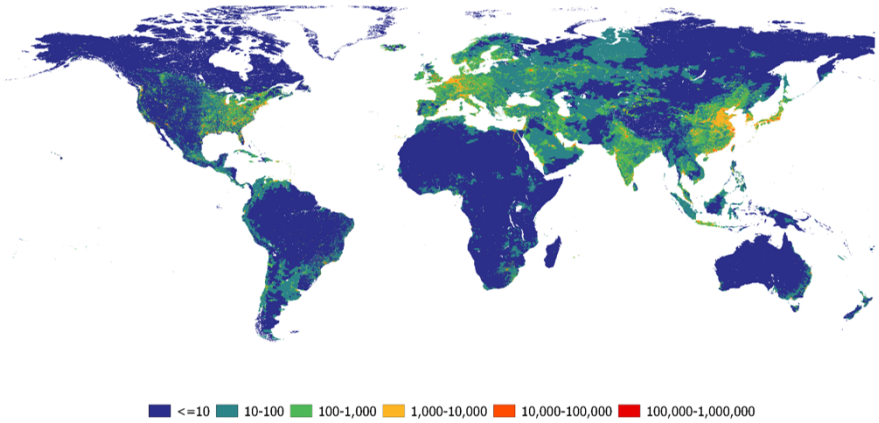
rid of an issue encountered by Lloyd et. al. (2017) regarding whether to do the backtransformation before or after summarizing the ensemble of predictions [51]. Finally, to get some insight into the uncertainty in the final predictions a pseudo-standard deviation was calculated from the prediction ensemble. An alternative quantile regression model was also estimated to generate percentile bounds.

The predictions in their current format performed reasonably well, as can be seen in Figure 4. Nevertheless there is obviously still some variance such that the predictions for certain countries or regions are poor. As such a dasymmetric allocation approach was employed whereby the 30-arc-second prediction layer was used as a weighting scheme to conduct within region allocation, rather than as the final data product. This was achieved by proportionally increasing or decreasing the predicted values of the cells in a given region such that they allow for the accurate recovery of the original "known" electricity consumption values that were used in the training dataset. This has the added benefit of making the final dataset consistent with the UN national database that formed the basis for the entire analysis. The final dataset can be seen in Figure 3.

Code availability

All coding was completed in R version 3.4.0 [53]. Code is contained in a single file that is available from the online data repository. The code is internally documented throughout. All input data used is publicly available. The dataset was created using a computer running Microsoft Windows 7, 64 bit operating system. Some figures were produced using QGIS version 2.16.1 [54].

Figure 3: Global Gridded Electricity dataset in kwh per sq. km (2015)



Data Records

The GGE data are freely available to download from the GGE data repository. A separate GeoTIFF file is provided for each year. Additional input raster data that was used in the analysis, such as the economic raster data, is available on request.

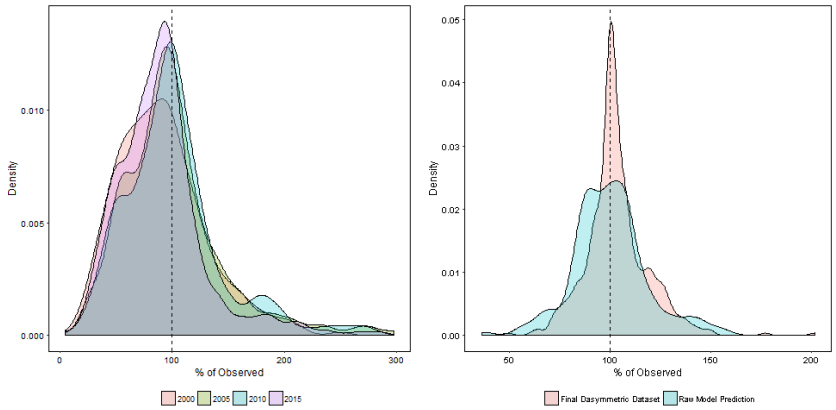
Technical Validation

The most common approach to validation in this context is to withhold known data from the estimation procedure and then examine the ability of the final dataset to predict out-of-sample. The use of a Random Forest estimation approach means this kind of procedure was essentially built into the creation of the dataset. As such the Out-Of-Bag (OOB) error of the estimated Random Forest regression already provides a robust indication of the quality of the dataset. The OOB error itself is difficult to directly interpret in terms of the expected error for the final predictions, both because of the change of spatial scale and the dasymmetric weighting. Nevertheless, the results for 2015 appear to indicate that the preferred Random Forest regression specification performs well, explaining 96.5% of the variance in the training dataset. The ability of the model to predict the original training dataset values can be seen in Figure 4a.

Because the predictions of the Random Forest model were used for dasymmetric weighting at a finer spatial scale than the initial training set there is likely still value to conducting additional out-of-sample validation checks. At the very least these can give a more transparent insight into the predictive quality of the final dataset. To do this one source of particularly high spatial resolution data was used that was not incorporated into the model estimation

process: the local authority level electricity consumption data for the UK (approximately 400 subnational units). The results of the validation against these data is shown in Figure 4b and indicates the dataset performs well, particularly after the dasymmetric adjustments.

Figure 4: Dataset Validation



(a) Raw Model Prediction Error (b) Out-of-sample Validation (2015)

Acknowledgements

I would like to acknowledge the valuable input provided by Prof. Duncan Callaway, as well as comments offered throughout by my UC Berkeley colleagues at the Energy Institute and the Energy & Resources Group.

Competing financial interests

The author declares no competing financial interests.

References

[1] United States Energy Information Administration. Electric power sales, revenue, and energy efficiency Form EIA-861. (2016).

[2] British Petroleum. BP World Energy Outlook 2017. (2017).

[3] United Nations Statistics Division. Energy Statistics Database. (2018).

[4] International Energy Agency. World Energy Balances. (2017).

[5] United States Energy Information Administration. State Energy Data System. (2016).

[6] Fridley, D. *et al.* China Energy Data Yearbook. Lawrence Berkeley National Laboratory (2016).

[7] Brazil Energy Research Company (EPE). Regional Electricity Consumption. (2016).

[8] United Kingdom Department for Business, Energy & Industrial Strategy. Sub-national electricity consumption data. (2017).

[9] General Commissariat for Sustainable Development Observation and Statistics Service (EIDER). Final Energy Consumption Statistics. (2013).

[10] France Electricity Transmission Network (RTE). Final annual regional electricity consumption. (2017).

[11] Australian Office of the Chief Economist. Australian Energy Statistics. (2017).

[12] Statistics Canada. Canada Energy Supply and Disposition Statistics by Province. (2017).

[13] Statistics Austria. Austria Energy Balances at Laander. (2017).

[14] German Country Working Group on Energy Balances. German Energy Balances at Laander. (2017).

[15] Statistics Sweden. Sweden Energy Balances by Region. (2017).

[16] Mexican Secretariat of Energy. Mexico Energy Information System, Electricity Sales by State. (2017).

[17] Italian Statistical Territorial Infrastructure Atlas (ATSI). Italy Province Electricity Consumption. (2017).

[18] Statistics Norway. Norway County Electricity Consumption. (2017).

[19] National Statistical Institute of Chile. Chile Regional Electricity Consumption. (2017).

[20] Netherland Central Bureau of Statistics. Netherlands Electricity Delivered by Province. (2017).

[21] Korea Energy Statistical Information System. South Korea Electricity Sales by Region. (2017).

[22] Statistics South Africa. South Africa Electricity Distributed by Province. (2017).

[23] California Energy Commission. Electricity Consumption by County. (2017).

[24] Poland Local Data Bank. Poland Regional Electricity Consumption. (2017).

[25] Argentina Ministry of Energy and Mines. Argentina Electricity Consumption by Province. (2017).

[26] India Ministry of Power. India State Power Supply Position. (2017).

[27] Spain Ministry of Energy, Tourism and Digital Agenda. Annual Electricity Statistics. (2017).

[28] World Bank. World Development Indicators. (2017).

[29] United Nations. Demographic Yearbook Statistics. (2017).

[30] United Nations. National Accounts Main Aggregates Database. (2017).

[31] Eurostat. Regional Statistics Database. (2017).

[32] United States Bureau of Economic Analysis. Regional Economic Accounts. (2017).

[33] Organisation for Economic Cooperation and Development. Regional Economic Database. (2017).

[34] Organisation for Economic Cooperation and Development. Market Exchange Rates. (2017).

[35] United States National Aeronautics and Space Administration. Gridded Population of the World Version 4. (2016).

[36] United States National Oceanographic and Atmospheric Administration. DMSP Radiance Calibrated Nightlights Dataset. (2017).

[37] United States National Oceanographic and Atmospheric Administration. VIIRS Nightlights Dataset. (2017).

[38] European Space Agency. Climate Change Initiative Land Cover Dataset. (2017).

[39] United States Geological Survey. Global 30 Arc-Second Elevation (GTOPO30) Dataset. (2017).

[40] Berkeley Earth. Gridded Monthly Land and Ocean Temperature Dataset. (2017).

[41] OpenStreetMap contributors. OpenStreetMap Substations. (2017).

[42] Hijmans, R., Kapoor, J., Wieczorek, J., Garcia, N. & Maunahan, A. Rala, A. Mandel, A. Global Administrative Areas Dataset v2.8. (2017).

[43] Geoportal of the European Commission, Eurostat. Nomenclature of Territorial Units for Statistics (NUTS) 2013 - Statistical Units. (2015).

[44] Office for National Statistics. United Kingdom Local Authority Boundaries. (2011).

[45] Statistics Canada. 2011 Canadian Census Division Boundaries. (2011).

[46] Australian Bureau of Statistics. Australian Standard Geographical Classification Statistical Area 4 Digital Boundaries. (2011).

[47] French Etalab. French administrative division at regional level. (2015).

[48] Tsendbazar, N., de Bruin, S., Fritz, S., Herold, M. & Schadt, E. Spatial Accuracy Assessment and Integration of Global Land Cover Datasets. *Remote Sensing* **7**, 15804-15821 (2015).

[49] Stevens, F.R., Gaughan, A.E., Linard, C. & Tatem, A.J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS ONE* (2015).

[50] Nordhaus, William D. Geography and macroeconomics: New data and new findings *PNAS* **30**, 10, 3510-3517 (2006).

[51] Lloyd, C. T. *et al.* High resolution global gridded data for use in population studies. *Scientific Data* **4:170001** (2017).

[52] United Nations International Recommendations for Energy Statistics *Statistical Papers* **M**, 93 (2016).

[53] R Foundation for Statistical Computing R: A Language and Environment for Statistical Computing v3.4.0 (2017).

[54] Open Source Geospatial Foundation QGIS Geographic Information System v2.16.1 (2017).