rgugik: Search and Retrieve Spatial Data from the Polish Head Office of Geodesy and Cartography in R

Krzysztof Dyba and Jakub Nowosad

1 Institute of Geoecology and Geoinformation, Adam Mickiewicz University in Poznań

Introduction

Currently, the open data market size is estimated at about 185 billion Euros in the European Union and is expected to grow in the coming years (Huyer & Knippenberg, 2020). It includes spatial data that can result in cost savings and create new, innovative products and services for the benefit of the society, environment, and economy. The public sector is one of the primary providers of vast amounts of valuable spatial data resources.

The Head Office of Geodesy and Cartography (Główny Urząd Geodezji i Kartografii, GUGiK) is the central government agency responsible for collecting spatial data in Poland. Their resources include various datasets, such as orthophotomaps, register of borders, 3D models of buildings, digital elevation models, and point clouds. Until July 31, 2020, spatial data acquisition was time-consuming, required filling-in forms, and paying a fee. However, the recent amendment of the Geodetic and Cartographic Law in Poland in mid-2020 made all of the current and future spatial datasets publicly available.

Poland’s spatial data is released on a dedicated website, Geoportal, which allows it to be browsed and downloaded. The Geoportal is one of the most popular government websites in the country, currently ranked 3rd with 5.5 million unique visits in 2020. Although the data is related to Poland’s area only, it is a popular resource for many other countries (e.g., Germany with 52,000, Great Britain with 40,000, and United States with 15,000 unique visits this year). In the first month after the change of law, 69 TB of data was downloaded, and by the end of October, this value grew to over 240 TB.

Statement of need

While the Geoportal gives access to some of the GUGiK data resources, it has several practical disadvantages. Datasets can only be downloaded individually and manually, limiting their practical use for studies over large areas or for many points in time. It is also problematic for the reproducible research process. Additionally, some GUGiK data is located on other associated websites or in the form of dedicated services, which makes finding and downloading certain datasets more difficult.

Therefore, there is a need to make all GUGiK data sources available in one place and to automate the data downloading and preprocessing.

*Corresponding author
Summary

rgugik is an R package (R Core Team, 2020) that attempts to tackle all of the shortcomings listed above by providing consistent tools for searching and retrieving of spatial data from GUGIK. It integrates multiple data sources (i.e., HTML websites, FTP servers, API services), allows for data search and download, and gives the ability to create reproducible scripts. In total, it provided access to ten datasets of various formats, including numeric, vector, and raster [Table 1].

The package contains 15 functions, including three functions dedicated exclusively to digital terrain models. The functions can be divided into three main groups indicated by their suffixes:

- `_request()` to obtain metadata and links to the data based on the provided location. Allows users to understand what sort of data is available, select only some of the metadata, and use the result as an input to the `_download()` functions.
- `_download()` to download the data files to a hard drive and unzip it.
- `_get()` to retrieve selected spatial datasets as R object of classes, such as `sf`/`data.frame`.

It is also possible to geocode addresses or objects located in Poland with rgugik. Additionally, the package includes objects containing names of the administrative units and their IDs to facilitate data retrieval.

| Polish                                      | English                                      |
|---------------------------------------------|----------------------------------------------|
| Ortofotomapa                                 | Orthophotomap                                |
| Baza Danych Obiektów Ogólnogeograficznych    | General Geographic Database                  |
| Baza Danych Obiektów Topograficznych        | Topographic Database                         |
| Ewidencja Miejscowości, Ulic i Adresów      | Register of Towns, Streets and Addresses     |
| Państwowy Rejestr Nazw Geograficznych       | State Register of Geographical Names         |
| Państwowy Rejestr Granic                    | State Register of Borders                    |
| Lokalizacja działek katastralnych           | Location of cadastral parcels                |
| Modele 3D budynków                          | 3D models of buildings                       |
| Cyfrowe modele wysokościowe                 | Digital elevation models                     |
| Chmury punktów                              | Point clouds                                 |

rgugik uses jsonlite (Ooms, 2014) for parsing JSON to R objects and sf (Pebesma, 2018) for processing spatial data in a user-friendly way. The package is released under the MIT open-source license and can be directly installed from CRAN, or from GitHub using the remotes (Hester et al., 2020) package. This package’s source code is thoroughly tested, with about 87% lines of the code executed using automated tests. The package also has an associated website at https://kadyb.github.io/rgugik, which contains installation instructions and three articles presenting different use cases of downloading and processing of orthophotomaps, digital elevation models, and topographic databases.

Three other products aimed at downloading data from GUGIK were recently developed — QGIS plugins by the EnviroSolutions and by GIS Support companies, and a commercial, general data acquisition purpose product made by Globema. However, all of them have certain limitations and offer a smaller subset of the GUGIK datasets compared to rgugik. They use graphical user interfaces, which, while they can be user-friendly, they also make it more laborious to download many files and use the data in reproducible workflows. Moreover, the QGIS plugins are in Polish, restricting potential users to Polish speakers only.
Example usage

```r
library(rgugik)
library(sf)
library(raster)
polygon = read_sf("search_area.gpkg")

The first example shows a search for available digital elevation models based on the input polygon and downloading a selected digital terrain model [Figure 1]. The DEM_request() function uses a dedicated API. As a result, a data.frame with available data and their metadata is returned. The output data.frame can be easily filtered and used to download the desired data via FTP.

# downloading a metadata of available digital elevation models
req_df = DEM_request(polygon)

# printing metadata
t(req_df)
#> sheetID     "M-33-58-A-d-1-1"
#> year        "2011"
#> format      "ARC/INFO ASCII GRID"
#> resolution  "1"
#> avgElevErr  "0.15"
#> CRS          "PL-1992"
#> VRS          "PL-KRON86-NH"
#> filename    "3982_154755_M-33-58-A-d-1-1"
#> product     "DTM"

# downloading DTM
tile_download(req_df)

# plotting the results
DTM = raster("3982_154755_M-33-58-A-d-1-1.asc")
plot(DTM)
```

[Figure 1: Fort Srebrna Góra (a mountain fortress hidden in the forest).]

Dyba et al., (2021). rgugik: Search and Retrieve Spatial Data from the Polish Head Office of Geodesy and Cartography in R. Journal of Open Source Software, 6(59), 2948. https://doi.org/10.21105/joss.02948
The second example presents how to get geometries of the highest-level administrative division of Poland (voivodeships) [Figure 2]. The names of administrative units can be obtained from the `voivodeship_names` object stored in the package. As a result, an object of class `sf/data.frame` is returned.

```r
defining names of voivodeships
voivodeships = voivodeship_names$NAME_PL

defining the data as sf object
voivodeships_geom = borders_get(voivodeships)

# plotting the results
plot(st_geometry(voivodeships_geom))
```

![Figure 2: Voivodeships in Poland.](image)

The third example shows the process of converting place names to spatial coordinates (geocoding) [Table 2]. As a result, an object of class `sf/data.frame` is returned.

```r
cdefining geocoding of a provided name or address
gcodePL_get(address = "Dąbrowa")
```

| city     | teryt | voivodeship | county | commune | geometry_wkt               |
|----------|-------|-------------|--------|---------|---------------------------|
| Dąbrowa  | 021302| dolnośląskie| miłicki | Krośnice| c(387236.148, 403862.917) |
| Dąbrowa  | 061002| lubelskie   | Łęczyński| Ludwin  | c(770342.296, 393839.750) |
| Dąbrowa  | 101709| łódzkie    | wieluński| Wieluń  | c(467414.612, 374431.514) |
| Dąbrowa  | 160402| opolskie   | kluczborski| Kluczbork| c(445709.237, 351749.657) |

**Table 2:** Geocoding results for the city of Dąbrowa.

**References**

Hester, J., Csárdi, G., Wickham, H., Chang, W., Morgan, M., & Tenenbaum, D. (2020). *Remotes: R package installation from remote repositories, including 'GitHub'*.[https://CRAN.R-project.org/package=remotes](https://CRAN.R-project.org/package=remotes)

Huyer, E., & Knippenberg, L. van. (2020). *The economic impact of open data. Opportunities for value creation in europe* (pp. 1–138). European Data Portal. [https://doi.org/10.2830/63132](https://doi.org/10.2830/63132)

Ooms, J. (2014). *The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects*. *arXiv:1403.2805 [stat.CO]*.[https://arxiv.org/abs/1403.2805](https://arxiv.org/abs/1403.2805)
Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal, 10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Dyba et al., (2021). rgeik: Search and Retrieve Spatial Data from the Polish Head Office of Geodesy and Cartography in R. *Journal of Open Source Software, 6*(59), 2948. https://doi.org/10.21105/joss.02948