censusTools - An R package for integrating and harmonising census statistics

Steffen Ehrmann, Ralf Seppelt, Navin Ramakutty, Carsten Meyer

2019-05-07

1 Abstract

Humans gather information about a wide variety of phenomena in the form of census statistics. Those are typically based on recuring surveys that relate some quantity of socioeconomic, land-use, or environmental variables to a particular territorial area. We are interested in those data out of an economic interest, to assess environmental impacts or to support political decisions.

Not only in the context of Sustainable Development Goals such data need to be harmonised across broad temporal and spatial extents up to the global scale. Mostly the issue is not to gather any data at all, but to make data across various sources compatible with one another. This obstacle has so far not been tackled to a degree that would be satisfying with respect to transparency and reproducability. However, to enable re-usability and thus future-proof the vast and often laboriously collected (and perhaps simplified) data, a transparent and reproducible workflow should be guaranteed.

Here we introduce a software package (written in `R`) that allows the user to build up integrated and harmonised databases of a wide variety of census statistics, which are typically related to some sort of territorial unit. Databases set up with this tool are both internally consistent and can be combined with other databases assembled with this tool. We exemplify the usage by showcasing some of the important steps we carry out for an ongoing project on spatio-temporal dynamics of land-use. The package can be used for any kind of census statistics that shall be connected to territorial units.

2 Introduction

Gathering census statistics to aggregate them is a laborious and cumbersome task. The aim of a census is to tally the overall population of some group of entities of interest. A census is

thus often perceived as the opposite of a sampling, which merely tries to estimate the overall population based on a representative subset.

To enable an accurate assessment of any current situation, systematic census is crucial. A census is carried out to assess, for instance, the livelihood of people, to monitor agricultural production for food security, to inventory all sorts of commodities or products out of economic reasons or to assess the environmental impact of human activities with the help of checklists of plant and animal species.

While we already learn quite a bit from national censuses, there is a large incentive to aggregate several of such efforts across larger spatial or temporal scales to infer on phenomena that can only be perceived at those larger scales. For example, to successfully achieve several of the sustainable development goals, we need to recognize and acknowledge global phenomena that have unfolded during the course of decades (*go into more details/examples here?!*). Moreover, we may want to combine census statistics from various sources that may cover the same spatio-temporal extent or that may be related to one another. For instance, combining several subsets that make up a larger area and considering the different sources of errors for small and large scale data lets us infer on uncertainties that may be introduced by all of the data sources.

This entails however that data from multiple, heterogeneous sources often have to be harmonised and integrated across several (spatio-temporal) scales. On top of the considerations and assumptions that go into individual nation level censuses, harmonizing them requires to consider yet another level of heterogeneity and assumptions.

Shall we also include a more theoretical section on how these census data are typically "meso-scale" data that fill an important gap between in-situ local scale data and larger scale data, cf Petr Keils topic?

The United Nations defines the essential features of population and housing censuses as "individual enumeration, universality within a defined territory, simultaneity and defined periodicity", and recommends that population censuses be taken at least every 10 years. *Agricultural censuses are defined by the FAO as follows...* Ecological assessments of plant or animal communities are typically not recorded in a census but merely sampled. (*more details on that*) (König et al., 2019)

Typically census data are counted per a specific spatial unit. In other words, a census can have various properties, one of which is the location at which the recorded variables are true. For monitoring or modelling system responses with a spatial dimension it is thus necessary to make census data and spatial information align perfectly.

A *sample* of subsets of a population should be treated different than the *census* of the

(complete) population out of statistical considerations (*should we mention details on that?*). The *degree of detail/precision/...* that comes with the difference between "sampling" and "taking a census" is an important property based on which we may want to select certain statistics under certain conditions.

CM: Also, refer to some specific data integration processes that are currently underway but slowed down due to lack of such tools. E.g. FAO and IFPRI spend personnel funds each year on people like Ulrike who currently do this by hand. All these institutes are underfunded and understaffed and should thereby benefit from such a tool.

A couple of words on how census data are typically sampled and relationships between various census efforts: http://www.fao.org/world-census-agriculture/methodology/en/

A couple of words on why census data are recorded: http://www.fao.org/fileadmin/templates/ess/documents/world_census_of_agriculture/chapter02_r7.pdf; needs more sources of information

Outline issues

Territory outlines may change with time. *More on how this becomes problematic*

Working with census statistics entails not only finding orientation in a vast number of different formats the data are provided in, but also associating those tabular information to the spatial information for which it often is recorded.

With the advancement of GIS know-how, many census stats are provided together with the spatial information based on which they have been derived. When census data are aggregated for a particular territorial unit, even if they are originally recorded in the form of point records, the spatial extent of that unit needs to be available. However, often those spatial information are not taken from a standardized set of spatial geometries, such as the GADM data-set [REF], but are derived or manually created from (outdated) local maps. Various sources of error make the spatial data deviate from a possible standard. Those could be:

1. Choice of the wrong or no particular coordinate reference system,
2. systematic errors when manually copying paper maps with digitizing tools,
3. deviations that emerge when vectorising and processing raster maps (either raster-outline as boundary of the spatial unit, or a deviating boundary when the raster-boundary is smoothed with some unknown smoothing function) or
4. disagreement on territorial boundaries.

These eventually result in a very general source of inconsistency, when including or excluding certain (point) sources or other parts of information due to deviating spatial extents, or when census data are innately related to the area or topological information of territorial units.

3 The challenge

This chapter shall clearly state/summarise the open aspects and what of that we want to solve with `censusTools`.

A wide variety of census data are available. *Outline two/three examples and their specifities*; *Outline what they have in common and which differences between them need to be considered.*

Census data and geometries must be associated to one another, possibly on a global scale and so that they are compatible across distinct efforts.

This entails:

1. Administrative/territorial units must be matched *across different languages*.
2. Likewise, particular key variables (species, commodities, basically any categories) might have different values in different languages or censuses and they must also be matched.
3. To make variables comparable across different efforts, there must be a standard for naming variables (*SE: I would thus suggest in this paper that naming shall follow the Darwin Core, this would take the burden of defining such a naming standard off of our shoulders and would also assert that future compatibility is managed by the Darwin Core rules*).
4. Moreover, we believe that such efforts must be *fully transparent and reproducible*.

A generalised framework for managing census data that would be related to spatial information, irrespective of the covered variables, was devised and shall be presented in this paper.

4 Description of `censusTools`

This section describes briefly what `censusTools` does. In the following subsections technical details are outlined.

Typically some sort of polygons outlining territorial boundaries (here called *geometries*) and some sort of tables outlining the quantities of particular variables (here called *census tables*) shall be handled with `censusTools`.

Set project variables

A typical workflow (Fig. 1) would initially consist of setting up project variables. Those variables are the categorical variables for which the census covers quantities, such as socio-economic groups of people, animal species or agricultural commodities. For each of the variables an *index* and a *translation table* are required. An index contains at least the

values of the target variable (e.g. 'maize', 'wheat', 'rice' for the variable 'commodities'), unique IDs for each value and arbitrary data that describe the values further (e.g. details on collecting/sampling the data, the scientific name or a description). A translation table contains at least the values that should be translated, the values of the translated target variable (same terms as in the index) and a statement as to how the translations have been created.

Register input tables

In a second step first of all some relevant data-series (i.e. specific series of data which are delineated from other series due to their data format) and then the source-files of geometries and census tables are registered in respective index tables each. While project variables are set up only once, this second step will be carried out as often as new data are added to the database. The underlying functions give instructions, monitor and document progress and assert that all files are available in the correct directory (*see section "The directory structure"*). Additionally, they create unique IDs for data-series, geometries and census tables. The resulting index tables are eventually an inventory of which geometry and census files are part of the data-base, where and when they have been stored and which ID they have.

Normalise data

In a third step, both geometries and census tables are harmonised and integrated (i.e. *normalised*) into a standardised data-base structure. Census tables typically contain information that are aggregated for certain territorial units and thus it makes sense, from an efficiency point of view, that the territorial units are first normalised, so that the census tables can simply be joined to them. Both, geometries and census tables are aggregated per nation. If data sources (geometries and census tables) contain information on several nations, they are dis-aggregated and the respective junks are appended to the nation specific files.

To normalise geometries a large computational effort is required, but much automisation is possible (*see section "Normalising spatial data"*). The geospatial data-base is built up from a basis that is provided by the user. It does not urge the user to employ for instance the GADM data-set [REF], even though this is the recommended starting point recently. This basic geo-spatial data-set provides the hierarchical administrative organisation and the names of territorial units by which the following input shall be organised. GADM comes with alternative names for numerous territorial units, it acts thus in a way like a gazetteer. However, since it can't be guaranteed that all data providers use only names that have been covered by the included index, `censusTools` allows that new mappings between so far not recorded terms and target terms are made or that additional gazetteers are read in (*see section "Managing translation tables"*).

Normalising census tables can only be automated to a certain degree. Most data providers

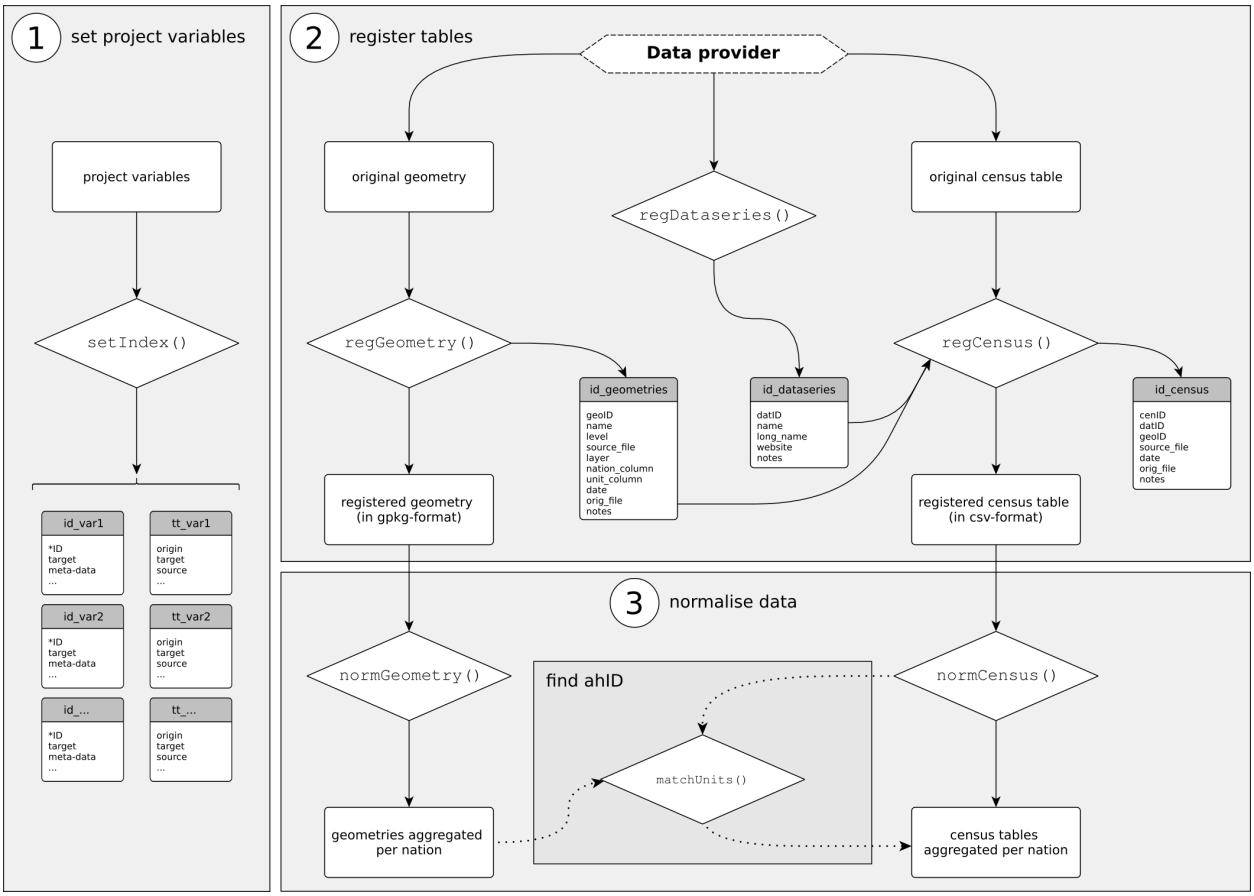

Figure 1: Fig. 1: Overview of the general workflow employed by `censusTools`.

don't follow a standard format when it comes to the provided data. All sort of messy tables need to be transformed and, to our knowledge, this is nothing a computer could recognise autonomously to date. `censusTools` relies on the *rectr* R-package [REF], which takes a list of instructions that describe where which data are located in a particular spreadsheet, and reorganises the information into a rectangular table that can be processed with a *dplyr*-based pipeline [REF]. Eventually, the values of categorical variables in the census tables are mapped to their indices to reduce the size of the overall database. Such variables may prominently be the territorial units (*see section "Territory ID"*), agricultural commodities or basically any variable that is defined by the user (*see section "Normalising census data"*).

4.1 The directory structure

`censusTools` relies on a rather rigid directory structure and thus creates this structure itself, when the function `setPath()` is called for the first time. Within the primary project directory the two directories `./cT_geometries/` and `./cT_census/` are created. Within both of those

directories, the following directories are created:

- `incoming/`: a tentative location for new data,
- `meta/` a place for related data,
- `original_datasets/` an archive of the original, unmodified files,
- `stage1/` and `stage2/`, which store different qualities of modified data. In `stage1` the registered data are stored in a standardised format, geometries are stored as GeoPackage files [REF, https://www.opengeospatial.org/standards/geopackage] and census tables are stored as comma-separated value files with a UTF-8 encoding. In `stage2` the nation specific database files are stored.

4.2 Normalising spatial data

Despite this, `censusTools` matches territorial units not based on the administrative names, as those are less reliable than the spatial extent. It matches administrative units by their spatial overlap, with a so-called *spatial join* [REF?]. (*go into a couple moer details*)

4.3 Normalising census data

Census stats are stored in tables, where the data source(s), territorial and temporal information and the recorded variables are present in the following columns:

- `cenID` (census ID) and `geoID` (geometry ID) hold IDs that relate the respective row to a particular census and geometry data provider. Often similar or the same data are provided by several distinct data providers or by data aggregators. Those two IDs thus allow quality assessments of the data and ensure that the data provenance is properly documented.
- `ahID` (administrative hierarchy ID) represents the ID that is also recorded in the respective spatial data and which relates census stats to territorial units.
- `time-step` would be a column that denotes the time for which the census has been recorded. It could take any form, for example a combination of year and day-of-the-year (doy), merely the year or the month within a year, etc. This may depend on the temporal resolution of the source data-sets or of the questions that shall be addressed with the acquired data.
- `variable(s)` can be any number of columns, each of which would contain either keys (typically categorical variables such as agricultural commodities or species) or values (typically some sort of quantity of the key, such as “production” and “yield” or “abundance”).

ID	cenID	geoID	ahID	time-step	variable 1 (key)e.g. commodity	variable 2 (value)e.g. production
1	1	1	070017008	2016	maize	15000
2	1	1	070017008	2016	wheat	12000
3	1	1	070017008	2017	...	
...						

Both, `timestep` and `variables` are recorded indifferent of their units, which are recorded separately in the meta data. (*SE: I still need to assert that the units are actually recorded; my idea would be that data should be recorded at the finest level at which they are reported so that aggregation only happens later, when the model is computed; which may not even be needed for some more advanced models in the future.*).

4.4 Managing translation tables

A couple of words about what "key variables" are and in general about the data types that might be recorded in census data

The key variables must all hold values that are part of the smallest common set. Terms are gathered in a list, where they are related to a set of generalized terms. For instance, commodities may either be labelled by their species, by the fao-ID or by a nation/language specific term. Hence, both, semantic synonyms and terms in other languages are translated to a common set of terms.

origin	target	notes
toTranslate	translated	translateTerms() on 2019-02-28
anotherOne	itsTranslation	translateTerms() on 2019-02-28
...	...	
	term1	target
	term2	target
	term3	target

This table is split into two parts. The upper part contains all original terms, as they appear in different census data-sets (`origin`), their unified translations (`target`) and `notes` on when and how (here, by the function `translateTerms()`) the translation has been carried out. The lower part contains all target terms to which original terms should be translated. When encountering new terms, those target terms are used for fuzzy matching. Here, a list of the

three most highly matched terms based on the Levenshtein distance [REF] is returned, from which the user has to chose. Moreover, before the translated data are finally entered into the database, they are checked against this list so that no new translations violate the ontological consistency (*SE: I still have to make sure that this feature is actually implemented*).

4.5 Territory ID

Administrative territorial units are often subsets of larger units and eventually of nations. To denote those units, we use a particular ID that is unique per unit, irrespective of the administrative level and which captures the hierarchical administrative information.

At each administrative level, territorial units are enumerated along their alphabetic sequence with a three-digit ID, starting from 1 through their count. The "children" within each higher level unit ("parent") likewise restart at 1 and are enumerated along their alphabetic sequence. The nestedness is represented by a sequence of those three digit IDs, where parent units are placed before their children. For example, `070'017'008` is the community of Tammelin (8th community) in Tartumaa (17th county), Estonia (70th nation). This code is extended with more sets of unit-IDs, when working on the fourth/fifth/... administrative level.

(*I have started a chapter in the discussion, where I outline how temporal and territorial shifts are dealt with in* `censusTools`)

I realised that the package also needs an option to join census and spatial data by an ID that may already be in the data, such as FIPS for the US census data.

5 Discussion

what do other packages that `censusTools` *doesn't do? -> CM: more precisely, what cool new things will now be facilitated with these new tools?), limitations (e.g. against other tools or in general, but focusing on the specific challenges the toolls are meant to overcome), etc*

what can `censusTools` *generally not be used for?* -> It is not a tool that automates extraction of census data from messy tables/spreadsheets. For that task, check out the R-package `rectr`. However, `censusTools` makes use of this package to organize a particular subset of messy data, agricultural census data that typically provide a narrow, predefined set of variables and meta information.

Discuss, based on parts that have been mentioned in the introduction, how `censusTools` *can actually deal with the outlined challenges and the requirements of different kinds of census*

data.

5.1 Dealing with temporal changes

As the territory ID does not depend on any topological information of the territorial units, it allows to assign individual IDs to units that are valid only for a certain period of time.

(*SE: I still have to make sure that this is properly reflected by what the code does*)

5.2 Dealing with disputed areas

As the territory ID does not depend on any topological information of the territorial units, it allows to assign individual IDs to units that may be disputed territories.

(*SE: I still have to make sure that this is properly reflected by what the code does*)

6 Outlook

Perhaps talk about how to implement this for use in citizen science and for building large and exhaustive open databases.

mention a couple of words about LUCKINet?! -> CM: Could be done in a paragraph on different initiatives that could make good use of this tool. Could also mention e.g. WorldPop for population census data, GIFT for regional checklists, etc.

How does `censusTools` *facilitate interoperability with other tools (that may not exist yet)?*

7 Session Info

this should probably be replaced by a proper list of packages that are used for `censusTools`

```
sessionInfo()
#> R version 3.6.0 (2019-04-26)
#> Platform: x86_64-pc-linux-gnu (64-bit)
#> Running under: Ubuntu 18.04.2 LTS
#>
#> Matrix products: default
```

```
#> BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
#> LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=de_DE.UTF-8        LC_COLLATE=en_US.UTF-8
#>  [5] LC_MONETARY=de_DE.UTF-8    LC_MESSAGES=en_US.UTF-8
#>  [7] LC_PAPER=de_DE.UTF-8       LC_NAME=C
#>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
#> [11] LC_MEASUREMENT=de_DE.UTF-8 LC_IDENTIFICATION=C
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] censusTools_0.1.0
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.1         knitr_1.22         magrittr_1.5
#>  [4] units_0.6-2        hms_0.4.2          tidyselect_0.2.5
#>  [7] R6_2.4.0           rlang_0.3.4        stringr_1.4.0
#> [10] dplyr_0.8.0.1      tools_3.6.0        grid_3.6.0
#> [13] checkmate_1.9.1    xfun_0.6           KernSmooth_2.23-15
#> [16] e1071_1.7-1        DBI_1.0.0          class_7.3-15
#> [19] htmltools_0.3.6    yaml_2.2.0         digest_0.6.18
#> [22] assertthat_0.2.1   tibble_2.1.1       sf_0.7-4
#> [25] crayon_1.3.4       bookdown_0.9       tidyr_0.8.3
#> [28] purrr_0.3.2        readr_1.3.1        glue_1.3.1
#> [31] evaluate_0.13      rmarkdown_1.12     stringi_1.4.3
#> [34] rectr_0.1.0        compiler_3.6.0     pillar_1.3.1
#> [37] backports_1.1.4    classInt_0.3-3     pkgconfig_2.0.2
```

8 Miscellaneous/Notes/Comments

8.1 Comments on editorial decisions

Steffen

1. After reading more on it, it seems that it would make sense to use *territorial unit*, which is at least in the European legislative speak the commonly used term: https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics. Also, "administrative unit" seems to be often used to describe a unit that deals with administrative stuff within an organisation, like a school, etc.
2. I think it is not needed to specify in detail in the title that we only deal with *census stats that are related to territorial units*, as this is typically the case for the concept of "census" (https://en.wikipedia.org/wiki/Census).
3. I am making a deliberate distinction between *attribute* and *property* (https://stackoverflow.com/a/21566583)

8.2 Known tools and pipelines (based on R)

this list is to gather information against what `censusTools` *could be compared.*

Implement mapping in shiny: https://shiny.rstudio.com/tutorial/written-tutorial/lesson5/

Combining Australian Census data with the Same Sex Marriage Postal Survey in R

Extract US Census 2010 data with data.table and dplyr

Creating beautiful demographic maps in R with the tidycensus and tmap packages

Compare US metropolitan area characteristics in R with tidycensus and tigris

A Guide to Working with US Census Data in R, R-packages

8.3 (Important) papers on the topic

Aalders & Aitkenhead (2006): *Modelling land use change is often constrained by imperfect and incomplete data sources. This paper explores three modelling methodologies and their ability to predict agricultural land use on the basis of information from the Scottish Agricultural Census.*

Forclaz (2019): *This article provides a history of the First World Agricultural Census of 1930, an ambitious international attempt to evaluate world agricultural resources through the compilation of global statistics on crops, livestock, and agricultural production.*

Monfreda et al. (2008): *Here we present land use data sets created by combining national, state, and county level census statistics with a recently updated global data set of croplands on a 5 min by 5 min (~10 km by 10 km) latitude-longitude grid.*

Imbach et al. (2015): *We present here an agricultural statistics database of the entire Amazonia region, with a harmonised description of crops and pastures in geospatial format, based on administrative boundary data at the municipality level. The spatial coverage includes countries within Amazonia and spans censuses and surveys from 1950 to 2012.*

Otto et al. (2015): *Subnational socio-economic datasets are required if we are to assess the impacts of global environmental changes and to improve adaptation responses. Institutional and community efforts should concentrate on standardization of data collection methodologies, free public access, and geo-referencing.*

Ricciardi et al. (2018b), Ricciardi et al. (2018a): *We examine variations in crop production by farm size using a newly-compiled global sample of subnational level microdata and agricultural censuses covering more countries (n=55) and crop types (n=154) than assessed to date.*

Waha et al. (2016): *Surveys for more than 9,500 households were conducted in the growing seasons 2002/2003 or 2003/2004 in eleven African countries: Burkina Faso, Cameroon, Ghana, Niger and Senegal in western Africa; Egypt in northern Africa; Ethiopia and Kenya in eastern Africa; South Africa, Zambia and Zimbabwe in southern Africa.*

8.4 links to look through

https://ajs.data-analysis.at/index.php/ajs/article/view/vol39,%20no4%20-%202

https://www.tandfonline.com/doi/abs/10.1080/136588198241590

https://www.aeaweb.org/articles?id=10.1257/aer.p20161061

https://www.journals.uchicago.edu/doi/abs/10.1086/693916

https://search.proquest.com/openview/eb6bdc49aa2678f832fcf014ca09ae21/1?cbl=105444&pq-origsite=gscholar

https://www.jstor.org/stable/1403545?seq=1#metadata_info_tab_contents

http://www.fao.org/economic/ess/countrystat

https://link.springer.com/chapter/10.1007/978-3-319-44421-5_3

https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-041715-033713

there is still more...

https://www.ipums.org/

https://datacatalog.worldbank.org/

https://en.wikipedia.org/wiki/Gazetteer

8.5 Packages

perhaps include some statistics about how many packages there are

explain briefly what the key packages are used for and how this relates to `censusTools`

tidycensus: https://walkerke.github.io/tidycensus/, https://juliasilge.com/blog/using-tidycensus/

censusapi: https://github.com/hrecht/censusapi, https://cran.r-project.org/web/packages/censusapi/vignettes/getting-started.html

tigris: An R Package to Access and Work with Geographic Data from the US Census Bureau

8.6 Sources of census statistics and associated geometries

Lots of information on processing census stats (in R) is based on the "American Community Survey" and "Decennial Data from the US Census".

List of national and international statistical services

Spatial Data in Geographic Information System Format on Agricultural Chemical Use, Land Use, and Cropping Practices in the United States

AGRICULTURE CENSUS IN INDIA

References

Aalders, I. & Aitkenhead, M.J. (2006). Agricultural census data and land use modelling. *Computers, environment and urban systems*. 30 (6). pp. 799–814.

Forclaz, A.R. (2019). Agriculture, american expertise, and the quest for global data: Leon estabrook and the first world agricultural census of 1930*. *Journal of Global History*.

Imbach, P., Manrow, M., Barona, E., Barretto, A., Hyman, G. & Ciais, P. (2015). Spatial and temporal contrasts in the distribution of crops and pastures across amazonia: A new agricultural land use data set from census data since 1950. *Global biogeochemical cycles*. 29 (6). pp. 898–916.

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J. & Kreft, H. (2019). Biodiversity data integration–the significance of data resolution and domain. *PLoS biology*. 17 (3). p. e3000183.

Monfreda, C., Ramankutty, N. & Foley, J.A. (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles*. 22 (1).

Otto, I.M., Biewald, A., Coumou, D., Feulner, G., Köhler, C., Nocke, T., Blok, A., Gröber, A., Selchow, S., Tyfield, D. & others (2015). Socio-economic data for global environmental change research. *Nature Climate Change*. 5 (6). p. 503.

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L. & Chookolingo, B. (2018a). An open-access dataset of crop production by farm size from agricultural censuses and surveys. *Data in brief*. 19. pp. 1970–1988.

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L. & Chookolingo, B. (2018b). How much of the world's food do smallholders produce? *Global food security*. 17. pp. 64–72.

Waha, K., Zipf, B., Kurukulasuriya, P. & Hassan, R.M. (2016). An agricultural survey for more than 9,500 african households. *Scientific data*. 3. p. 160020.