censusTools - Standardising census statistics

Steffen Ehrmann, Ralf Seppelt, Navin Ramakutty, Carsten Meyer

2019-04-11

```
library(censusTools)
```

1 Introduction

Definition of Census data: "A census is the procedure of systematically acquiring and recording information about the members of a given population." (wiki)

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.

-> *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.

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

With the advancement of GIS knowhow, 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a standardised set of spatial geometries, such as the GADM dataset [REF], but are derived or manually created from (outdated) local maps.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 choice of the wrong or no particular coordinate reference system, systematic errors when manually copying paper maps with digitising tools (*SE: I guess it makes sense to go a bit more into detail here, but I have not the greatest knowledge, as that was before my time...*) or 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).

This poses a very general source of inconsistency, when including or excluding certain (point) sources of information due to deviating spatial extent or when census data are inately related

to the area or topological information of the territorial unit.

A bit bla about what "key variables" are and in general about the data types that might be recorded in census data

SE: Which term should we use for the geometries? There is so far 'territorial unit', 'administrative unit', 'geometry' 'spatial information' and derivations thereof.

1.1 The challenge

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

This entails:

1. Any effort must be *fully transparent and reproducible*.
2. Administrative/territorial units must be matched *across different languages*.
3. Likewise, particular key variables (species, commodities, basically any categories) might have different values in different languages or cenuses and they must also be matched.
4. 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*).

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.

2 Known tools and pipelines (based on R)

Full list: https://crantastic.org/search?utf8=%E2%9C%93&q=census

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

An example workflow: https://medium.com/@miles.mcbain/combining-australian-census-data-with-the-sam

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

2.1 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

3 What this R-package provides

In a nutshell:

- Builds a unique ID that relates census and spatial information.
- Standardises both census and spatial information so that they are compatible across large swaths of data sources.
- Standardises ontologies of variable names and values.

3.1 Territory ID

This is an ID per territorial unit which is unique per unit captures at the same time the hierarchical information. Territories, especially administrative units, are often subsets of larger administrative units and eventually of nations. At each administrative level, territorial units are enumerated along their sequence (*starting from one through their count*) with a three-digit ID, restarting within each parent (*I guess this needs some refinement*). The nestedness is represented by a sequence of those three digit IDs, where larger units are placed before units nested therein, such as `070'017'008` (which is, when sorting administrative units alphabetically, the community of Tammelin in Tartu, Estonia). This code is extended with more sets of unit-IDs, when working on the fourth/fifth/... administrative level. The number of unit-IDs indicates thus the administrative level of the territorial unit.

3.2 Standardised Census data

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

ID	censusSourceID	geometrySourceID	territoryID	timestep	variable(s)
1					
2					
...					

`censusSourceID` and `geometrySourceID` are important for data provenance. Often similar or the same data are provided by different dataseries providers.

`territoryID` relates territorial units to census stats.

`timestep` would be a column that denotes the time for which the census has been recorded. It's temporal resolution should be adapted accordingly.

`variable(s)` can be any number of columns, each of which would contain either keys (typically categorical variables such as "commodities" or "species") or values (typically some sort of quantity of the key, such as "production" and "yield" or "abundance").

3.3 Standardised spatial data

3.4 Standardised ontologies

The key variables must all hold values that are part of the smallest common set. *note: there is a difference between 'variable names' and 'values of key variables'. Yet, both need to be translated/unified. In the former case this would be following the Darwin Core standard, in the latter case this would be case-specific, but ideally still following some (external) standard. for LUCKINet this would be the FAO naming-standard.* Terms are gathered in a list, where they are related to a set of generalised 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 (*SE: I have a more elgant word for that on the tip of my tounge, but can't come up with it right now*).

origin	target	notes
toTranslate	translated	translateTerms() on 2019-02-28

origin	target	notes
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 datasets (`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. 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 implement this feature*).

4 Discussion

what does `censusTools` *that other packages don't do?*

what do other packages that `censusTools` *doesn't do?*

what can `censusTools` *generally not be used for?*

5 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?!

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