**Software paper for submission to the Journal of Open Research Software**

To complete this template, please replace the blue text with your own. The paper has three main sections: (1) Overview; (2) Availability; (3) Reuse potential.

Please submit the completed paper to: editor.jors@ubiquitypress.com

**(1) Overview**

Title

Regions – an R package to validate, convert, aggregate and disaggregate sub-national statistics.

The title of the software paper should focus on the software, e.g. “Text mining software from the X project”. If the software is closely linked to a specific research paper, then “Software from Paper Title” is appropriate. The title should be factual, relating to the functionality of the software and the area it relates to rather than making claims about the software, e.g. “Easy-to-use”.

Paper Authors

1. Antal, Dániel; *(Lead/corresponding author first)*

2. Popovic, Milos;

Paper Author Roles and Affiliations

1. First author role and affiliation

2. Second author role and affiliation *etc.*

Abstract

A short (ca. 100 word) summary of the software being described: what problem the software addresses, how it was implemented and architected, where it is stored, and its reuse potential.

Keywords

keyword 1; keyword 2; *etc.*

Keywords should make it easy to identify who and what the software will be useful for.

Introduction

An overview of the software, how it was produced, and the research for which it has been used, including references to relevant research articles. Our software validates, and if necessary, corrects or converts regional statistical area codes, name labels, and helps simple imputation when regionally aggregated or disaggregated data is missing. Our software was developed as a side-project of a scientific research project when the conversion of Eurostat’s regional data from the NUTS2013 to the NUTS2016 definition of European sub-national statistical regions made broke our replicable research guideline.

The most similar software is the well-known and widely used countrycode R package in R <let’s find a Python one, too>. The country-code package provides a translation between various country-level geographical codes and names.

We were aiming to create a similar solution on sub-national level, which is a far more complicated process. While national boundaries are relatively stable (for example, in the last 40 years, only the division of the Soviet Union in 1991, the division of Czechoslovakia in 1993 and the multi-step divison of the former Yugoslavia) craeted hiearchically well defined, new boundaries in Europe. In these cases, for example, a longitudional dataset is easy to create, for example, by adding together, or creating a weighted average of statistics for Czechia and Slovakia. Countries however have a complete freedom to change their internal administrative, statistical or electorial boundaries. In the European Union alone, several hundred changes take place every 3-5 years on regional, provincial, and similar levels, and the even more on municipal and metropolitan area level. This makes the creation of longitudional dataset from statistical aggregated data (for example county or regional population, Covid-related excess death) or disaggregated ( US, Australia or Germany ‘state’ level GDP, or provincial GDP for the Netherlands, etc.) very difficult.

Our software has the following functionality:

* Validate
* Recode / rename
* Impute and project

A short comparison with software which implements similar functionality should be included in this section.

**Implementation and architecture**

How the software was implemented, with details of the architecture where relevant. Use of relevant diagrams is appropriate. Please also describe any variants and associated implementation differences.

**Quality control**

Detail the level of testing that has been carried out on the code (e.g. unit, functional, load etc.), and in which environments. If not already included in the software documentation, provide details of how a user could quickly understand if the software is working (e.g. providing examples of running the software with sample input and output data).

**(2) Availability**

***Operating system***

Please include minimum version compatibility.

***Programming language***

Please include minimum version compatibility.

***Additional system requirements***

E.g. memory, disk space, processor, input devices, output devices.

***Dependencies***

E.g. libraries, frameworks, incl. minimum version compatibility.

***List of contributors***

Please list anyone who helped to create the software (who may also not be an author of this paper), including their roles and affiliations.

***Software location:***

***Archive*** (e.g. institutional repository, general repository) (required – please see instructions on journal website for depositing archive copy of software in a suitable repository)

***Name:*** The name of the archive

***Persistent identifier:*** e.g. DOI, handle, PURL, etc.

***Licence:*** Open license under which the software is licensed

***Publisher:*** Name of the person who deposited the software

***Version published:***The version number of the software archived

***Date published:*** dd/mm/yy

**Code repository** (e.g. SourceForge, GitHub etc.) (required)

***Name:*** The name of the code repository

***Identifier:*** The identifier (or URI) used by the repository

***Licence:*** Open license under which the software is licensed

***Date published:*** dd/mm/yy

**Emulation environment** (if appropriate)

***Name:*** The name of the emulation environment

***Identifier:*** The identifier (or URI) used by the emulator

***Licence:*** Open license under which the software is licensed here

***Date published:*** dd/mm/yy

***Language***

Language of repository, software and supporting files

**(3) Reuse potential**

While our R package originally aimed to solve a problem that is relevant to about 35 countries using the European Union’s sub-national topology, we quickly realized that our methodology can be extended to other similar uses.

Currently, our software can create regional statistical aggregates. For example, in <xxx>, we created new statistical variables for library use in Europe <xxx>

Our first addition to the software was a conversion table from the international sub-regional standard ISO 3166-2 to Europe’s NUTS 0, 1,2,3 topologies to almost all European subdivisions. The ISO 3166-2 is not a statistical standard – it is most widely used in public administration – but in several parts of the world that is the most useful sub-national boundary definition to create sub-national (regional, provincial) statistics. Our software is easily adopted to be used with any sub-national topologies that are described in sufficient detail to make a conversation table (containing the metadata of the sub-national division) available. Our aim was to convert the Google Mobility Report’s European data to NUTS, and allow a combination with sub-national statistics, such as regional population or GDP figures.

Because of the frequent changes of the ISO 3166-2 standard, an API access is desirable to the latest country-specific definitions to the xxxxx.

We created an example of the harmonization potential of this variable by creating a harmonized statistic on attitudes to climate changes in xxxx European and xxx African regions. < this will overlap with the retroharmonize, but I think that we should keep it here >

Our second addition to the first release of regions is an extension to the LAU topological level of Europe, which consists of xxxxx. This extension is possible because the LAU system is part of the hierarchical European statistical nomenclature <reference description>. Here, we rely on Eurostat’s annual pre-releases and validated releases of the NUTS-LAU correspondence for each participating country. (Besides members of the European Union, several other European countries harmonize their statistical systems with Eurostat and its ESSNet network.)

**Acknowledgements**

Please add any relevant acknowledgements to anyone else who supported the project in which the software was created, but did not work directly on the software itself.

**Funding statement**

The software had no funding from any external source.

**Competing interests**

The authors declare that they have no competing interests.

**References**

Please enter references in the Harvard style and include a DOI where available, citing them in the text with a number in square brackets, e.g.

[1] Piwowar, H A 2011 Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE* 6(7): e18657. DOI: http://dx.doi.org/10.1371/journal.pone.0018657.

**Copyright Notice**

Authors who publish with this journal agree to the following terms:

Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a [Creative Commons Attribution License](http://creativecommons.org/licenses/by/3.0/) that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.

By submitting this paper you agree to the terms of this Copyright Notice, which will apply to this submission if and when it is published by this journal.