Longform

Daniel Antal, CFA

5/8/2020

library(regions)

Most economic, social and environmental situations and developments have a specific territorial dimension. Their statistical representations, which generalize many individual data points by aggregating, averaging or other means corresponds to a territory.

We are used to relatively constant territories in the forms of national boundaries. While national boundaries do change almost every year somewhere on Earth, lower level territorial units, such as administrative divisions within a country, ranked below a province, region, or state, are changing very frequently. Such changes make the creation of time series, data panels, or data joins a very painful task in practice. For example, the regional boundaries within the European Union change on average every three years, so a panel or time series of data has to account for about three boundary changes that raise various comparability problems.

The European Union harmonises the national statistical systems of 27 members states, 4 EEA countries and several other potential member states. Its regional typology, the hierarchical NUTS system is the most complex international system, and the regions package is mainly developed to help working with European regional data. However, the tasks related to such regional typologies are not unique the European Union. They are present in almost any national statistical system, and they can be similarly complex in federal states like the United States or Australia.

The [Methodological manual on territorial typologies 2018 edition](https://op.europa.eu/en/publication-detail/-/publication/d96ef933-1ec5-11e9-8d04-01aa75ed71a1/language-en/format-PDF/source-106441306) introduces three types of typologies.

* Regional typologies on NUTS1 level (states of Germany, provinces, macroregions), NUTS2 level (regions) and NUTS3 level (small regions.) Similar typologies exist in the United States or Australia.
* Local Administrative Units (LAUs) which exist in almost all nation states that are sufficiently large, and have a territorially divided public administration and statistical system.
* Grid typologies which are beyond the scope of regions.

The availability of the sub-national statistical products is lower in terms of data table completeness or timeliness. This is often due to the fact that the creation of sub-national statistical products poses significant methodological challenges.

The changes of national or provincial boundaries are rare – within a single polity. However, when we work with global data panels, or data panels of many sub-national territorial units, boundary changes are rather frequent. On a national level, just within Europe we have witnessed in the last three decades the dissolution of federal states like the Soviet Union, Yugoslavia and Czechoslovakia and the unification of Germany. Especially the break-up of the former Yugoslavia took a long time, meaning that annual national statistics contain different configurations over the years for Serbia, Montenegro and Kosovo.

On sub-national levels, within Europe regional boundary changes are so frequent that they happen on average every three years. Of course, at one time, this procedure leaves most regional boundaries intact, and change only a relatively small number of boundaries. Whenever we work with European regional data, we still have to adjust panels of statistical aggregates every three years; similarly, we must validate any regional time series every three years.

In either case, producers of statistical products do not necessarily re-cast historical statistical data for the new national or sub-national boundaries. For example, when the regional boundary definition of Central Hungary changed in xxxx, Eurostat did not require the Hungarian national statistical authorities to re-cast Central Hungary data to the new boundary definition (i.e. leaving Budapest out.)

As a result of frequent boundary changes and estimation problems, global and regional databases often contain many missing data. These datasets cannot be imputed with algorithms that were designed to impute non-aggregated, individual data.

The xxxx package is mainly aimed at validating and imputing regional statistical datasets.

* The impute\_down function and it helper functions help the projection

## Typical Problems

### Conformity With Various Typologies

The validate function family helps in finding out to conformity of your regional coding with various typologies. In the first CRAN release, only Eurostat typologies are coded, but some of the functions can be used with outher typologies, for example, with the ISO-3166-1 and ISO-3166-2 typologies. OECD uses the NUTS typologies for the EU member states, so these typologies are the most frequently used in international statistics.

Eurostat offers so-called correspondence tables to follow boundary changes, recoding and relabelling for all NUTS changes since the formalization of the NUTS typology. Unfortunately, these Excel tables do not conform with the requirements of tidy data, and their vocabulary for is not standardized, either. For example, recoding changes are often labelled as recoding, recoding and renaming, code change, Code change, etc.

The data-raw library contains these Excel tables and very long data wrangling code that unifies the relevant vocabulary of these Excel files and brings the tables into a single, tidy format , starting with the definition NUTS1999. The resulting data file nuts\_changes is included in the regions package. It already contains the changes that will come into force in 2021.

The vignette Validating Your Typology gives a real-life example to the use of these functions. Validations are the starting points for [Recoding and re-labelling](#recoding), [Aggregation And Re-Aggregation](#aggregation) and [Disaggregation And Projection](#projection) operations.

### Disaggregation And Projection

Another good example is statistical data that is derived from representative surveys. Such surveys in the European Union are usually designed to be representative of the natural or legal persons within a national boundary. The sample design and the sample size usually does guarantee that sub-samples remain representative. For example, the Eurostat statistical product xxxxx is based on the xxxxx. This survey is designed to be representative for the German population, and with certain compromise, it is a reliable estimate of the federal states of Germany like Bavaria. However, it is unlikely to be representative of further subdivisions of Bavaria – neither the sampling nor the sample size make it likely that the respondents from xxxxx represent well the population of xxxx.

For example, there is a considerable interest in US and Australia state-level, or EU regional level GDP and other national accounts aggregate variables. However, the system of national accounts is, as the name tells, ‘national’ – the complex data structures behind these statistical products require a complete aggregation of data in a tax jurisdiction, for example, which is usually the federal or nation state. Attribution of certain components of national accounts to sub-national territorial units often requires estimation and various methodological compromises. For example, an important source of the national accounts is tax report of corporation, which are usually not required to precisely attribute their financial statements to sub-national territorial units like regions in the EU, or states and territories in Australia.

The function impute\_down creates a generic solution to the problem when we want to project data from a larger territorial unit to a smaller one. For example, some state level data is missing for certain Australian or German federal states, and we want to use the national (average) values for imputation.

A special case of this projection is the creation of comparable data for small countries. Within the European Union, some small members states do not have NUTS1 and NUTS2 level divisions because of their size, but it is appropriate to keep their national data in comparision with similar sized units. We can add national (often referred to as technically NUTS0 level) data about Andorra, Cyprus, Estonia, Iceland, Kosovo, Liechtenstein, Luxemburg, Malta, Montenegro to NUTS1 and NUTS2 level datasets. Their data is often present in Eurostat datasets even they are not members of the European Union.

### Recoding and re-labelling

In the European Union, the most frequent regional statistical change is ‘recoding’ and ‘relabelling’. Such changes do not affect the statistical data, only the statistical metadata. For example, when Greece is changing the definition of two regional units, it may re-code or re-label all other unchanged units to avoid misunderstandings in regional data tables.  
Many of Eurostat’s regional datasets contain recoded geographical IDs that violate the definition of tidy data. For example, a dataset may contain xxxx for xxxx, and xxxx. Without knowing the vocabulary of the metadata for the NUTS2013 and NUTS2016 regional boundary definitions, the users of such data tables do not know that in fact these data make a time series, because they refer to the same region under a different regional code. The xxx functions of the xxxx package take care of this problem. In these cases, naivly bringing a data table to a tidy format creates a lot of seaming missing data. These data, however, should not be imputed because the actual, seemingly missing data can be found under different labels. By harmonizing the row (or column) IDs, we can create a truly tidy, and often complete data table.

## Aggregation And Re-Aggregation

The European Union sometimes publishes data on NUTS2 level regions, but not on the larger NUTS1 level macroregions or provinces. In many cases, this information is fully present in the NUTS2 datasets, and a simple aggregation can produce a NUTS1 level dataset, because each NUTS2 region is strictly assigned to one broader NUTS1 marcoregion. For count-type data, we can simply aggregate the constituent NUTS2 level data to NUTS1 level.

A frequent problem is the re-assignment of a smaller territorial units, such as lower level administrative unit or a lower NUTS unit to another larger unit, or a pre-existing larger unit. This happened in the case of xxxx, when the former xxxxx unit was re-assigned from xxxxx to xxxxx. Obviously, this created a comparability problem, because the old xxxxx data are aggregated for a different part of xxxx than the xxxxxx. The aggragate\_over function helps to solve such problems.

### Additive Reconfiguration of Territorial Units

Xxxxxx Again, in these cases what we are doing is not imputation in the strict sense. Adding together Vilnius and xxxxx creates the xxxxxx; xxxxx and Budapest creates xxxxxx. Because this is an error-prone procedure, we create tidy, filled dataset that clearly indicate the ex-post data manipulation in this case.

### Boundary Changes With No Solution

In some cases the boundary change cuts across all territorial typologies that re-configuration is impossible from aggregated data. Unless we have access to the individual level data (which is not within scope of the tasks helped by this package), it is not possible to create time-wise comparable data for these units. We must treat the observations on these units logically missing, which may or may not be imputed with any techniques. It is usually helpful to mark territorial observations in panels or time series that have such a structural change.

## Boundary definitions

Our package uses a number of boundary definitions that are easily expanded. Our functions are generic, i.e. they treat boundary information as an input variable, which, if well formatted, can describe for example the local administrative units of any country on Earth.

All the boundary definitions are valid for a certain time period and certain boundaries. It would be tempting to use ISO-3166-2 definitions for sub-national boundaries, however, the ISO-3166-2 itself is changing relatively often, and in time series or panel data we can only refer to a certain version of the ISO-3166-2. (Which is, though public, is not open source data, and it is rather costly to access the original definitions.) However, the producers of statistical products do not necessarily follow the ISO-31660-2 standards, or, if they do, they publish the same information on their own.

There is no universal solution to incorporate all boundary definitions for every country in an R package, but we believe that it is possible to use the most commonly used boundaries and geographical vocabularies as metadata in our package. This package is itself an offspring of certain functions added to the eurostat package of the rOpenGov community. It was detached from the package because even in Europe boundary and metadata changes are frequent, but also, because the programmatic solutions needed to address them can be applied to the statistical products of OECD, the United States or Australia, too. We try incorporate boundary metadata in our package so taht users working with data from these sources do not have to input the boundary metadata information for each use of our functions.

Of course, we are welcoming contributors from anywhere in the world to provide typology descriptions and vocabularies to ease this work.

For example, the European Union is publishing its NUTS system every three years, and it is also publishing information on countries that are officially participating in ESSNet system of Eurostat, including the former EU member states United Kingdom and the prospective member states, or the European Economic Area countries. Australia xxxxx, the United States, and OECD….

Our package allows the use of any custom structural metadata information, and in the xxxx vignette we give a simple example of xxxxx, ass.. We would like to encourage users and future contributors to provide us structural metadata tables that we can incorporate into the package data. However, we believe that the priority should be given to the UN, European Union and OECD formats, because these organizations produce many data tables with heterogenous and changing boundaries. Federal and national statistical offices, for example the US, the Australia or the German are usually consistently creating their own sub-divisional data; adjustments are likely to be needed with international statistical products.

## Projection from larger units to smaller units

After creating a truly tidy dataset, we can devise various imputation strategies that are well-suited for sub-divisions of statistical aggregates. Simple or complex imputation methods, such as imputing the median value, are not valid when the data has a particular structure that breaches the imputation assumptions. For example, if we know that Bavaria is a (relatively small) part of Germany, we will not impute the mean population value of Germany and France to this missing unit, because we are aware of the fact that Bavarian population is strictly and significantly smaller than that of Germany’s.  
We are projecting mean values from larger units when we know that the smaller, constituent unit’s data is hidden in the larger unit. For example, if we want to impute a typical, for example, median survey answer to Bavaria, we know that German typical (median) value is calculated from various sub-samples that include the Bavarian responses. One possible, and often correct imputation strategy is to use Germany’s median value as an estimate for Bavaria’s.

Is this a valid imputation approach? For non-additive regional statistics in Europe, generally it is a good approach, because the European statistical sub-divisions are designed in a way that smaller (constituent) territorial units have, as a general rule, a diminishing intra-territorial heterogeneity for most social and economic variables. This is particulary true for smaller subdivisions of Bavaria itself. The first-level subdivisions in Europe are often historically set boundaries, but the lower sub-divisions are usually created with this homogeneity-criterion. (This is the reason that the EU uses NUTS or statistical regions, and not local administrative units for statistical purposes. While xxxxx may be set on historical precedents and legal considerations, the statistical aggregation is done in the NUTS system. ) The approach is obviously not valid in the case of additive data, for example, with population or GDP data, but it is a valid approach for GDP per capita values. Needless to say, if we have access to microdata, and the amount and sampling of the microdata is adequate, the best way fill out the missing information is to re-aggregate the data for missing units. However, for a number of practical and legal reasons this is almost never the case. In these cases, completely different methods should be used. [See: }

## Addition from smaller units to larger units

In some cases, when the enumeration of the statistical information is bottom-up, like in the case of population data, we may have very precise information on the lowest subdivision level. For example, in the EU xxxx population. In the case of strictly aggregate data, any imputation method will create terrible estimates, while we can just simply re-aggregate the missing data following the xxxxxx. Of course, we must make sure that we are always following the correct boundary definition.

## Time-series

Because of the complexity of sub-national data, there are likely to be missing elements in time-wise comparisons. For example, we may find that some data is available for many years on a German federal level, and for most states, but a particular annual data is missing for Bavaria. It would be tempting to use a simple interpolation of the available data, but often a naïve interpolation is not the best method, because, as we have seen, often the data is available at a different aggregation level or under a different metadata label. It is imperative to first make sure that we exploit all information that is present in a different territorial structure for any given year, and then start inter-temporal imputation with interpolation, or carry forward/backward, forecasting and backcasting algorithms. Xxxx At last, we may go back to the cross-sectional data, and may fill in some cross-sectional data gaps with interpolated or extrapolated data. Aggregating an estimate for Bavaria from previous or next-years constituent xxxx data is very likely to be a far better estimate than any imputation results, for example, a simple substitution of a median value from all German or EU NUTS1 units. The territorial aggregation structure gives plenty of exploitable information. Vignette xxxxx shows a case study for the Eurostat xxxx sss. .