`cat2cat`: An R Package for Handling an Inconsistent Coded Categorical Variable in a Panel Dataset

Maciej Nasiński[1]

Abstract

This paper introduces `cat2cat`, an (R Core Team 2012) package that provides users flexible tools to handle an inconsistent categorical variable in a panel dataset. The categorical variable levels could changing in time, some categories are grouped and others are separated, new ones are added or old ones are removed. Examples of a dataset with such inconsistent coded categorical variables are ones linked with the The International Standard Classification of Occupations (ISCO) and the International Classification of Diseases (ICS). The `cat2cat` package enable unifing of an inconsitent coded categorical variable between two different time points in accordance with a mapping table. The idea is to take categorical variable levels from a specific period and apply them to the neighbouring period. It is done by replicating an observation if it can be assigned to a few categories, and then use simple frequencies or statistical methods to approximate the probabilities of being assigned to each of them. The `cat2cat` package extends the scope of the available statistical analyses in a panel dataset with an inconsistent coded categorical variable, as normally such variables are removed or force the dataset aggregation.

Keywords: categorical, panel, longtitudal, statistics, unify, inconsistent, map

In many scientific projects where a panel dataset contains a categorical variable, one of the possible obstacles is that the data provider is changing such variable levels in time. The categorical variable levels are regularly changing if the variable has an evolutionary nature. For example, some categories were grouped and others were separated, a new one is added or an old one is removed. Examples of a dataset with such inconsistent coded categorical variables are The International Standard Classification of Occupations (ISCO) and the International Classification of Diseases (ICS). Both classifications are regularly updated to adjust to, for example, new scientific achievements. Specifically, new scientific achievements can create new occupation types on the market or enable discovery of new diseases types.

The core idea for the R `cat2cat` package comes from the procedure proposed by (Broniatowska, Majchrowska, and Nasiński 2020), which is to apply statistical

Email address: `maciej.nasinski@uw.edu.pl` (Maciej Nasiński)

analysis to unify an inconsistent coded categorical variable in a panel dataset. The referenced paper was a practical application with a concise description of the procedure. This paper is the first place where the procedure is detailed described, extended and implemented. The procedure is offered to the scientific community in the `cat2cat` R package. This package and the related details are available on CRAN at https://CRAN.R-project.org/package=cat2cat. `cat2cat` R package was designed to offer an easy and clear interface to apply a mapping table provided by a data maintainer or built by a researcher. The main objective of contributed `cat2cat` package and procedure is unifying a categorical variable with an inconsistent encoding over time in a panel dataset. In other words, we want to map the categorical variable levels from one period to another in accordance with a mapping table. If the mapping table contains only one-to-one mappings then the unification process is straightforward, it is a more complex process when working with one-to-many mappings. The main rule is replicating the observation if it can be assigned to a few categories. Specifically, for each observation, we look at the mapping table to check how the original category could be mapped to the opposite period one. Then, we use simple frequencies or statistical methods to approximate the probabilities of being assigned to each of them. For each observation that was replicated, the probabilities have to add up to one. The procedure distinguishes the different mechanisms for panel data with and without unique identifiers.

A typical strategy to handle categorical variables with inconsistent encoding across time is removing it. The usual impression is that this is the only solution but we decided to remove a potentially important variable from the analysis. Another possible method if the categorical variable has a hierarchical structure is to simplify it for each level and then aggregate it. A hierarchical categorical variable is, for example, an occupation or disease codes, where each next character of the code provides a higher level of detail. Then, we can simplify the hierarchical variable by considering only the first n characters. Another assumption is that the first n characters maintain their meaning over time. The last step is aggregating the data for each simplified level, for example, by taking an average of the continuous variables. This simplification-aggregation method is causing a loss of information and is questionable in many scenarios. The `cat2cat` procedure is an alternative for these imperfect solutions.

The study adds to the literature by providing a new way of handling an inconsistent coded categorical variable in a panel dataset, a topic that is hitherto under-discussed. The `cat2cat` procedure extends the scope of the available statistical analysis in panel datasets such as ISCO or ICD based ones. The algorithm can be applied to any type of panel dataset regardless of the scientific field. As the presented algorithm is a novel solution so is vulnerable to be incomplete.

References

Broniatowska, Paulina, Aleksandra Majchrowska, and Maciej Nasiński. 2020. "Age Structure of Employment and Wages. An Analysis Across Occupational Groups." Artykuły / Articles. *Central European Journal of Economic*

Modelling and Econometrics, no. No 3: 227–50. https://doi.org/10.24425/cejeme.2020.134747.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.