Bridging macroeconomic data between classifications

Topic: Micro data
Author: Mattia Cai
Co-Authors: JosÃ© M. RUEDA-CANTUCHE

In applied research and policy analysis work, it often becomes necessary to link datasets that adhere to different statistical classifications. Most commonly, this occurs as a result of the periodic revisions that industry and product classifications are subjected to. In the late 2000s, for example, the national accounts of European Union member states switched from revision 1.1 to revision 2 of the â€˜Statistical classification of economic activities in the European Communityâ€™ (NACE, from the French acronym). When the boundaries of an industry shift, comparability over time is lost for important economic variables such as value added or employment. Then, obtaining consistent time series for the industry-level variables of interest requires conversion between classifications.
The need for reclassification, however, can also arise for other reasons. Sometimes, for instance, data need to be reorganized according to a classification other than that in which primary collection took place before they can be used as an input to a certain economic modeling exercise of interest. Consider data on final use by households. In the supply and use framework, each transaction is categorized according to the characteristics of the good or service that is being exchanged. In a European context, this means that data on final use by households have to be organized according to the â€˜Statistical Classification of Products by Activityâ€™ (CPA). Household surveys, however, typically collect information about the purpose for which expenditures are made, and not about the type of goods or services that are acquired. These surveys usually adopt the â€˜Classification of individual consumption by purposeâ€™ (COICOP). Before the data can be incorporated in the IO framework, they must undergo conversion from COICOP to CPA.
In the context of their institutional activities, national statistical offices do construct conversion factors that allow bridging between classifications. That kind of information, however, is not typically released to the public. Furthermore, even when conversion factors are available, it is rarely the case that the degree of aggregation is aligned to the needs of the analyst. In practice, when it comes to classification issues, independent researchers are generally left to their own devices.
To the best of the authorâ€™s knowledge, the academic literature provides little guidance as to how to handle data reclassification problems. Like other common data management tasks, classification issues are rarely discussed. By and large, it appears that in applied work practitioners predominantly use expert judgment to establish best-guess correspondences between aggregates of the source and target classifications. The process of specifying such correspondences is often tedious and its outcome somewhat subjective.
This paper describes a simple, mechanical and reproducible approach to the construction of bridge matrices under conditions of data availability that are likely to be met in most circumstances. From a practical standpoint, the essential requirement is that there exists an earlier or later time period â€“ or a geographical area that is similar enough to the one of interest â€“ for which the relevant economic variable can be observed in both the source and the target classification. Using this information, we estimate a contingency table that links the two classifications by means of bi-proportional scaling methods. Finally, data reclassification is carried out using conversion factors computed from that table.
Estimating an unknown matrix by proportionally scaling an initial guess â€“ typically referred to as the seed or prior matrix â€“ using known marginal totals is a routine practice in a variety of fields. In input-output economics, the procedure is known as RAS. What is challenging about the specific RAS application discussed here is that it is not obvious how to construct a plausible seed matrix. In the spirit of Lenzen et al. (2012) and Lenzen and Lundie (2012), a simple option would be to use a binary seed matrix based on a readily available qualitative table of correspondences between classifications. In fact, we argue that from the same table of correspondences a more informative

prior matrix can be constructed just as easily. In a nutshell, the proposed seed matrix is compiled by counting the number of fundamental items (i.e. items defined at the most disaggregated level of the classification) that simultaneously contribute to a given pair of source-classification and target-classification aggregates.

We examine two case studies in which the conversion factors used by the statistical office are known and find that, in spite of its simplicity, the proposed approach yields encouraging results. We then try to assess the performance of the procedure in a more general context using Monte Carlo simulation.