MikroSim – Developing a Microsimulation Data Center

Keywords: microsimulation, data analysis, official statistics, data center, confidentiality

1. INTRODUCTION

National Statistical Institutes face major challenges in the Digital Era. There is an expectation of the Stakeholders of official statistics that more timely and integrated statistics will be produced by the statistical system for an ever-growing list of users and uses, especially for policy analysis. Additionally, institutions and researchers aim at investigating own studies on microsimulation data which urges the needs of implementing carefully access rules up to an adequate legislation.

In this context, the project "Multi-sectoral Regional Microsimulation Model" (MikroSim) funded by the German Research Foundation is jointly conducted by the universities of Trier and Duisburg-Essen in cooperation with the Federal Statistical Office of Germany. The aim of the research group is the scientific foundation of a Microsimulation Center and the implementation of a Simulation Data Center in analogy to Research Data Centers. The presentation will provide an overview of the results already achieved within the MikroSim-Project as well as outlooks on future work, including the establishment of a Microsimulation Data Center.

The MikroSim project aims at providing an assessment base for evidence-based choices between different socio-political options for action. In the first funding phase of the research group (2018-2021), one of the core tasks is to generate and update the simulation base data set. By estimating transition probabilities, the base population is projected into the future. These transition probabilities are estimated using theoretically grounded models, e.g. care dependency, fertility, income, etc., based on longitudinal data sets of empirical social research and official statistics. The two central areas of application of the first funding phase are the impact of demographic change on care and the integration of migrants into the labor market.

The objective of the second project phase is to build up a sustainable, integrated and constantly updated database to enable open and reproducible research. The MikroSim simulation infrastructure created within the first phase builds the foundation for the conception of a nationwide Simulation Data Center in Germany. The MikroSim model can easily be supplemented with new simulation modules and thus promotes further research cooperation. For the sustainable continuation and expansion of the research group, the development of three concepts is essential: (1) a meta data concept, (2) a roles and rights concept, and (3) a security concept to guarantee statistical disclosure control. The focus is particularly on ensuring data security, access and availability. Furthermore, the thematic extensions of the second phase include the development of a housing module for MikroSim that reflects the housing situation in Germany.

2. BUILDING A MICROSIMULATION INFRASTRUCTURE

In the second phase of the research group an implementation of the data stock as available data pool for the research community is foreseen. The goal is the sustainable development

of a Simulation Data Center. This also requires a thematic extension. Therefore, the research work is divided into three blocks. A first block contains the technical core to be able to build the (statistical) methodology for a data center. This block contains also basic work related to confidentiality procedures and availability of data for researchers. Two further blocks are application-oriented thematic priorities to expand the data pool, which on the one hand deal with medical-sociological topics and on the other hand, deal with sociological-economic infrastructural topics.

The focus of subproject 1 is the development of statistical methods to build up a Simulation Data Center. Suitable methods of editing as well as statistical evaluation and modeling will appropriately be applied and extended. Moreover, methods of statistical confidentiality are required for the development of such a Simulation Data Center (Subproject 1). Microsimulations on large data sets require considerable computing times due to their high degree of complexity. In Subproject 2, methods for complexity reduction allow for more efficient analysis based on appropriate sampling methods. In addition, a simulation-based methodological procedure is developed to model hypothetical educational trajectories and latent decision processes in school transitions (Subproject 3).

The synthetic data set, the simulation environment and the estimation methods is to be made available to other content-oriented research groups. The microsimulation infrastructure created this way is internationally unique in its scale and allows for investigations of socially and politically relevant by the international research community.

2.1. Meta data concept

The aim to build up a Simulation Data Center requires the development of a sophisticated metadata concept for simulation studies including a versioning system. Metadata provide structured information on contents and characteristics of data sets to make them easier to handle and use [1]. Metadata can be categorised into structural and reference metadata. Structural metadata are necessary for the identification and understanding of statistical data. They include variable names, the title of the data set, information on units of measurement (e.g. Euro, Dollar, etc.) and time dimensions [2]. Structural metadata should be delivered with the data set or integrated into it (e.g. by using appropriate variable labels) [2]. Reference metadata (also called ‘explanatory metadata’) provide information about the content and quality of statistical data or statistical time series data [3]. Reference metadata include information about concepts, methods and the data quality [4]. The versioning of the provided metadata shall ensure that for each (updated) database in the Simulation Data Center as well as for all performed simulation tasks clearly assignable metadata are available.

2.2. Role and rights concept

In order to enable the data and simulation models for the general opening of the Simulation Data Center as infrastructure, a role and rights concept for the regulation of data access must be developed. For this purpose, three access options are created. Using an internal workstation at the University of Trier, researchers can access the internal server structure on a secured computer and implement their own models. In order to guarantee the data protection of the simulation results, automated functions as well as on-site staff will ensure direct control. Using the external server structure, researchers from other locations can run simulations on anonymized or limited data. As a third option, simulations can be

programmed based on the freely accessible, structure-equivalent data and implemented on the overall data by an employee on site (similar to controlled remote data processing).

2.3. Confidentiality

To provide the MikroSim databases and simulation structures within the Simulation Data Center, comprehensive security concepts are required as a supplement to the role and rights concept. The basic data set should be anonymized so that data access via an external server is possible.

In order to prevent a re-identification of units of the base population, appropriately adapted secrecy concepts have to be developed and implemented. The adapted secrecy methods aim at the anonymization of the individual values, which effectively prevent the identification without covering or changing basic characteristics of the base population. The focus is on sensitive information, such as the income size of households (disclosure of attributes) and general protection against identification (disclosure of identity). In addition to common concepts based on *k*-anonymity such as cell blocking and rounding procedures, data-modifying methods such as smoothing approaches and stochastic overlay are applied [5].

3. OUTLOOK

Modern dynamic microsimulations are increasingly performant thanks to improved computer architectures and the increasing availability of data. They provide the ideal link between research and *Policy-Oriented Research*. Both Eurostat and the European Commission strongly support *Policy-Oriented Research*, and the topic becomes more and more important particularly with regard to regional issues. Relevant questions cover predominantly sociological and economic aspects. The current discussion regarding Covid-19 shows that a variety of topics are of interest, ranging from medical, and currently epidemiological, to infrastructural questions. The focus of the research group MikroSim on regional microsimulations and the provision of small-scale data allows for interdisciplinary evidence-based research on a broad range of relevant issues.

This presentation outlines the objectives and challenges in developing a Microsimulation Data Center in analogy to Research Data Centers, as part of the MikroSim project. Work has already started within the first project phase financed by the German Research Foundation and concrete action plans for the second project phase are currently being developed. Considerable efforts will be spent in the definition of technical solutions and legal frameworks for the Microsimulation Data Center in particular related to the security concepts that secure data access and allow small-scale research while ensuring confidentiality.

REFERENCES

[1] Richter A. / Weil, S. (2005): Metadaten. Eine Grundlage für die Auswertung amtlicher Statistiken durch die Wissenschaft. In: Statistische Monatshefte Rheinland-Pfalz, Vol. 1/2005, pp. 12 – 18. URL: https://www.statistik.rlp.de/fileadmin/dokumente/monatshefte/2005/Januar/01-2005-012.pdf.

[2] Eurostat (2020): ESS Reference Metadata Reporting Standards. URL: https://ec.europa.eu/eurostat/de/data/metadata/metadata-structure.

[3] Statistical Data and Metadata eXchange (SDMX) (2016): SDMX Content-Oriented Guidelines. February 2016. URL: https://sdmx.org/wp-content/uploads/SDMX_COG_2016_Introduction.pdf.

[4] Simeoni, G. (2018): ESS Standards for reference metadata and quality reporting (ESMS, ESQRS, SIMS). In: ESTP Training Course "Information standards and technologies for describing, exchanging and disseminating data and metadata", Rome, 19-22 June 2018.

[5] Templ, M. (2017): Statistical disclosure control for microdata: methods and applications in R. Cham: Springer.