2.2. THE KRTK'S LINKED ADMINISTRATIVE PANEL DATASET - ADMIN4

ANNA SEBŐK

In the summer of 2022, the Databank of the Centre for Economic and Regional Studies[1] created the newest public administration panel database, the Admin4, which is the fourth admin dataset in a row. Linked Administrative Panel datasets, including Admin4 (covering the period 2003-2021), were created using a data integrational method, anonymised, but at the same time include health, education, labour market and unemployment data for half of the Hungarian population at an individual level, as well as the characteristics of Hungarian enterprises as reported in the Wage Survey. The datasets do not contain natural identifiers neither household tables, however, are uniquely detailed from a research point of view. Corresponding to the criteria of modern international scientific data management systems, the data are being applied solely for scientific research with secure server connections under controlled conditions. The previous waves of the dataset (Admin1, Admin2 and Admin3) are widely known and respected among researchers in the labour market, health-economics, and in other fields of economics, both nationally and internationally. This is described in more detail by János Köllő in subsection 2.1 of the chapter In Focus.

Admin4

The research of information accumulated in public administration organisations has a long history in Hungary. The possibility and the method of creating research databases for the combined examination of several registers have so far been laid down in Act CI of 2007 on ensuring the availability of data for decision-making (*Scharle*, 2019). The data integration necessary for the creation of combined databases could be initiated by the heads of the budgetary bodies specified in the above Act (for scientific purposes, the President of the Hungarian Academy of Sciences), and technically could only be carried out by the National Infocommunications Service Company (NISC Ltd).

The basic principle of all waves of the Linked Administrative Panel datasets is to unite all research-relevant registers that are available and linkable at the time of the linking. Thus, in the most recently linked Admin4, the data of the National Health Insurance Fund Administration (NEAK), the Hungarian State Treasury (MÁK), the Educational Authority (OH), the National Office for Vocational and Adult Education (NFSZH–SZIR) and the Ministry of Technology and Industry (TIM) were linked at individual and enterprise level. Thanks to the linking, the following sets of data can be researched in their context:

[1] The Databank of the Centre for Economic and Regional Studies (KRTK), Hungarian Academy of Sciences

Healthcare: home address data (at regional level), data on social security status, public health care, general practitioner, outpatient and inpatient care, mortality, change of prescriptions, drug expenditure, cash benefits, CT/MRI and dental care, as well as data on greater interventions, at individual level.

Labour market: data sets on employees, labour market, public employment and labour intermediation, at individual level.

Social transfers: data on pension payments, family allowances, cash benefits, data on unemployment and labour market programmes, at individual level.

Field of education: higher education training, higher education status, public education status, school-leaving certificate, National Assessment of Basic Competences data, at individual level.

Enterprise data: data from corporation tax declaration (tao) and Wage Survey records, at enterprise level, linkable to individuals.

Secure linkage of the data

The data integration process is based on the systematic linking of the different state administration registers on individual level, possibly with some other unique identifier. As a result of data integration, the contents of the different administrative registers can be analysed collectively, on an individual or entreprise level, yet anonymously. The data from data holders are linked by codes and then the natural identifiers are deleted. A link code can be any unique identifier that is available at the data holders'.
The KRTK Databank's Linked Administrative Panel Dataset, unlike those of the state administration, is not comprehensive, but covering a 50% sample of the population with a social insurance number (Government Decree 335/2007 [XII.13.]). The sample was sorted from a file covering the entire Hungarian population, containing those who held a Social Security Number in 2003 and was executed by the National Health Insurance Fund.
The selection of the population started with the establishment of a list of link codes (social security numbers) known to the other data holders listed above. Using a hash algorithm[2] generated by NISC Ltd. for anonymous data capture specifically for the given linkage, the registry operator who selected the base population assigned unique technical identifiers to the original codes. In the next step, the other data providers also extracted the data related to the

[2] Hash-algorithm is a one-way coding routine, which forms output data from input data according to the following conditions: it always provides the same output from a given input information, as well as the output data clearly refers to the input data, but the input data cannot be generated from the output data. In this procedure, the smallest change of the input data results in a completely different output. Hash methods are also used for compression, password storage, and searching procedures. In this case, the procedure serves the generation of anonym, technical identifiers.

base population and passed it in *hashed* form to the NISC Ltd. which eventually merged and anonymised the database. Thus, the resulting database does not contain any natural identifiers. In the case of the linked public administration panel datasets, in addition to the linking of individual data (holders of social security number), the employer data are also hashed: the employers' tax number is used to link the data of the Ministry of Finance, the Ministry of Technology and Industry and the Hungarian State Treasury. The KRTK Databank receives this data without the original identifiers (e.g. social security number and employer tax number), but with the data of the individuals in a raw form, not yet suitable for research. Once it has been received, the cleaning of the data begins, in accordance with research questions.

Data integration inevitably involves and perpetuates those mistakes in data content or recording that originally come from the data providers. The preliminary cleaning and interpretation of data is required as administrative registers have a specific content, terminology and structure, i.e. they initially have to be interpreted within the logic of each administrative regsitry. Understanding this logic is also essential for data processing and analysis in terms of research purposes.

After linking the data sets, the KRTK Databank creates a longitudinal panel, by harmonising hundreds of raw datafields of the latest linkage. Professional data cleaning and harmonisation, covering a period of almost 20 years, takes a long time. Therefore, the earliest date for a scientific analysis of the currently linked database is around 2023-2024.

Data cleaning

As to the characteristics of administrative data detailed here, both longitudinal and cross-sectional consistency testings are part of the data cleaning process. A cross-sectional consistency testing involves comparing the content of different data sets at a given point in time. Seeing the arising issues and questions, we can gain an understanding of the limitations of the data. At the same time, by longitudinally monitoring, we observe the content and changes in the data series from year to year and then record them in a metadatabase for the entire observation period. In the cleaning process, following the above-detailed groundwork, analytical decisions are made to extract the data content appropriate to the research questions from cells that are in themselves would only have administrative meaning.

The first round of the cleaning starts with a review of the data received from the data providers and the NISC Ltd.

Afte the validation, the raw data fields are iteratively transformed to create variables corresponding to different research questions, first only within the register content of each data source.

Once the data has been cleaned, explored and harmonised by each data source, the construction of a great database begins, which is reduced in size but combines the most important variables. The resulting database contains the monthly status of individuals, in the case of Admin4: from 2003 to 2021. The database inconsistencies that arise after the linking will be checked and managed.

This creates a cleaned, huge database. Using the variables from complex sources, mini-analysis are launched in order to identify potential mistakes, gaps and interpretation

limitations that may arise when the data would be used later. Issues and solutions that arise in this way are also incorporated into the corrected database.

This is followed by a collective phase of harmonisation, during which experts from different areas of the scientific community are given the opportunity to use the Linked Public Administration Panel Database, and to incorporate their previous experience into the cleaning process and improve it. The information gathered during the collective harmonisation phase (even program codes) will be incorporated into the first round of database-cleaning of the next wave of Admin, organically improving it. In a similar way, all questions, problems and feedback that arise during subsequent data use, as well as the program codes written as solutions, are incorporated into the Admin-file cleaning procedure.

Protection of personal data

The KRTK Databank's Admin4 database covers a 50 percent sample of the Hungarian population. When linking the data, the NISC Ltd. pseudonymizes this sample, thus distorting the social security number and tax number. Data holders do not provide natural identifiers (name, place and date of birth, mother's name and surname). In order to reduce any risk of identification, the body responsible for the linkage (National Infocommunications Service Company) also anonymises. In doing so, it aggregates special categories with a critically low number of cases (e.g. codes of different diseases, educational attainment).

The Databank then performs further codings, deletions, and merge categories in the database to reduce the risk of identification. The KRTK Databank only allows research work on the resulting database in a secure and closed server environment by researchers and thesis-writers working in scientific institutions or with an appropriate and verifiable scientific objective. Research requiring more detailed data can be carried out in a camera-monitored Research Room of the Databank, also in accordance with the above conditions. Professional protection of individual data and scientific ethics criteria together provide protection against the risk of identification.

An independent, data-driven, cross-disciplinary social research practice that pays particular attention to the verifiable use of data is necessary for policy making, for the development of a well-functioning scientific and public administration sector, and for the development of data guidelines applicable to all areas of life.

Contact

The Databank of the Center for Economic and Regional Studies conducts surveys and creates administrative-based datasets in uniquely diversified themes, runs a scientific research room that makes possible the broad analysis of microdata, sets up register-based datasets applying a data-integrational method, organizes training courses of STATA and maintains a research laboratory with its own computers. For general information please contact the adatbank@krtk.mta.hu email address, for data request use the following: adatkeres@krtk.mta.hu.

References

335/2007. (XII. 13.) Government Regulation on the implementation of the Law of Act CI. of 2007 on ensuring access to data required for decision preparation

Law of Act CI. of 2007 on ensuring access to data required for decision preparation

Ágota Scharle (2019): Hungary: a case study on improving access to administrative data in a low-trust environment. Published in *Crato, N,-Paruolo, P.* (ed.): Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design. Springer, pp. 119-130.