BUDAPEST WORKING PAPERS
ON THE LABOUR MARKET
BWP – 2019/2

The Panel of Linked Administrative Data of CERS Databank

ANNA SEBŐK

BWP 2019/2

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES
BUDAPEST, 2019

Budapest Working Papers on the Labour Market

BWP – 2019/2

Institute of Economics, Centre for Economic and Regional Studies

The Panel of Linked Administrative Data of CERS Databank

Author:

Anna Sebők
junior research fellow
Institute of Economics, Centre for Economic and Regional Studies
E-mail: sebok.anna@krtk.mta.hu

December 2019

The Panel of Linked Administrative Data of CERS Databank

Anna Sebők

Abstract

The Databank of the Centre for Economic and Regional Studies has established Admin3, the third round of the Panel of Administrative Data. The data source contains administrative data of the National Health Insurance Fund Administration, the Hungarian State Treasury, the National Tax and Customs Administration, the Ministry of Finance and the Educational Authority on a sample of 50% of the Hungarian population between 2003 and 2017. Based on these data an individual-and firm-level, anonymized panel dataset will be created which contains monthly individual observations on educational activity, employment status, occupation, gross wages, health events etc.1

Keywords*:* Register data, data integration, labor market data, health data, educational data, National Assessment of Basic Competencies data

JEL codes: C8, C80, C81, C82, C89

[1] Special thanks to Bálint Mónika, Czethoffer Éva, Köllő János, Hönich Heléna, Sinka-Grósz Zsuzsanna, Szabó Endre and Tir Melinda, the fellow workers of the Databank of the Centre for Economic and Regional Studies.

A KRTK Adatbank Kapcsolt Államigazgatási Paneladatbázisa

Sebők Anna

Összefoglaló

A Közgazdaság- és Regionális Tudományi Kutatóközpont Adatbankjában létrejött a egújabb Kapcsolt Államigazgatási Paneladatbázis, az Admin3.
A különböző államigazgatási nyilvántartások személyi szintű adatösszekötése – a korábbi hullámokhoz hasonlóan (Admin1 és Admin2) – lehetővé teszi a magyar lakosság 50 százalékos mintáján a népesség munkaerőpiaci, munkanélküliségi, oktatási és egészségügyi jellemzőinek tudományos vizsgálatát 2003 és 2017 között. Az egyéni és vállalati szintű, hosszú idősoros, ugyanakkor természetes azonosítókat nem tartalmazó paneladatbázis egyedülállóan szerteágazó tartalmú. Az Admin3 forrásregiszterei között szerepelnek a Nemzeti Egészségbiztosítási Alapkezelő, a Magyar Államkincstár, az Oktatási Hivatal, a Pénzügyminisztérium és a Nemzeti Adó- és Vámhivatal adatbázisai.

Tárgyszavak: államigazgatási adat, adatintegráció, munkapiaci adatok, egyészségügyi adatok, oktatási adatok, OKM adatok

JEL kódok: C8, C80, C81, C82, C89

Introduction

In the summer of 2019, the Databank of the Centre for Economic and Regional Studies[2] created the third Linked Administrative Panel dataset. The Linked Administrative Panel datasets, as well as the Admin3 (2003 – 2017) have been created using a data-integrational method. The dataset is anonymized, however it contains half of the Hungarian population's medical, educational, labour market and unemployment data at an individual and firm level. The datasets do not contain natural identifiers neither household tables, nonetheless they are uniquely detailed for scientific research.
The previous waves of the dataset are widely well-known and acknowledged on various national and international scientific platforms. Corresponding to the criteria of the international scientific data management systems, the data are being applied solely for scientific research with a safe server connection under controlled conditions.

Data aggregation

The research of information accumulated in public administration organisations has a long history in Hungary. The practice of administrative data management of the Hungarian Central Statistical Office, and the possibility to research and apply those mentioned accumulated data are all secured by national and international laws (see: GDPR). Based on this fact, research databases aiming to analyse the original register or other multiple registers collectively, can be produced. The latter is made possible by Act CI. of 2007 on ensuring access to data required for decision preparation. The necessary data integration to generate the linked datasets can be initiated by the executives of budgetary bodies recorded in the law mentioned. Technically, this can solely be executed by the National Infocommunications Service Company Ltd (NISC).
The principle of all waves of the Linked Administrative Panel datasets is to unite all research-relevant registers that can be found and are linkable at the time of the linking. Thus, individual- and firm level data of the National Insurance Fund Administration, the Hungarian State Treasury, the Educational Authority, the Ministry of Finance, and the National Tax and Customs Administration were linked in the latest Admin3 dataset.
Thanks to the linking, the following data – similarly to the previous Linked Administrative Panel datasets – can be researched in their context:

2 Formerly: The Databank of the Centre for Economic and Regional Studies, Hungarian Academy of Sciences

Healthcare: an anonymized identifier generated from the Social Security Number, data about the Social Security Number register, home address, term of social insurance, public health care, general practitioner, in-and outpatient care, death, prescription redemption, social security and- monetary provisions, on an individual level.
Labour market: data about the employee, labour market, referral of public employment and labour force, on an individual level.
Social transfers: data about pension payments, monetary provisions, unemployment, data related labour force programs, on an individual level.
Education: data about higher educational training, higher educational relationship, public educational relationship, maturity exam, National Assessment of Basic Competencies, on an individual level.
Firms: data from corporation tax declaration and NES Wage Survey records, at a firm-level, although connected to individuals.
The basis of the data-integrational method is to link together various registers of public administration on an individual level, or based on other unique identifier units. This results in the situation whereby the context of different administrative registers become appropriate to be analyzed collectively, by each observation, however in an anonym way. Data of data-owners are linked together by generated anonymized linkage codes of the identification number, then the natural identifiers are deleted. A joint code can be any unique identifier that may be found at multiple data-owners, making it possible to analyse the final dataset collectively, which contains data from diverse registers.
As the Linked Administrative Panel dataset is a universal material concerning the whole population of Hungary, the number of observations in the sample can not go beyond the 50% level of the whole population (335/2007. (XII. 13.) Government Regulation). The sample was sorted from a file containing those who held a Social Security Number in 2003, and was executed by the registry covering the quasi-entire Hungarian population, namely the National Insurance Fund Administration. Sorting the population starts with creating a list of joint codes (Social Security Number). The NISC Ltd. is responsible for the anonym data linking by using a generated “hash-algorithm”. The operator of the register that sorts the population orders unique, technical identifiers to the original codes. In the next step, the other data providers also sort their data that relate to the population and then transmit the hashed data to the NISC which later merges and anonymizes the dataset. Thus the dataset does not contain any natural identifier. In the case of the Linked Administrative Panel datasets besides the individual-level data (population holding Social Security Number), the employment data are hashed as well: the employers’ data from the Ministry of Finance Nes Wage Survey, the National Tax and Customs Administration, and the Hungarian State Treasury are also merged to the

dataset by using the employment tax number. After merging and linking the individuals' data while dropping the original identifiers (e.g. Social Security Number, employment tax number), the raw dataset is sent to the CERS Databank for data cleaning and harmonization.

About the administrative data

As a result of the data-integration, the contextual mistakes and those that were made during the data recording process by data providers inadvertently remain in the raw dataset. In the process of preliminary data cleaning and data-interpretation necessary for data analysis, the fact that the administrative data have their own structure, terminology and purpose-driven content, should be taken into consideration. That is, the administrative data are based on the public administrations' registers' own logic system. Regarding the research question in connection with the relevant data themes, it is certainly necessary to take the mentioned information into consideration before research-focused data analysis and data processing.

In the following, the characteristics of the datasets based on administrative data will be presented by fitting them into the traditional research terminology. The validity of the harmonized dataset shows: to what extent are the aggregated variables able to reflect those questions evolved through the research concept. That is to say, how possible is it to define or interpret the subject of analysis by the given data. The degree of validity can be elevated by the harmonization and cleaning process, by fixing the interpretational frames correctly, and by handling the data according to those frames. However, such data content that would answer all research questions cannot be fully accessed. The data content in terms of each research question has quite a high confidence level, it barely changes after the sampling is repeated. However, due to the internal dynamism and fluctuation of the datasets, minimal differences can be observed here as well. When using administrative data sources, the conceptualization and the operationalization – in opposition with those traditional surveys that are typically initiated from scientific question-making – are executed on a parallel and iterative way, plus by making multiple steps. The final realisation, that is, the actual calculation process of the generated variables can be recorded afterwards, usually after a long experimentation.

After the data-linking, the CERS Databank creates a long time series panel dataset, which is produced by harmonizing hundreds of raw fields of the latest linking. The professional data cleansing – embracing more than ten years – takes a long time. That is why the scientific analysis of the current linked dataset will be possible only in 2020-2021 at the earliest.

Data cleaning

The data cleaning and harmonization work are particular and are adjusted to the characteristics of the administrative data. The data of the linked panel datasets are partly structured, are statical concerning the dimension of time, however are appropriate for retrospective longitudinal research. The observable unit is the change of administrative statuses, thus instead of opinions, concrete behaviours can be examined.

Because of the characteristics of administrative data detailed above, both longitudinal and cross-sectional consistency tests are part of the data cleansing process. In cross-sectional consistency testing, we compare the contents of different fields at a given time. The occurring anomalies during this process reveal the limits of the data.

Given the administrative nature of the data, the examination of the data environment in a broader sense is inevitable concerning the whole research time period – if the research target is to follow processes, changes over time –, in a retrospective way. (Veroszta, 2015). In parallel with this, we observe the contents and variations of the yearly changing data blocks by using longitudinal checking. Besides that, in order to interpret the data it is indispensable to update the dictionary of codes (meta database) related to the whole observational period. After the above mentioned tests as part of data cleansing, the data content corresponding to the research questions can be obtained upon analyst decisions, from those cells of solely administrative meaning (framed by the purpose and context of the given data collection).

The data cleansing executed by the Databank begins with reviewing the received data, the work of data providers and the NISC Ltd. The supervision of data transfer is also performed by taking into account the characteristics of administrative data detailed above. After checking the received data, variables corresponding to the different research questions are generated while transforming the raw data fields on an iterative way, first by each data source.

After the cognition, cleaning and harmonization of the data by each data source, the set-up of the enormous dataset begins by linking the shortlisted but most important variables. The resulting database, the Admin3 contains monthly data of individuals' statuses between 2003 and 2017. The longitudinal (temporal) and cross-sectional (based on the comparison of different data hosts) wave of harmonization, supervision and cleaning also gets done in this phase. In addition the variables derived from multiple data sources are generated at this point. The inconsistencies of the database revealing after the linking is made are managed.

A largely consistent, enormous database is hereby set up. Applying the more complicated variables, complex tests or mini surveys are commenced using the database in order to

reveal potential defaults. The occurring issues and solutions are also built into the revised database.
The collective phase of the harmonization comes next, in which the experts of various fields of the scientific community are given an opportunity to use the Linked Administrative Panel dataset, together with adding their previous experiences into the cleaning process so that it can be improved. The collected information (like written codes) in the phase of collective harmonization are built into the first-round cleaning of the forthcoming wave of Admin, in an organic way, continuously improving it. In a similar way, every issue, question and feedback occurring when later using the data – as well as the written codes – is built into the cleaning program of the Admin files.

Access to the data

The NISC Ltd. anonymizes the data during data-integration, thus they are no longer possible for follow-up identification. It also publishes the raw data content of the created databases. The relating annex of the contract contains the partners list and variables describing the raw content of datalinkings. To which the NISC Ltd. ensures access if requested, based on the above mentioned law of 2007.
The large database containing typically monthly data of individuals' statuses - cleaned in line with the above detailed aspects - is solely open for the researchers of the CERS in the phase of collective data cleaning, who then participate in the process of expertised data cleaning.
After the expertised data cleaning, access to the Linked Administrative Panel datasets is given for doctoral dissertations or thesis in every case, as well as in the case of appropriate affiliation and research purpose. Currently the Admin1 (2002-2008) and Admin2 (2003-2011) are available to access. Access to the databases is made possible through a safe server-and STATA based software, which can be used until a previously recorded deadline marked on the data request form.

Relevance and significance of the research

The Admin databases are uniquely rich in Central and Eastern Europe and are widely used by national and international researchers. They are particularly suitable for producing scientific results and publications in high-quality international journals, as they provide opportunities for complex longitudinal and cross-sectional analysis across a wide range of disciplines.

The currently available waves of the Linked Administrative Panel datasets (Admin1 and Admin2) have been used in more than 80 research projects, by nearly fifty researchers. There are existent and currently conducted Admin-based research projects and international scientific publications in the fields of health science, health policy, regional studies, labour economics, firm research, migration research, agronomics, and social policy. Besides national journals, the set up scientific products are also published in international scientific papers, such as in the *American Economic Journal* (Lindner & Reizer, 2019), the *Quarterly Journal Of Economics* (DellaVigna & Lindner & Reizer & Schmieder, 2017) the *Health Economics* (Bíró & Elek, 2018), the *IZA Journal of European Labor Studies* (Czafit & Köllő, 2015), the *Scandinavian Journal of Public Health* (Scharle & Adamecz-Völgyi & Lévay & Bördős, 2018) and the *Research in Labor Economics* (Csillag, 2019) book series.

The latest wave of the Linked Administrative Panel datasets, the Admin3 contains larger time series and more punctual data, as the data content of the public administrative registers are continuously improving. The degree of this is primarily relevant in the field of education, particularly in the case of the National Assessment of Basic Competencies, as it still has a short history.

Contact

On a national level, 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 STATA workshops and maintains a research labor force with its own computers. For general information please use 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 https://net.jogtar.hu/jogszabaly?docid=A0700335.KOR

Law of Act CI. of 2007 on ensuring access to data required for decision preparation https://net.jogtar.hu/jogszabaly?docid=a0700101.tv

Adamecz-Völgyi, A. & Bördős, K. & Lévay, P. & Scharle, Á. (2018). Impact of a personalised active labour market programme for persons with disabilities. Scandinavian Journal of Public Health 2018; 46 (Suppl 19): 32–48

Bíró, A. & Elek, P. (2018). How does retirement affect healthcare expenditures? Evidence from a change in the retirement age. Health Economics 27(5), 803-818.

Csillag, M. (2019). The Incentive Effects of Sickness Absence Compensation – Analysis of a Natural Experiment in Eastern Europe. Research in Labor Economics, Vol. 47 (Health and Labor Markets), Emerald Publishing Limited, pp. 195-225. https://doi.org/10.1108/S0147-912120190000047007

Czafit, B. & Köllő, J. (2015). Employment and wages before and after incarceration - evidence from Hungary. IZA Journal Of European Labor Studies 4 pp. 1-21., 21 p.

DellaVigna, S. & Lindner, A. & Reizer, B. & Schmieder, J. F. (2017). Reference-dependent jobsearch: evidence from Hungary. Quarterly Journal Of Economics 132 : 4 pp. 1969-2018. , 50 p.

Lindner, A. & Reizer, B. (2019). Front loading the unemployment benefit: an empirical assessment. American Economic Journal: Applied Economics – *Due to be published*

Veroszta, Zs. (2015). Management of administrative data in social research. Educatio, Vol. 23., No. 3. http://ofi.hu/sites/default/files/attachments/educatio_2015-3_web_0.pdf

---

[1] 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 our case, the procedure serves the generation of anonym, technical identifiers.