2. ADMINISTRATIVE DATA IN ECONOMIC RESEARCH

2.1. THE BENEFITS OF LINKED ADMINISTRATIVE PANEL DATA AN EXAMPLE OF THE KRTK DATABANK ADMIN-DATABASES

JÁNOS KÖLLŐ

On 25 July 2007, the "Act CI. of 2007 on Ensuring the Availability of Data for Decision-Making" entered into force, opening a new era in the field of empirical social studies in Hungary, alongside governmental decision-making and impact assessment. The act was initiated by the Economic Research Department of the Ministry of Finance, recognising that by linking public administration registers, longitudinal databases of unprecedented richness and size could be built, but also noting that ministries and authorities were often reluctant to share these data with governmental partner institutions. The Act CI. *obliged* governmental bodies to share data and established a procedure to link registers and data for different years within each register, while respecting the protection of personal data. On the road from the first attempts to the law, see the paper by the then head of the research department, the initiator of the law (*Scharle,* 2019)!
In the decade since then, several public administration bodies have carried out professionally sophisticated analyses with linked data: first the Ministry of Finance itself (see for an example *Elek et al.,* 2009), later the Educational Authority, which now uses linked panel data as the basis for the Diploma Career Tracking System (see *OH,* 2021), which was previously based on survey, or the Hungarian National Bank (*Borsos–Stancsics,* 2020).

Data linkings initiated by the HAS (Hungarian Academy of Sciences)
The implementing decree of the Act (Government Decree 335/2007 (XII. 13.)) listed the President of the Hungarian Academy of Sciences among those entitled to request data. Taking advantage of this opportunity, the Databank of the Centre for Economic and Regional Studies (KRTK Databank), formerly under the HAS, initiated data linking on four occasions[1].

As *Table 2.1.1* shows, the Admin2–4 databases now link the registers of not only three – as previously –, but five datahosts, and track 50% of the population as of January 2003 on a monthly basis until the end of 2011, 2017 and 2021. Moreover, Admin4 also covers the population entering the registers after 2003. In addition, the databases observe employers for whom at least one observed person has worked at least once during the whole period, on an annual basis. The main data sets in the Admin3 and Admin4 are the following:

[1] However, the HAS does not have a role of full right: while public administration bodies can link entire registers, the HAS can take a maximum of 50 percent random sample from the population

– occupational status, job title, paid working days per month, pension-entitled earnings; date of entering and leaving the employer,
– the employer's characteristics (balance sheet data, indicators relating to the workforce composition),
– registered unemployed, benefits, participation in programmes, job mediation (from 2009),
– pensions, child benefits, welfare benefits,
– public work participation (from 2011),
– participation in education, complete from 2009, with back data before; educational attainment of young people; test scores and background questionnaire variables from the National Assessment of Basic Competences (from 2008),
– health status variables: out-patient and in-patient care, change of medication, sick-pay, codes for type of illness (from 2009),
– permanent residence, at district level.
On the data process and harmonisation, anonymisation, pre-testing and continuous improvement of ready-to-use databases, and ensuring the protection of personal data, see Anna Sebők's contribution to the volume in *subsection 2.2*.[2]

2.1.1. Table 3: Linked public administration panel databases in the KRTK Databank

	Observations		Period	Primarly data holders
	Individuals	Employers	(monthly data)	
Admin1	4 602 000	-	2002-2009	OEP, ONYF, VAT
Admin2	4 601 999	886 425	2003-2011	OEP, ONYF, NAV, VAT, OH
Admin3	5 174 486	1 126 343	2003-2017	OEP, ONYF, NAV, NMH, OH
Admin4[a]	6 393 158	1 537 329	2003-2021	NEAK, ONYF, NAV, NMH, OH

[a] *The Admin4 database is in the making, the case number of employers is not yet verified. For details, see* Databank.
OEP: National Health Insurance Fund, ONYF: Central Administration of Pension Insurance. ÁFS\: National Employment Service, NAV: National Tax and Customs Administration.
NMH: National Labour Office, OH: Educational Authority, NEAK: National Health Insurance Fund.

What makes linked administrative panel data different from the traditionally available data sources?
It is not difficult to see how a database of this structure and size represents a radical change from the data sources most commonly used in academic research and policy analysis in the 1980s and 1990s.

[2] *For more information on the KRTK database and the files created so far, see adatbank.krtk.mta.hu.*

The KSH *Labour Force Survey* (LFS) is an extremely valuable database for the study of labour market processes, with its quarterly waves providing information on families, educational participation, employment, unemployment, inactivity, transfers and much more since 1992. Participants in the register can be tracked for six quarters. However, the LFS does not include data on wages, says very little about employers (sector, size, geography), and the follow-up period is one and a half years, pointwise, with quarterly observations, unlike the Admin datasets, where observations are monthly (or even daily) and have been following the sample participants for nearly twenty years, providing detailed data on both their current employers and employees. The sample size of the LFS is less than one per cent of that of the Admin databases, unsuitable for studying small groups.

The NMH and, from 2019, the HCSO's Wage Survey **(**WS) **is** also a source of particular importance, which has collected data in 1986, 1989 and every year since 1992, in essentially the same structure, initially on companies with 20 employees, from 1995 onwards on companies with 10 employees and from 2000 onwards companies with 5 or more employees, budgetary institutions and larger non-profit organisations. (Individual data are complete for firms with fewer than 20 employees and for government agencies. For larger employers individual data is available on the basis of a roughly 10 percent sample.) Each wave includes 150,000 to 170,000 observations at the business sector level, nearly 700,000 observations at the individual level and more than 20,000 observations at the employer level for the government and nonprofit sectors, until 2019 on settlement level. At its launch, the Wage Scoreboard was one of the first linked employer-employee data (LEED**)** databases worldwide and, complemented by company balance sheets, opened up unprecedented opportunities to study employer behaviour and earnings inequality. However, the data are not longitudinal, companies can be followed but not individuals.

Company balance sheets and their compilations (Amadeus, Opten, Orbis) are another important source of data for economic research, in which companies can be followed over a longer period, even if it is a perfectly impossible task to filter out companies splitting and companies merging. However, we do not look deeper than the firm level, having information only about financial variables, number of employees and average wages, which severely limits the range of questions that can be researched.

Databases on education, health and demography provide a detailed picture of the areas concerned, but the analyst cannot see through: known information is about what school someone went to, what test results they achieved or what kind of illness they have, but it is not clear what preceded these conditions and what followed by.

The Unemployment Register also provides good quality and detailed data on the activities of the employment offices, but does not give a picture of the path to unemployment and getting out of it.

Admin databases "know" almost everything that the above types offer, and much more than the research that uses them can dream of. The 'almost' part is justified by two circumstances: the educational attainment is not known except for young people and the unemployed, (as opposed to LFS and WS), and no information about who lives in a household together (as opposed to LFS).

The former can be roughly estimated from the jobs occupied, however, the latter cannot be recovered.

Scientific yield

The longitudinal databases created by linking public administration registers have significantly contributed to the unfolding of the 'credibility revolution', which has previously placed difficult, often unachievable demands on empirical research seeking to verify causal interactions (Angrist-Pischke, 2010). With such data, the new standards are easier to meet: it is possible to study variation within observation units instead of (or next to) levels, and to filter out the effect of unobserved heterogeneity, even by using multiple fixed effects simultaneously. It is easier to select an appropriate control group from large samples rich in variables.

Public administration data not only allows deeper analysis, but have also expanded the scope of questions that can be researched. At the same time, we can study characteristics that could previously only be studied separately, such as employment, earnings, job and occupation change, unemployment, welfare benefits, health status, educational participation and school achievement. We can see through the factory fence or the gates of the institutions at such a level that was unimaginable in previously available databases. Inactive people of working age, who are not included in any government register, now can be identified in the linked panel data: by excluding students, the employed, the unemployed (and, if necessary, the retired and child benefit recipients).

Perhaps what is most important: the extraordinary size of the databases makes it possible to monitor a wide range of critical events and small groups: people who have left school at the same time, became unemployed or found a job, people who became ill at the same time; companies that have expanded or reduced output or have reduced the number of the staff at the same time, and so on. Beyond the edges of the window of time, in every moment, we know the pre-history and the subsequent story of the people and firms observed. The case numbers are considerable even for such small groups as conscripts, early retired police officers or prisoners.

Research with Admin1-3 databases

The wordcloud in Figure 2.1.1 used those research that were based on Admin data and which have already resulted in publications or at least appeared in workshop studies. The cloud was made out of the words in the title of the publications, the size of the headings is proportional to their frequency of occurrence. For a detailed list of the 48 publications, see bpdata.eu, where you can also access the content summaries of each paper (if available) by clicking on the link in the bibliographic section. The list in Figure 2.1.1, which refers only to the subject of the research illustrates well the wide range of analysis possibilities offered by the linked administrative panel data.

Figure 2.1.1: Words in titles of publications based on KRTK Admin databases

Note: Most of the titles are originally in English. Only the words referring to the subject and method have been left and the conjunctions have been omitted. For a list of titles, see bpdata.eu.

The word "impact" is the most frequent word in the titles, clearly indicating that large, linked panels are well suited to conduct methodologically established impact analysis. Among the areas of research, the labour market (and related employment, activity, labour demand, work ability, substitutability), health (healthcare, doctors, hospital, antidepressants, caesarean,, mortality, morbidity, life expectancy, deaths, rehabilitation), earnings (inequality, young people, the elderly), pensions and corporate behaviour (industries, managers, productivity, value chains, property), but there are also studies on education, police and prison services, personal and business relations, emigration and public subsidies.

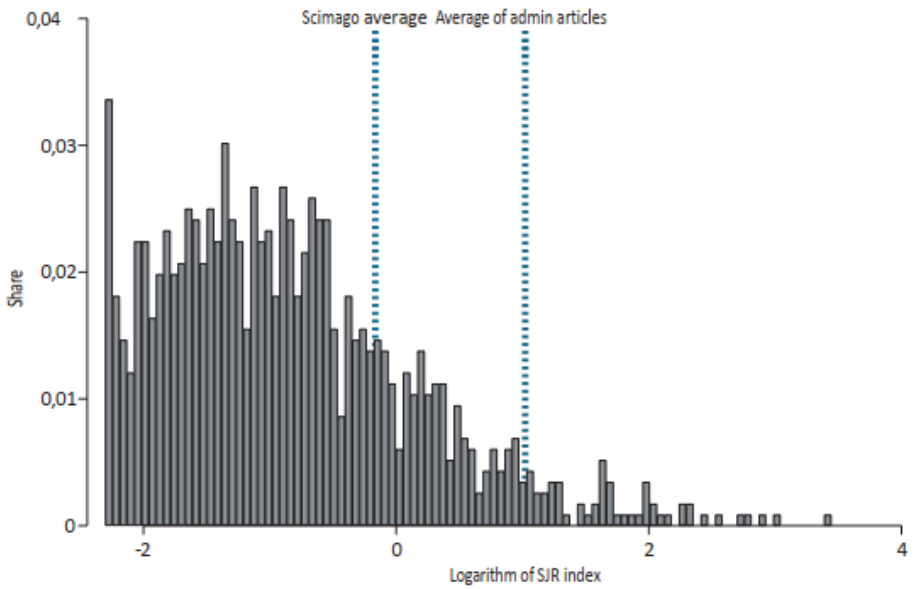
The world-class Admin3 database will be available until 2021 (it is still being tested and improved as I write this), so very little of the research on it is ready for publication in the journal. So far, 21 articles based on the Admin panels have been published in Scimago listed journals, in places with much higher than average citation. There are two statistics to mention. The IS (article influence score) indicator for Admin articles is 3.7, which is much higher than the average (IS = 1.0), suggesting that the scientific impact of the journals hosting them, measured by citation, is much stronger than average.[3] The other indicator chosen is the SJR (Scimago Journal Rank), which has an average value of 0.85 with a standard deviation of 1.85, among the 1160 journals in the field of economics, econometrics and finance measured in Scimago. For Admin articles, the SJR is 2.78, more than one standard deviation unit higher than the average.[4]

[3] For IS statistics, see cornell.edu.

[4] The SJR indicator measures the scientific impact of journals, which is both the impact of journals, both in terms of the number of citations received by the journal and the importance or prestige of the journals from which the citation come. Its logic is similar to the Live Score used in chess.

The distribution of Admin articles in the field is clearly shown in the histogram in *figure 2.1.2*, where the horizontal axis shows the SJR values (in natural logarithm so that the extreme cases do not overwhelm the histogram), while the columns show the field share of journals with a given SJR value. Of the two broken lines, the left one shows the Scimago average and the right one shows the average of the Admin articles. It can be seen that the distribution of SJR values is highly asymmetric, which is the reason why only five percent of the journals listed in Scimago achieved an SJR index above the average of Admin publications.

2.1.2. Figure 3: Average SJR of journals with publications based on Admin data of 1160 journals in the field of economics, econometrics and finance listed in Scimago, histogram



The science metrics indicators will certainly improve in the future thanks to the richness of Admin3 and Admin4. The unique data background attracts many foreign researchers. The authors of these publications and prepublications include 34 foreign researchers and 25 Hungarian researchers working abroad. The latter suggests that the rich data background helps to maintain the relation of Hungarian researchers working abroad with their home country.

Costs

Linked administrative panel costs a lot: on top of the 15-35 million forints paid to the data hosts and the coordinator responsible for the linking, that is also at most what it takes to build a research-ready database over several years, and that amount is without taking into account the cost of the machine capacity and the sofware needed to handle the extraordinary size of data. But if we consider how much it would cost to produce the data for the studies in our list using questionnaire methods, even on samples of a tenth of the size, we can say the administrative panel is cheap. At the time of writing, 64 research projects are using the Admin3 database. If we assume a cost of 100 million, this means a data cost of one and a half million forints per project – the cost of a few focus group discussions.

Practical information

The spell-structured source files underlying the Admin databases qualify as public data, accessible to anyone for scientific purposes.[5] The versions prepared for research by the KRTK Databank can be analysed on a secure server with dedicated remote access, while more detailed health data can be researched in a strictly controlled dataroom. Institutional users can access the databases for a fee, depending on the cost of server hosting and any additional server capacity required. The data may be used free of charge by individual researchers working in accredited research centres and their thesis writing students for scientific purposes only, subject to the limits of available computer capacity. Priority is given to KRTK researchers and projects lead by them. The data is stored in Stata format, but is also available for R and Python programs. User support is provided for Stata users. We do not release the data for reasons of size (the Admin3 database including the additional modules is 120 gigabytes) and data protection. No research can be published in a serious place if data are not available. To ensure this, the estimation samples will be made available - with the minimum data content required - to lecturers, or researchers wishing to replicate the results, while preserving anonymity if requested.

The future of administrative panel databases

In the summer of 2021, Act XCI of 2021 on National Data Assets repealed Act CI of 2007. Administrative data management was taken over by the then established National Data Asset Management Agency (www. navu.hu). Neither the law nor the related implementing decree[6] addresses the issue of access to data for scientific purposes. The decree does not specify the institutions (persons) entitled to initiate data linkage, nor does it describe the procedure itself in detail.

It is very much hoped that the advantage gained by Hungarian social science research through the affiliated administrative panels will not be lost: the high quality of the data, which is also outstanding in global terms, will continue to attract Hungarian and international research community, helping to secure international funding and laying the foundations for major scientific achievements. At the same time, we hope that the methodological knowledge accumulated over the last ten years by the experts who built the databases and the researchers who use them, as well as the hundreds of million forints of computing capacity that the KRTK, the MTA and later the ELKH have built up specifically to manage large panel databases, will not be wasted. We are also confident that this knowledge will be used to a greater extent than today by public administrations in data management, evidence-based decision making and professional impact assessments.

[5] The observation unit of a source file consists of a start and an end date, and a variable indicating what happens in between: "spell". From that set of data the KRTK Databank creates datamatrix in fix format consisting of individual-month observations, with more than a billion rows and hundreds of columns.

[6] See Government Decree 607/2021 (XI. 5.) on certain detailed rules for the use of national data.

References

Angrist, J. D.–Pischke, J-S. (2010): The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. Journal of Economic Perspectives, Vol. 24. No. 2. pp. 3–30.
Borsos András–Stancsics, M. (2020): Unfolding the Hidden Structure of the Hungarian Multi-Layer Firm Network. Occasional Papers, No. 139. Hungarian National Bank, Budapest.
Péter Elek–Ágota Scharle –Bálint Szabó – Péter András Szabó (2009): Wage-related tax devasion in Hungary. Public Finance Papers, No. 23. ELTE TáTK, april.
OH (2021): Diploma tracking system 2020 – merging administrative databases. Quick report. Educational Authority, Budapest.
Ágota Scharle (2019). Hungary: a case study on improving access to administrative data in a low-trust environment. Published in: Crato, N–Paruolo, P. (editors): DataDriven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design. Springer, Cham, pp. 119–130.