UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Creating ready-made research datasets from national administrative registers.

Päivi Kankaanranta and Aino Melakari (Statistics Finland)

paivi.kankaanranta@stat.fi; aino.melakari@stat.fi

Abstract

In order to improve access to national registers and administrative data for research purposes, Statistics Finland produces ready-made datasets that are less labourious to disseminate than datasets tailored to researchers' requests. In contrast to sample surveys made readily available to researchers by many national statistical institutes as well as Eurostat, these ready-made datasets contain a standard set of variables with information at the level of individual business, household or person extracted from the underlying administrative register. The use and production of the ready-made datasets are regulated by the Finnish Statistics Act, and the EU statistical legal acts. Hence the data are pseudonymized so that all the direct identifiers have been excluded from the data. To be able to provide micro data with highly detailed information, the ready-made datasets are divided into smaller modules that include a limited number of variables related to specific topics. Researchers use only modules relevant for the subject area of their research, which allows Statistics Finland to fulfill the GDPR requirements for data minimization. All the ready-made data modules can be linked to each other and to micro data from other sources with individual pseudo-identifiers to build rich datasets. The set of variables included in each module is described in the metadata catalogue on Statistics Finland website in a way that allows researchers to identify the data files suitable for their purposes. The ready-made datasets can be used only via secure remote access system, and each output file is checked for confidentiality. The ready-made datasets are one of the basic functions of the research services of Statistics Finland. Almost 70 percent of all research projects are using ready-made datasets. There is a separate team, who maintains these datasets on a full-time basis. Together the ready-made datasets and the remote access system ensure that the use of unit-level register-based data for research purposes complies with statistical confidentiality and data protection policies.

Creating ready-made research datasets from national administrative registers

Päivi Kankaanranta[*] and Aino Melakari[**]

[*] Statistics Finland, paivi.kankaanranta@stat.fi
[**] Statistics Finland, aino.melakari@stat.fi

Abstract: In order to improve access to national registers and administrative data for research purposes, Statistics Finland produces ready-made research data modules that are less laborious to disseminate than datasets tailored to researchers' requests. In contrast to sample surveys made readily available to researchers by many national statistical institutes as well as Eurostat, these ready-made modules contain a standard set of variables with information at the level of individual business, household or person extracted from the underlying administrative register. The use and production of the ready-made datasets are regulated by the Finnish Statistics Act and the EU statistical legal acts. Hence, the data are pseudonymized so that all direct identifiers are removed from the data. In order to be able to provide microdata with highly detailed information, the ready-made modules include a limited number of variables related to specific topics. Researchers use only modules relevant for the subject area of their research, which allows Statistics Finland to accommodate the GDPR requirements for data minimization. All the ready-made data modules can be linked to each other and to micro data from other sources with individual pseudo-identifiers to build rich datasets. The set of variables included in each module is described in the metadata catalogue on Statistics Finland website in a way that allows researchers to identify the data files suitable for their purposes. The ready-made data modules can be used only via secure remote access system, and each output file is checked for confidentiality. The ready-made data modules form one of the basic functions of the Research Services at Statistics Finland. Almost 70 percent of all research projects are using ready-made datasets. There is a separate team, who maintains these datasets on a full-time basis. Together the ready-made data modules and the remote access system ensure that the use of unit-level register-based data for research purposes complies with statistical confidentiality and data protection policies.

1 Introduction

One of the strategic goals of Statistics Finland is to improve access to national registers and other administrative data sources for scientific research and statistical purposes. Administrative records are used extensively to produce official statistics in Finland (see Statistics Finland 2004). Therefore, the majority of data reserves at Statistics Finland consist of register-based data that is of interest to researchers too. Since tailoring research datasets from in-house reserves and external sources according to researchers' requests is laborious, the Research Services at Statistics Finland releases ready-made data modules. These ready-made data modules are made available to researchers on an "as is" basis: they contain a standard set of variables with unit-level information on all enterprises, households and persons extracted from administrative records that a wide range of authorities maintain. Typically, national statistical institutes and other data

providers, such as Eurostat, offer access to microdata that include information on a sample of survey respondents instead of the whole target population.

Ready-made research data modules are time- and cost-effective to produce. A single file folder stored in a secure environment is enough to satisfy the needs of all users of a specific module. The data comprising a ready-made module is compiled only once and saved to a dedicated folder which is made accessible to authorised users. Hence, there is no need to replicate data files each time information in a specific module is requested or updated.

Statistics Finland was inspired to develop a modular design that provides researchers with an easy yet flexible access to confidential microdata by its experiences with the Finnish Longitudinal Employer-Employee Data (FLEED) and a variety of business data modules. This dataset that spanned over 28 years (from 1988 to 2016) included all persons aged 15 to 70 residing in continental Finland. Over 160 register-based variables in the widely used FLEED dataset contained information on person's basic characteristics, family, living conditions, employment, relationships, periods of unemployment, income and education. In response to national and EU level demands for data protection, this large dataset was later divided into smaller subsets by topic. Researchers were thus prevented from gaining access to information unnecessary for their specific purposes. The subsets of the former FLEED dataset are now offered as FOLK ready-made research data modules.

2 Ready-made research data modules at Statistics Finland

2.1 Basic elements of ready-made research data modules

Statistics Finland designs ready-made research data modules on the basis of researchers' needs. Most frequently requested administrative records on demographics, society and businesses are turned into ready-made data modules. Direct feedback from the research community is paramount in identifying the data that correspond to the researchers' current and future needs. A few researchers working primarily in other organisations hold a part-time position at the Research Services at Statistics Finland. Together with Statistics Finland's own data experts and scientists, these researchers contribute their subject-specific expertise to the development of new ready-made research data modules. They also serve as a bridge between Statistics Finland as a data producer and the research community as a data user. In addition, Statistics Finland explores researchers' preferences by conducting customer surveys and analysing trends in tailor-made research datasets. Lately, researchers have been increasingly asking for access to as raw data as possible and as quickly as possible after its release.

Each ready-made module released by Statistics Finland belongs to a thematic collection and is centred around a specific topic within the theme. The general themes include population, education, business activity, income, labour market, international trade,

transport, housing as well as wages and salaries. In order to control users' access to confidential microdata, each module consists of a standard set of variables whose number is limited. Even though the number of variables is a subset of the source data, all of the original statistical units are contained in ready-made data modules. Hence, the modules represent the entire population in the administrative source.

All ready-made data modules have a name consisting of a prefix and a label. The prefix refers to the theme and the label describes the topic of the module in more detail. These module names are widely used in the research community. Currently, Statistics Finland has some 56 ready-made research data modules on offer in total. Table 1 provides a summary of these modules.

Prefix of the collection	**Theme of the module collection**	**Selected topics of the modules (Data Source)**	**Target population**	**Number of modules**
FOLK	Population	Income Education Family Degrees and qualifications Cohabitation	Person	12
EDUC	Education	Admissions Students Degrees and qualifications	Person	3
FIRM	Business activity	Business operations Financial statements	Enterprise	23
FLOWN	Ownership	Shareholder information Dividend information	Enterprise, person	1
SES	Earnings	Structure of earnings statistics Harmonised structure of earnings statistics	Person	1
TAX	Income	Wages, pensions and benefits (National Incomes Register)	Person, enterprise	3
INFRA	Spatial data	Spatial data of residents	Person, building	2
Data from other organisations	Labour market International trade Education Transport	Employment service statistics (Ministry of Economic Affairs and Employment) International trade on commodities (Finnish Customs) Matriculation examination results (Matriculation Examination Board)	Person, enterprise	11

Table 1.1 Statistics Finland's currently available research ready-made data modules.

The metadata catalogue on Statistics Finland website, known as Taika, consists of comprehensive documentation that describes the ready-made data modules and the variables in them (see Statistics Finland, Taika – research data catalogue). The documentation is available in Finnish and in some cases in English, at least in abbreviated format. The data and variable descriptions can be browsed separately and queried for search terms and/or certain attributes. Data documentation can thus be tailored to one's needs and downloaded in a file format of one's choice. The metadata catalogue is useful not only for current users of ready-made modules but also for future ones. The catalogue enables researchers to familiarize themselves with ready-made modules and identify the ones suited for their purposes before they submit the application for licence to use statistical data.

2.2 Linking of ready-made data modules

Ready-made research data modules can be used to build rich datasets that correspond to the specific needs of a research project. Different authorities in Finland use unified identification codes for persons, businesses and organisations as well as buildings and dwellings. Ready-made data modules can therefore be linked to each other with pseudo-identifiers derived from these direct identification codes. This type of data linking is not feasible with survey-based microdata except for longitudinal studies. Once the data is harmonised, unified identification codes allow individuals and businesses to be tracked over their life span. The data harmonisation typically consists of naming variables consistently over time and making most-commonly used classifications comparable.

Unified identification codes make it possible to combine ready-made modules even with unit-level data from sources beyond the control of Statistics Finland, such as data on the use of healthcare services or prescription drug purchases. These data are sent to Statistics Finland for pseudonymization before they are made available to researchers. Survey data collected either by Statistics Finland or researchers themselves can be integrated with unit-level register-data to a limited degree. Survey respondents need to be informed before the interview which administrative records will be combined with the information they give for the survey (see e.g. Törmälehto 2008).

2.3 Application to use ready-made data modules

Ready-made data modules released by Statistics Finland provide researchers with a relatively quick access to confidential microdata. Since these modules are offered in a standard format and the data files are readily available in a secure environment, it takes less time for a licence application to be processed compared to tailor-made research data. According to current practice, the licence covers all future updates to ready-made data modules specified in it for as long as the licence is valid (maximum five years with an option to renew). Hence, researchers do not need to separately apply for a new licence to gain access to the latest release to data modules at their disposal. However, a new licence application is required whenever new ready-made modules or other data are added to an existing research project.

Because ready-made data modules contain information on the total sample of the target population, licence applications and the accompanying research plans are screened carefully. The application must specify requested ready-made data modules and other data along with the purpose of the research project. When access to microdata containing information on persons is requested, the application must provide a justification for the need to use information on the total population. Qualifying reasons include for example following families over time or producing control groups for the study population. If the use of the total data is not justified, a sample corresponding to the demands of the research design will be offered instead.

3 The data protection and data security measures

3.1 General preconditions for releasing ready-made data

Ready-made research data modules provide a compromise between easy access to confidential microdata and compliance with the legal framework on the processing of unit-level information. The conditions for the production and use of the ready-made data modules are defined in the General Data Protection Regulation (GDPR) and the supplementary national Data Protection Act (1050/2018) as well as in the national Statistics Act (280/2004). These legal arrangements define what highly detailed unit-level information contained in administrative registers can be released and how this microdata can be accessed (see Statistics Finland 2013 and UNECE 2007 for previous descriptions of national legislation on the release of microdata in Finland).

Since ready-made datasets consist of standard sets of variables that are made available as such without tailoring, the information content of each module is reviewed thoroughly in the planning phase to prevent researchers gaining access to unnecessary information. The proposal for a finalized ready-made data module is submitted for evaluation to Statistics Finland's statistical ethics committee, which consists of representatives of various departments. The final decision on the release of a new ready-made research data module is made by the directors in charge.

In order to meet the data security and data protection requirements, direct identifiers are excluded from ready-made research data modules. Unique identification codes, such as personal identity codes, business identity codes as well as domicile codes for buildings and dwellings, are replaced with pseudo-identifiers. Because ready-made data modules contain a limited number of variables, a final dataset linked together with these pseudo-identifiers complies with the GDPR demand to minimize data: researchers only use modules relevant for their project. Although pseudonymization ensures that data subjects of the modules cannot be identified directly, they can still be identified indirectly. Hence, the ready-made data can only be used via a remote access system, where processing of data can be controlled.

To further protect the privacy of data subjects, sensitive information is removed from ready-made data modules. Like exceptional information, personal data belonging to certain special categories is excluded from these modules as well. The processing of information that belongs to these categories is generally prohibited. Such data reveals racial and ethnic origin, political opinions, religion or philosophical beliefs, trade union membership as well as sexual orientation or activity. Special categories of personal data also cover information concerning to health as well as genetic and biometric data that directly identifies the person. Statistics Finland does not maintain records of these data. However, health as well as genetic and biometric data are available via other authorities and can be processed and linked to Statistics Finland's ready-made datasets if GDPR, European law or national legislation provides an exception to the prohibition.

3.2 Secure solution for using the ready-made data

Because data subjects can indirectly be identified from the confidential microdata, ready-made modules are only accessible via the remote access system FIONA that Statistics Finland has developed together with the state-owned company IT Center for Science. This system offers a secure way to handle sensitive data. FIONA is a closed environment without external internet connection, and researchers have no possibility to take out or upload any data or any files into the system. It is possible to use FIONA only within the EU region and the third countries to which EU commission has allowed to transfer microdata, and from the organisation-level IP-addresses which Statistics Finland has accepted as secure ones. In situations where it is impossible to get access to FIONA, researchers can visit the research laboratory at Statistics Finland's facilities in Helsinki.

Before researchers can get access to the remote access system and the ready-made data, they must fill in a remote access commitment with which they commit to the terms and conditions of the Research Services of Statistics Finland. In the commitment it is mandatory to report for example the addresses and describe the data security measures of the workspaces. The commitment is signed by a representative from the researcher's organisation, who takes responsibility for researcher's remote access use. In case of misuse, Statistics Finland is allowed to disconnect entry into the remote access system.

3.3 User's guidelines for protecting confidential data

Researchers using pseudonymized data are subject to obligation of secrecy and must pledge not to disclose to other persons than mentioned in the license any information that they are legally obliged to keep secret. Researchers must also make sure, that outputs do not disclose unit-level information or contain information that enables a data subject to be identified.

There are several guidelines for different types of datasets and outputs. The main rule in protecting frequency and magnitude tables is a threshold value of three, which means that the output can be published only if it is based on at least three observations (with a few exceptions, such as the threshold value of ten for the ready-made data based on

Incomes Register). There are also other rules in addition to this threshold value, such as the dominance rule (1.75), which is used with the most up-to-date enterprise data. Regarding the most sensitive datasets, there are more specific data protection rules to be considered. For instance, certain threshold values apply to the outputs published from the spatial data. When the ready-made module includes data authorised by other authority, Statistics Finland follows the data protection rules of that organisation.

Researchers are required to follow data protection requirements while working in the remote access system. All data transfer between FIONA and researchers goes through an output checking protocol, which is done manually by the personnel of the Research Services. Before the outputs can be sent to the researchers, they are checked for possible breaches of confidentiality. Since the disclosure review engages Statistics Finland's resources and results in delays to researchers, options to partially automate the protocol are currently explored.

4 Production and user costs of ready-made research data

Statistics Finland charges a fee for both processing the application for licence to use statistical data and the research data itself. Additional charges are collected for the use of the remote access system on an annual basis. The processing of the licence application has a fixed rate and includes one hour of preparatory work. Additional hours of preparatory work are charged at an hourly rate. The processing fee is collected even if the application is not approved.

Modular design is a time- and cost-effective way to provide researchers with access to microdata. This efficiency is reflected in the relatively inexpensive prices charged for ready-made data modules. Because a ready-made data module is released only once in a standard format, the production and maintenance costs are distributed equally among all authorised users of the module, both current and future ones. The costs of compiling a tailor-made dataset, on the other hand, are solely incurred by the research project in need of the customised data.

The production method also allows Statistics Finland to fix the prices of ready-made data modules in advance which contrasts with tailor-made research datasets. Therefore, the data costs for a research project relying mainly on ready-made data modules are fairly predictable. At present, data fees are collected only once, after the user licence has been granted, and include future updates to data. Furthermore, data fees cover guidance on the use of the ready-made research data and any additional data maintenance. Currently, there are four different price classes that are roughly based on the size of the data module and the frequency of updates. The costs of output checking, on the other hand, are covered by the annual fees charged for the use of remote access system. The fees for processing of the licence application, the data and the use of remote access system are subject to periodic changes.

5 Conclusions

In response to researchers' requests, Statistics Finland releases ready-made research data modules from wide range of topics. The modules provide a relatively easy yet flexible access to administrative microdata for purposes prescribed in the national legislation, that is scientific studies and statistical surveys on social conditions. This data release policy does not compromise the confidentiality of the data. It is also cost- and time-effective for both Statistics Finland as a data provider and researchers as data users. The data files constituting a module are formed only once and stored in a secure environment where the processing of unit-level data can be controlled in a fairly effortless manner.

The ready-made data are a remarkable data source for studies which focus on individuals and enterprises. Over hundred research projects and four hundred researchers are currently using the ready-made data in their studies. In 2020, Statistics Finland processed nearly hundred applications related to ready-made modules, and the number for these kinds of data requests is rising. Most of the projects have access to the total population data, and the most frequently used ready-made data are the FOLK modules containing a wide range of population related data.

The ready-made data can be expected to constitute the core of the microdata services provided for researchers by Statistics Finland. In order to be able to respond to the increased demand, more resources have been allocated for maintaining current ready-made modules and developing new ones. As a consequence, a separate team with seven members responsible for these operations has been formed. Since research projects have expanded their purposes out to new phenomena, the number of ready-made modules is anticipated to increase in the future. In line with the current trend, research projects are expected to become larger in terms of the number of researchers as well as the number of ready-made modules they want access to.

According to the feedback, researchers appreciate the regular updates they get access to automatically without an application process. They also value that ready-made modules are suited for the needs of different kinds of researchers, whether they are beginners or experienced ones. On the other hand, some of the experienced researchers consider certain modules too processed and would like to have access to raw data instead. In these cases, it is possible to apply for tailored data in addition to the ready-made modules. Statistics Finland aims to create new modules with more detailed information to satisfy these needs. For example, during the past year, Statistics Finland has introduced a couple of new modules with little processed data on wage incomes and benefits extracted from the Income Register. These new modules have played a significant role in recent studies about the economic consequences of the COVID-19 pandemic.

References

Statistics Finland. Taika – research data catalogue. Available at <https://taika.stat.fi/en/>.

Statistics Finland (2013). "Development and challenges of on-line micro-data usage. United Nations Economic Commission for Europe Conference of European Statisticians, Geneva.

Statistics Finland (2004). *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*. Handbooks 45. Helsinki.

Törmälehto, V.-M. (2008). "Social statistics – integrated use of survey and administrative data at Statistics Finland." International Association for Official Statistics Conference on Reshaping Official Statistics, Shanghai. Available at <https://www.iaos-isi.org/papers/CS_26_3_Tehto.pdf>.

UNECE (2007). *Managing Statistical Confidentiality & Microdata Access. Principles and Guidelines of Good Practice.* Available at <https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf>.