Social statistics - integrated use of survey and administrative data at Statistics Finland[1]

Veli-Matti Törmälehto*

* Social Statistics, Income and Living Conditions, Statistics Finland (veli-matti.tormalehto@stat.fi)

1. Introduction

This paper reviews the exploitation of administrative and statistical registers in the person and household sample surveys at Statistics Finland. First we describe the infrastructure of Statistics Finland's population and social statistics which consist of the entirely register-based statistics complemented by interview sample surveys. Then we go through the phases of the interview survey process and describe the use of registers focusing mainly on the Labour Force Survey (LFS) and the Survey on Income and Living Conditions (SILC). The final section offers some concluding remarks in a broader quality framework.

2. Surveys and register-based social statistics at Statistics Finland

2.1. Overview of the system

Statistics Finland has a well-functioning, co-ordinated system of statistical registers and Finland has drawn its population and housing census entirely from registers since 1990. Social Statistics[2] are produced mainly from the register-based system or from independent person or household interview sample surveys. Both administrative data and survey data must be used to meet all user needs. Figure 1 gives an overview of the sample social surveys, the register-based population and social statistics, and the units and the links between them in the underlying administrative and statistical registers.

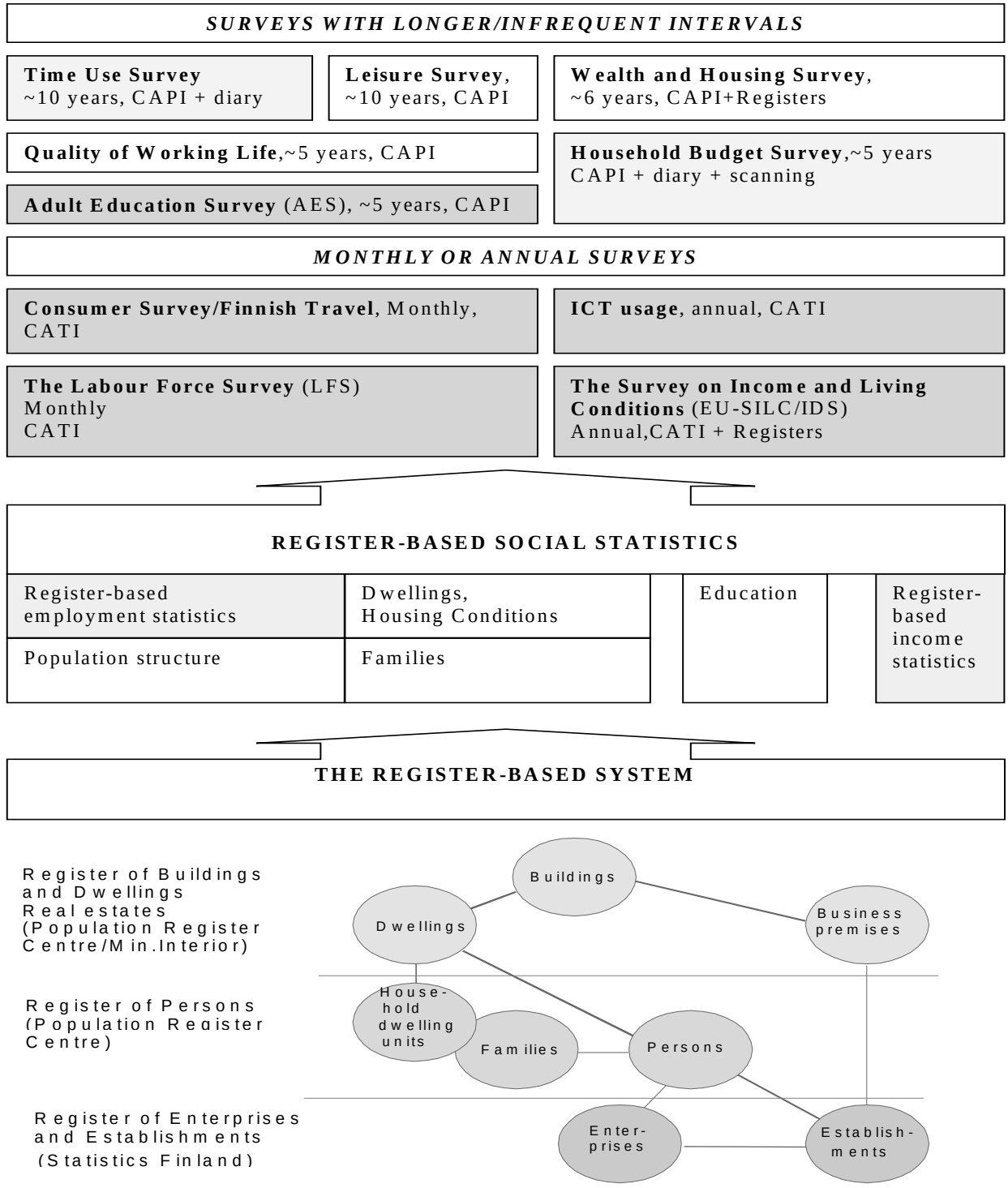
The register-based production is based on one-way traffic from the administrative data to the statistical registers maintained at Statistics Finland. The basic registers are the population register of persons, the register of buildings, dwellings and real estate, and the register of enterprises and establishments. The register system also includes a number of specialised registers, for example the personal income registers, education registers, and labour market registers. Both the basic and the specialised registers may be inter-linked with the unique identifiers: the personal identification number (PIN) for persons, the domicile code for dwellings, and the business identification code (BIN) for enterprises and establishments. Persons sharing the domicile code comprise a dwelling-unit, i.e. the register-based household based on the criterion of co-residence.

---

1 Paper prepared for the International Association for Official Statistics Conference on Reshaping Official Statistics, Shanghai, 14-16 October 2008.

2 The word "social" is used here only to imply that the statistical units are individuals and households. The phenomena described by the current sample surveys seem to fall more into the economic sphere (Figure 1).

Figure 1. Overview of survey- and register-based social statistics at Statistics Finland



The register-based system is the basis for the entirely register-based population and housing census as well as for the annual or more frequent subject matter statistics. The register-based subject matter statistics relevant for the sample surveys are those on population, employment, dwelling-units and housing conditions, families, education, and income. See Statistics Finland (2004) for further details on the register-based statistics at Statistics Finland.

2.2. The surveys

Surveys are used to meet the user needs that cannot be met with register-based statistics. Total (e.g. consumer opinions and preferences, time use) or partial (e.g. household wealth) lack of administrative data on some topics together with timeliness and comparability within the European Statistical System usually dictate the relevance of a survey. All the main surveys at Statistics Finland are either EU regulated or EU harmonised, although they usually have some national content as well.

Statistics Finland conducts two monthly and two annual interview sample surveys on persons/households, four surveys with approximately five year intervals, and two surveys with 10 year intervals[3]. The surveys together with their frequencies and main mode of collection are shown in Figure 1.

The Labour Force Survey (LFS) and the Survey on Income and Living Conditions (SILC) are the two main social survey instruments in the European Statistical System. Both are based on an EU regulation and are carried out regularly in all European countries. The Finnish implementations of LFS and SILC are based on long-running national surveys[4] which have been modified to take the EU demands into account. The LFS sample is also a platform for collecting primary data for statistics on rents.

The ICT use and the AES are EU regulated as well, and would fall under the scope of the planned European Household Survey (EHS)[5]. The Household Budget Survey and the Time Use Survey are EU harmonised but there is no regulation. The HBS is conducted in all European countries but with varying periodicity. The HETUS database of harmonised EU Time Use Surveys currently includes data from fifteen countries. The EU-AES was conducted in more than 20 countries in 2005-2007.

Despite being connected to the European Central Bank's plans of euro area wealth surveys, the Wealth Survey is a national survey. Quality of working life is a national survey.

Official statistics on employment and statistics on income are based on both entirely register-based sources and interview sample surveys (Register-based Employment Statistics and the LFS, the Total Statistics on Income Distribution and SILC). The surveys are ESS surveys, while the most important function of the register-based employment and income statistics is to provide very detailed regional data for national purposes.

2.3. Preconditions for joint use

The Statistics Act enforces the use of administrative data in surveys by decreeing that existing data sources must be used whenever possible. To use register data in surveys, informed consent from the survey respondents is needed, i.e. the survey respondents must be informed that unit level data from registers will be linked to the data they give for the survey. At Statistics Finland, the information on the use of registers must be contained in the advance information sent to the selected respondents before the actual interview takes place.

---

[3] The surveys considered here comprise those surveys where the data collection unit is individual or household, the mode of data collection is personal or telephone interview, and Statistics Finland publishes official statistics from the collected data.

[4] LFS 1957, SILC 1977 (Income Distribution Survey/EU-SILC). SILC is one instrument which serves both national (Income Distribution Statistics) and EU purposes (EU-SILC). For simplicity, we use only the abbreviation SILC here.

[5] The EHS (European System of Social Statistical Survey Modules) project aims to rationalise the existing household surveys and to reinforce the general social statistics infrastructure so that Eurostat is better able to respond to new Commission demands. The EHS is conceived as a system of social statistical survey modules wherein it would be possible to either use the existing national surveys or to use the EU recommended schema.

Before data integration, the issues concerning statistical units, reference periods and record linkage have to be solved. These are discussed below.

2.4.1. Statistical units

The statistical units in social statistics are persons and households. A person is a natural unit, but for practical reasons two different definitions of a household have to be applied. In the entirely register-based statistics, a household is defined as the dwelling unit, comprising all persons registered at the same address. In the surveys, the criterion of common housekeeping (shared income/expenses) is used. The ESS surveys (e.g. EU-SILC) use the latter definition.

Practical reasons are enough to dictate the use of two definitions. Register-based statistics cannot use a shared income/expenses criterion. In interview surveys, it is necessary that all persons from whom data must be collected with interviews must correspond to the actual household members. The dwelling-unit is not a sufficient proxy for the household because the two definitions are identical in approximately 90 percent of the cases (Epland & Törmälehto, 2007). The deviations in the remaining 10 percent are explained by definitions and measurement errors, both in registers and surveys.

Although the dwelling-unit is not used as a statistical unit in the sample surveys, the register definition is used as auxiliary information in data collection. All household surveys feed the register information into their questionnaires and correct the dwelling-unit to the household definition in the interview (see section 3.3).

2.4.2. Reference periods

Different reference periods remain a key problem when linking registers to survey data. The two common reference periods are “current” and “usual”. In the surveys, current is usually defined as “at the time of the interview” which implies moving reference periods in continuous surveys (e.g. LFS, HBS). In the registers, current is usually fixed to a given time, usually set to “at the end of the year”. The typical longer reference period in the registers is a calendar year. In the surveys a longer reference period may refer to the last 12 months preceding the interview (HBS), or the calendar year (SILC).

Internal consistency is a problem with combined use of data from registers and interviews. For example, labour variables from the register-based employment statistics indicate labour status at the last week of the calendar year while in the LFS they indicate labour status at the week preceding the interview. In SILC the reference periods have been set to match those in the registers to facilitate the combined use of register and interview-based variables[6].

2.4.3 Record linkage of register data

The linkage of register variables to surveys is straightforward if the survey units also have valid identifiers. This is the case with linking data on persons. Because all samples are samples of persons from the population database (see section 3.2), all persons selected from the sampling frame by definition have valid identifiers. Probabilistic matching or name-based search algorithms are not needed.

Missing or invalid PINs arise only when interview data must be collected for all household members in addition to the person initially selected from the sampling frame. In this case the dwelling register is essential as it greatly facilitates the record linkage. All the PINs for the persons registered at the same address as the person selected from the frame are known. The surveys create their household roster by changing and correcting the register-information, i.e., by deleting and adding members according to the information given by respondent(s). PINs must be searched from the population register only for the

[6] For example, household composition is fixed to the end of the year composition. The fieldwork period is early in the year to reduce the negative effect of retrospective questioning.

added members in case the PINs they report in the interview are invalid (for example, only date of birth is given) or missing. See Inglic (2007) for record linkage in the Slovenian SILC, which is register-based but without a dwelling-unit register.

The industry and sector of the local unit are very important in the LFS and needed in the other surveys as well (e.g. SILC). These are available as census variables in the register-based employment statistics, based on register-based linkage of employment and business register data (Ruotsalainen, 2005). None of the surveys take industry and sector as such from the register-based employment statistics. The variables are not timely enough even for annual surveys, the reference times may not match, and the validity may have some shortcomings[7]. Instead, the business register data are used as auxiliary data in coding in the surveys.

3. Exploitation of registers in the survey process

Table 1 summarises the uses of registers in the phases of a sample survey process. In ESS surveys the details of the phases (sampling, variable definitions, quality control, etc.) are usually controlled by the relevant regulation. The table also assesses the advantages and disadvantages in the different phases. The advantages are related to relevance, respondent burden, data collection costs and mean square error, while the disadvantages are related to increases in the processing costs.

3.1 Design of the survey

The register data have two types of functions in the data collection: they may be used either as auxiliary information or as the statistical target variable in itself. The register data, which may be used to replace survey variables, need to be determined in the design phase of the survey before questionnaire programming and data collection. Sometimes new data requirements may be replaced with register data. For example, in the Labour Force Survey ad hoc module 2008 on the labour market situation of migrants, the required variables can be taken from the registers without additional questions on the questionnaire.

The basic strategy is to seek the register-based systems for variables where the gap between the register-based and survey-based concept is non-existent or small, data are personal, the reference times are not in conflict, and register data are timely enough. Most of the surveys are able to use data on basic personal demographics (e.g. year of birth, country of birth, sex, citizenship, legal marital status), completed education and annual incomes from the registers as statistical variables[8]. These also are included in the set of core variables of European social surveys as defined by Eurostat. By contrast, labour variables are usually useful only as auxiliary information, because the variables in the register-based employment statistics are not adequately valid (e.g. ILO definition, reference times), detailed (e.g. part-time work is not a category in the register-based variable on current economic activity), or timely enough even for the annual surveys.

[7] E.g. treatment of labour rental workers.

[8] Income data may have to supplemented with interview-based data and imputations for those surveys where income is an analysis variable (SILC, HBS, Wealth Survey).

Table 1. The phases of a survey process and utilisation of register data

Phase	Action	Advantages	Disadvantages
Design of the survey	Replacing survey questions with register-based data The choice of mode of collection and reference periods	Respondent burden	May complicate questionnaire design
Frame creation and sampling	Specification of target population Stratification Simulations, methodological studies	Minimises under- and over-coverage Precision of estimates Pre-tested methods	
Data collection	Feeding data to electronic questionnaire Tracing of sample units	Respondent burden Less measurement error Less attrition	Fragmented and complicated questionnaires More proxy answers
Data processing	Coding Logical checks Micro editing, consistency editing Imputing for item non-response Record linkage of statistical variables Statistical matching	Less manual coding Less measurement error Less need for imputations Improved imputation models Relevance Flexibility	Increased consistency checks and editing Register dependence
Estimation	Non-response analysis Calibration to auxiliary information Methodological studies	Accuracy Coherence Pre-tested methods	
Quality control	Assessment of measurement errors Assessment of estimation errors	Less measurement error Improved estimation Feedback to the registers	
Dissemination	Enhancing dimensions of micro data with registers	Relevance	Confidentiality

Variables on the relationships between household members, such as family units, are somewhat complicated because the status of a person is influenced by the status of other household members, and the register-based household and the survey-based household are not equal because of different reference times, definitions, and measurement errors. Technically, it is straightforward to link all data from e.g. the register-based family statistics to a survey database, and this is done for example in SILC. The variables are mainly useful as auxiliary variables in data controls, rarely as statistical variables in themselves.

The use of register data instead of interview-based data decreases the length of the questionnaire because fewer questions need to be asked and questions on some sensitive topics, such as income, need not be asked at all. This is most evident in EU-SILC which allows for and takes the use of register data

into account in its design. For example, the mean interview duration in the "register countries"[9] in EU-SILC was 22 minutes while in the "survey countries" it was around 60 minutes in the 2006 EU-SILC operation.

While the use of registers reduces respondent burden, the questionnaire itself may become more fragmented and hence more complicated to program and less comprehensible for the respondents and the interviewers if "only the holes are filled in". The survey and register variables should be consistent at unit level within a topic area regardless of the source of the data. To avoid internally inconsistent data, the use of registers in a survey should be considered at the level of subject areas, not at the level of variables.

Eliminating questions from the questionnaire because some variables are available in the registers is not always straightforward. The interplay between register variables, questionnaire variables, and processing costs must be considered. All surveys use electronic questionnaires which may be quite complicated and heavily filtered, i.e., programmed to ask questions on sub-populations. The sub-populations are defined with filter variables. To ask about part-time work and the reasons for it (LFS, SILC), one first needs to ask questions on labour status for questionnaire routing; to ask about housing costs (SILC, HBS),questions on dwelling characteristics must be asked and the actual housing cost questions must be conditioned on these to reduce respondent burden.

For example, if register variables on tenure status and dwelling type were used and interview questions on these dropped, the respondent burden would probably increase because some questions would have to be unnecessarily asked. Using the register housing variables with interview-based housing data (e.g. housing costs) would lead to internal consistency problems and generate costs of consistency checks and editing. It may thus be cost effective to ask the filtering questions and use them as statistical variables, and to use the register variables only as auxiliary data for example in imputations or estimation.

The use of registers affects the mode of interview data collection because the decrease in the length of the questionnaire makes computer assisted telephone interviewing (CATI) a viable option. Consequently, the regular monthly and annual surveys at Statistics Finland are mainly telephone interviews while the more infrequent surveys use predominantly CAPI. CATI seems to be the prevailing mode of collection in the Nordic Countries, especially in SILC, while face-to-face interviews are more prevalent in countries with less extensive statistical registers.

The register data linked to the survey data may be seen as an additional mode of data collection. With the use of registers, all sample units normally have the same data source for a given variable, but the source may vary from variable to another. For example, arrears and over-indebtedness may be based on interview data but debt amounts are taken from the tax register. This is somewhat different from the standard discussion on mixed mode effects, e.g. whether the use of CATI for some sample units and CAPI for some for a given variable causes bias to the results.

There are valid research questions on the effects of different modes of collection on the international comparability of the data. For example, there is evidence that income data based on registers yields lower inequality and monetary poverty estimates than income data collected with interviews (Nordberg, 2003). From the indicators published by Eurostat we find that the countries which use register incomes also have the lowest monetary poverty rates. The flexible use of data sources may reduce the intra-country total survey errors but introduce additional bias to the across-countries comparison.

[9] The countries which use (mainly) register data on incomes: Denmark, Finland, Iceland, Norway, Sweden, the Netherlands, and Slovenia in the latest EU-SILC operation.

3.2. Frame creation and sampling

The population database is the basic sampling frame for the social surveys, and auxiliary information from the registers is incorporated to the population elements for the sampling procedures. The population database is an up-to-date and accurate sampling frame for the surveys, resulting in very limited over- and under-coverage in the surveys. Over-coverage may be excluded with register information such as the age of a person (e.g. the LFS) or the domicile code of the dwelling register (e.g. institutionalised population in SILC). The observed under-coverage in surveys is mainly explained by the time lag between sample selection and fieldwork and reference periods. The samples for the monthly surveys (LFS and the Consumer Survey) are drawn twice a year.

The general strategy in the major surveys is to draw stratified one-stage probability samples of persons from the frame and then link dwelling-unit members to the selected persons for household surveys. Drawing samples of persons is common in the Nordic register countries (e.g. in the LFS and EU-SILC) whereas in the other countries the final sampling units usually are dwellings/addresses. In household surveys the selected person and members of the person's dwelling unit define a cluster, and data are collected from members of the economic household within the cluster (one-stage cluster sampling).

Because a sample of persons rather than a sample of addresses is selected, only the persons registered in the population register may have positive inclusion probabilities. The coverage of the population register, which is known to be excellent, is vital for sample surveys when samples of persons are drawn. At the end of the day, a valid PIN is the necessary precondition for a respondent to be selected into the sample, and in household surveys for any member to be included in the sample[10].

Registers are used to stratify the samples. When the frame is sorted by the domicile code of the dwelling register and systematic sampling is applied (LFS, SILC, HBS), there is implicit regional stratification. In household surveys, sorting also implies sampling with inclusion probability proportional to the size (πPS) of the dwelling unit. Additional stratification is usually conducted with register information, e.g. stratification by type of income and income class (SILC) and further regional stratification (the HBS). The sample may be allocated to over-sample specific population sub-groups. It is easy to pretest different stratification variables and sample allocation options with the register data.

Both the LFS and SILC use rotating panels for cost reasons, and SILC uses them also to provide longitudinal data. Rotating panels may be used for flow statistics, e.g. labour dynamics from the LFS or in- and outflows from poverty from the EU-SILC. For national uses the entirely register-based sources are much more useful due to their accuracy. For EU-wide flow statistics, the EU harmonised surveys with rotating panel feature may be useful, as demonstrated e.g. by Romans (2008).

3.3 Data collection

In the data collection phase, register data are used as auxiliary information to feed data to electronic questionnaires and to trace sample units.

A priori information from the registers may be pre-filled to the electronic questionnaire, in a similar fashion as data from previous waves are fed forward for dependent interviewing in rotating panel settings (such as the LFS and the SILC). The pre-filling from the population register is used in all surveys for location and personal demographics, and in household surveys for automatic feeding of dwelling-unit composition. In the interview the place of residence is verified and the dwelling-unit corrected to a consumption unit (household).

---

[10] In SILC, the number of persons who in the interview reported to be member of a household but for whom no valid PIN could be established from the registers is restricted to very few cases, indirectly confirming the very high quality of the coverage of the population register.

One way to exploit register data is to feed information on missing register data to the questionnaire to improve imperfections in register data. Questions are then asked only from the persons with missing register data. For example, in the SILC questionnaire register data on education is fed onto the questionnaire. More detailed questions are only asked if register data are missing; otherwise the questions are bypassed. This method is used to get more detailed information on those respondents with no or only basic education (coded missing in the registers) and to improve serious register under-coverage in the year of completing the degree.

Because all samples are samples of persons, the persons initially selected to the sample (the selected respondents) must be traced if they have moved. Tracing of persons is also necessary in panel settings, such as the four-year longitudinal EU-SILC. The population register is used for tracing of selected respondents and to update changes in personal information in longitudinal surveys before the fieldwork of the second and subsequent waves.

3.4 Data processing

Data processing includes record linkage of register data to the survey units. This is quite straightforward once the valid identifiers for the survey units are confirmed. The subject matter statistics within the area of population and housing census are combined into one MS SQL-database called Herttua (“the Duke”). The persons responsible for the survey may determine the usable variables within the census database, and a restricted survey-specific SQL view can then be created for those responsible for the record linkage. The major surveys operate in MS SQL-databases as well.

The surveys which need to code sector and industry (LFS, SILC, HBS) use a coding application with a name-based linkage established by matching the name of the employer/enterprise given by the survey respondent with the corresponding name in the business registers. Manual checks and coding are needed to cope with matching failures and to find local units for multi-establishment enterprises and local government's operating units[11]. Register-errors noted in the monthly coding of industry and sector in the LFS are reported to the business register.

To edit incorrect and/or outlying values, suitable register data may be used as auxiliary information or as replacement data. In data fusion of different registers, decision rules have to be set in case of conflicting information. The problem is magnified with combined use of interview and register because more conflicting information is typically revealed. After integrating data from different sources and micro-editing variable by variable, an additional round of consistency checks and editing is often needed. A clear set of preference ordering or automated editing rules are needed in case of conflicts. These rules are usually survey and variable dependent.

The use of register data reduces the need for imputations especially in surveys where income plays a key role as a statistical variable (SILC, HBS, Wealth and Housing Survey). Incomes are sensitive data and typically suffer from large item non-response. Countries with no or only limited access to register-based income data have to devote quite some effort to imputation of non-response. Imputations are still needed, for example for working hours in the LFS and SILC and some income and expense items in SILC. The imputation models may be improved by using register data as auxiliary information in distance hot deck, regression- and mean imputations, and even in deductive imputations. For example, the unit or sub-unit (within household) non-response of occupations (typically as a result of uninformed and/or proxy answers) may be imputed with register variables on occupation, labour, education, and income.

[11] The link may also be established based on income data; this option is used in the national part of SILC (the IDS). For the surveys with other reference periods (current, i.e. at the time of the interview), this is not feasible.

For the monthly surveys (the LFS, Consumer Survey) the production time and the reference periods give less room for using registers to check and correct interview data. A polar opposite is SILC where the reference periods in the survey part are set the same as in the registers (end-of-year, calendar year) in order to edit and impute interview-data with register data and to maximise the use of register variables. Using registers brings along more derived variables compared to a pure interview survey (where questions can just be asked according to a classification), and hence more derivation and workload to the processing phase.

Since the same variables from the register source can be linked to all the surveys, these variables could be used in statistical matching, in other words, to donor data from one source to another (for example consumption data from the HBS to income survey SILC) using the commonly observed register variables. This technique has not been used yet in production of official statistics at Statistics Finland, however.

3.5. Estimation

The register variables are available for both the respondents and the non-respondents, and therefore may be used to describe and analyse unit non-response. The results of the analysis may be used to identify problematic sub-groups in order to develop targeted measures for the fieldwork process or to over-sample problematic sub-groups, and to adjust the calibration models. Longitudinal register data have been used in methodological studies of attrition in panel surveys.

Re-weighting is used to compensate for effects of unit non-response and in some instances for item non-response. Calibration estimators are used in all surveys at Statistics Finland. The sampling weights based on reciprocals of inclusion probabilities are modified to exactly reproduce certain population marginal distributions, i.e. by calibrating them to marginal distributions of register variables. In calibration, it is essential that the external control variables are strictly comparable to the variables available in the sample. This, of course, is the case when the control variables are record linked from the registers to the sample units.

The calibration variables need not be the survey target variables but the two must be highly correlated to obtain more accurate estimates. Basic demographic distributions, e.g. distribution of population by age and sex, are always used as calibration variables in the surveys. Depending on the survey, other survey specific variables may be used: register-based job-seeker status in the LFS, income in cross sectional SILC, longitudinal demographics in longitudinal EU-SILC, and so forth.

Calibration to register data is a powerful estimation technique. It improves coherence and consistency of sample surveys with other surveys, entirely register-based statistics, and other sources such as National Accounts. There are gains through improved accuracy of the estimates for given sample sizes, or through reduction in the required sample sizes for pre-specified levels of accuracy. The calibration technique is also used for quality studies to control for the effects of non-response bias.

3.6. Quality control

The register information may be used for quality controls of survey interview data, and sample surveys may be used for quality controls of the registers.

For the sample units, errors of measurement may be evaluated with the use of record linked data from the registers, for example by linking labour variables, e.g. occupation from the occupation register, to the LFS sample. The errors of estimation, i.e. the biases and uncertainties involved in going from the observed sample to the whole target population can be compared because for the register variables both

sample estimate and true population value are known. For example, the distribution of register-based incomes in SILC and the entirely register-based income statistics may be compared.

Surveys are used to monitor register data quality as well. The quality of permanent address, tenure status, native language, and occupation data in the Finnish Population Register is monitored with the Labour Force Survey (Hokka & Nieminen, 2008). The SILC sample is used to check for errors and to validate income registers, and it provides regular data on the differences between the housekeeping households and the dwelling households.

3.7. Dissemination

In addition to standard dissemination of the results by Statistics Finland, anonymized micro data for outside use has for quite some time been the essential outcome of sample surveys. Increased accessibility markedly increases the relevance of survey data. Given the costs of the surveys, resorting to basic output in the form of descriptive tables and basic analysis only would not be justified. The surveys need to perform well also as analytical surveys in research; the connections and interdependencies of phenomena are important. For the purposes of official statistics or EU-indicators, the sample surveys are often more descriptive in nature, aiming at the estimation of totals, averages, and proportions at population and sub-group level. The use of data for analytical as well as for descriptive purposes has to be taken into account in the survey process.

On a national level, micro-data from all surveys have been in active research use at least since the early 1990s, for example the national version of the SILC data is used as the basis for national micro-simulation models to study tax and transfer systems. At the European level, the LFS and EU-SILC micro-data are delivered to Eurostat. Eurostat produces cross-national datasets and disseminates them for research use[12]. The Finnish Household Budget Survey is also delivered to Eurostat as microdata.

Apart from the standard anonymized micro-data sets, new user needs may be met by creating user-specific micro-data sets from survey and register data. Merging of register and interview data yields broad information sets. It must be ensured that the respondents have been informed about the uses of register data, and care must be taken to eliminate or minimise the risk of disclosure (e.g. when register incomes are merged to a survey). Because of sample sizes and non-response, sample surveys have smaller disclosure risks compared to samples taken from the entirely register-based systems as these are larger and may be longitudinal. However, the entirely-register based micro-data sets are not limited by the principle of informed consent.

4. Summary and conclusions

The register-based statistical system influences the sample surveys in many ways. For efficient data integration it is first of all essential that the statistical registers are included in a co-ordinated system to be exploited and that the surveys are inter-linked to this system; this is the case at Statistics Finland.

The main challenges in using register variables in the surveys are related to the definitions and validity of the register variables, to the timeliness of the registers and the punctuality of the surveys, and the interplay between questionnaire design and register contents. The key feature of the survey concepts is that they are usually determined in an international context and need to be internationally comparable. The data coming from the European social surveys, such as the LFS and SILC, can only be relevant if they are comparable across the EU Member States.

[12] Similar dissemination is planned for the Adult Education Survey AES.

A set of register-based variables on personal demographics, completed education, and income are virtually always used in the surveys to replace direct data collection. Geographical information and household composition are fed to the questionnaire from registers and corrected in the interview. The register variables commonly used in the surveys cover 10 of the 16 core variables for the European social surveys as defined by Eurostat. Labour-related data and housing data as well as the business register data are useful mainly as auxiliary data.

In monthly surveys, such as the Labour Force Survey and the Consumer Survey, timeliness as well as the subject matter restrict the exploitation of register data in the data collection and processing phases. The most elaborated case is the annual Survey on Income and Living Conditions which is built explicitly on the fusion of interview- and register-based data.

The registers are utilised in a standardised way in frame creation and sampling, estimation and dissemination. The use of common register variables and calibration estimators improve accuracy in terms of mean square error as well as coherence with other statistics. Replacing some of the survey contents with register variables decreases respondent burden and improves data quality. On the downside, the survey questionnaires may become more fragmented and more complicated to program, and additional rounds of consistency checks and editing may be needed.

An interview sample survey remains a flexible, reactive, accessible and timely instrument with the combined use of survey and register data, although some of the control over data content is lost. When the definitions change, the statistical instrument must be capable of reacting to the changes. The use of registers may delay the production in annual surveys. However, the delay is justified considering the reduction in data collection and processing costs as well as the quality of the estimates, and the fact that the annual and more infrequent surveys produce structural rather than short-term statistics.

References:

Epland, Jon & Törmälehto V-M (2007): *From Sample Surveys to Totally Register-based Household Income Statistics: Experiences from Finland and Norway*. Paper prepared for the conference of the European Survey Research Association, Prague, 25-29 June 2008.

Hokka, Päivi & Nieminen, Markku (2008): *Measuring the Quality of the Finnish Population Register with a Survey*. Paper prepared for the European Conference on Quality in Official Statistics, Rome, 8-11 July 2008.

Inglic, Rihard (2007): *Administrative data and registers in EU-SILC*. Paper presented at the Seminar on Registers in Statistics, Helsinki 21-23 May 2007.

Nordberg, L. (2003): *An Analysis of the Effects of Using Interview versus Register Data in Income Distribution Analysis Based on the Finnish ECHP-surveys in 1996 and 2000*, Chintex Working Paper #15, Work Package 5, December 22 2003.

Romans, Fabrice (2008): *Measuring labour market flows in Europe with the Labour Force Survey*. Paper Prepared for the 30th General Conference of The International Association for Research in Income and Wealth, Portoroz, Slovenia, 24-30 August 2008.

Ruotsalainen, Kaija (2005). *Combining enterprise data to employment data in register-based employment statistics*. Paper prepared for the Siena Group meeting on Social Statistics, Helsinki.

Statistics Finland (2004): *Use of Registers and Administrative Data Sources for Statistical Purposes: Best Practises of Statistics Finland*. Statistics Finland, Helsinki.