Statistical Journal of the IAOS 30 (2014) 411–424
DOI 10.3233/SJI-140803
IOS Press

The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics

Bart F.M. Bakker[a,b,*], Johan van Rooijen[a] and Leo van Toor[a]
[a]*Statistics Netherlands, Den Haag, The Netherlands*
[b]*VU University Amsterdam, Amsterdam, The Netherlands*

Abstract. More and more countries are using register data to replace traditional Censuses. Moreover, official statistics as well as research are increasingly based on register data or combinations of survey and register data. Register-based statistics offer wonderful new opportunities. At the same time, they require a new approach to how data are processed and managed. In this article, we present the System of social statistical datasets (SSD), a system of interlinked and standardized registers and surveys. All production processes within Statistics Netherlands that pertain to social or spatial statistics converge in the SSD, which thus constitutes a shared output-oriented system. The SSD contains a wealth of information on persons, households, jobs, benefits, pensions, education, hospitalizations, crime reports, dwellings, vehicles and more. In the Netherlands it is the most important source for official social statistics and, because the data are available on request by means of remote access, also very popular in the social sciences. This article describes the contents of the SSD as well as the underlying process and organization, and demonstrates its possibilities.

Keywords: Registers, administrative data, data processing, micro-integration

1. Introduction

Administrative data are quickly becoming increasingly popular in the production of official statistics as well as in social research. A number of developments have contributed to this mushrooming growth.

Until the early 1990s, the usual instrument for collecting statistical data on persons, households and businesses was the sample survey. However, the steady rise in non-response rates in household surveys in that decade raised serious doubts about the quality of the survey outcomes [7,10,14,39,40]. This furthered the use of administrative data either for weighting variables to correct for non-response bias, or to replace sample survey data altogether.

Added to this, political pressure and budget cuts have forced statistical offices to reduce the number of (sample) surveys in order to lower the reporting burden and work more efficiently. Besides being an efficient way for statistical offices to collect a lot of valuable information, the use of administrative registers substantially lowers the reporting burden for companies, institutions and households.

Developments in information technology have led to a growing number of digital administrative registers with relevant information. Moreover, more and more countries have introduced a personal identification number for administrative purposes. As digital registers containing personal identification numbers can be linked fairly straightforwardly, users are

*Corresponding author: Bart F.M. Bakker, Statistics Netherlands, P.O. Box 24500, 2490 HA Den Haag, The Netherlands. Tel.: +31 703375707; Fax: +31 703877429; E-mail: bfm.bakker@cbs.nl.

quickly becoming increasingly aware of the new possibilities offered by administrative data. In addition, users' needs have changed rapidly in recent decades: today they want relevant and authoritative statistical information, providing insight into the complex relationships between different aspects of social and economic life. This information should contain enough detail to specify the situation of small groups in society and to enable estimation of phenomena with a low incidence. Lastly, the information should be provided regularly so that important developments can be monitored. By using administrative registers, large numbers of records can be obtained at one go: for example the population register, social security and tax data. Therefore, studies of regional phenomena and small sub-groups as well as longitudinal studies and small domain statistics are possible without placing an additional burden on respondents and encountering problems associated with panel attrition. Moreover, registers provide accurate measurements of some phenomena that are difficult to measure with questionnaires because of social desirability issues, like criminal behaviour.

Evidently, adequate legislation is a key precondition for the use of administrative data sources for statistical purposes. Thus, the use of administrative data by Statistics Netherlands (SN) would not have been possible without adjustments in existing legislation as well as the development of new legislation. Current legislation on the one hand authorizes Statistics Netherlands to use administrative data from all government institutions. On the other hand it obliges SN to take adequate technical and organizational measures aimed at data security and privacy protection.

In 1996, SN carried out a first feasibility study to examine the possibilities offered by joining administrative data and survey data. Data from the population register and the administration of employee insurance schemes were processed and subsequently linked to data from the Labour Force Survey. The emphasis of the study was on elementary issues, particularly the quality of the matching process, which is obviously a critical success factor. The results were promising and marked the beginning of the development of the System of social statistical datasets (SSD). Since then, the SSD has expanded enormously. A major milestone was the 2001 Census which was based on the SSD [32]. The 2001 Census cost approximately 3 million euros, next to nothing given the costs of a traditional Census: an estimated 300 million euros [31]. Nowadays, the term SSD primarily refers to a system of linked statistical registers and surveys which cover a broad range of demographic and socio-economic subjects: from labour force participation to social security, from health care to crime, from housing to migration. The content of the SSD is inextricably bound up with an elaborate support system consisting of an organization, processes, metadata, software tools, standardization and coordination principles, procedures and privacy protection measures. This system has been developed to ensure the efficient and secure use of the data in the SSD as well as to control, as much as possible, several quality aspects of the derived output.

This article aims to provide valuable information for developing register-based statistics, founded on almost twenty years of experience in the Netherlands. Furthermore, it can provide insight into the content and processes of the SSD for external users. The core elements are:

1. That data are centrally stored in a standardized way.
2. The different unit types (persons, buildings, households, companies) can be easily linked because of assigned linkage keys.
3. Coordination is crucial to obtain consistent outcomes. Coordination comprises organizational, technical, and content related aspects.

The development of Dutch register-based statistics did not begin from scratch. The fundamentals were already available from the Nordic countries and the United States. The Nordic countries started using data from registers very early on [36,44]. Denmark was the first country to fully base a census on administrative register data. Today, Norway, Finland and Sweden use large numbers of administrative registers for their Censuses. The United States were already using administrative data for their business statistics in the 1980s.

During the development of the SSD a lot of methodological, logistical and practical problems were solved. However, a few important ones still remain. In the discussion, we will give attention to these remaining problems, in particular challenges pertaining to methodology and to integral quality management.

2. The SSD

2.1. A brief overview

The transition from traditional survey-based to register-based social statistics has gone hand in hand with an enhanced necessity to combine and integrate data sources. The primary reason for combining sources is

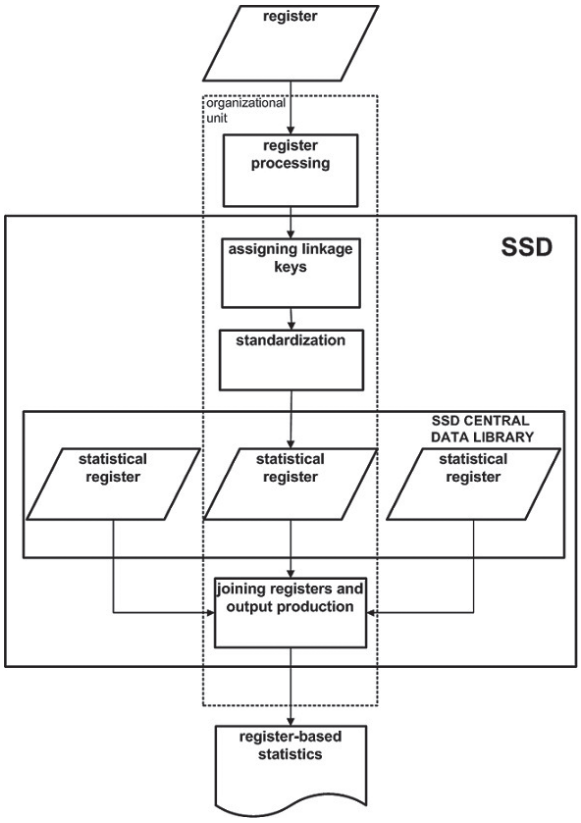

Fig. 1. Broad overview of the SSD. Organizational units process administrative registers and store the resulting statistical registers in the central data library of the SSD. Statistical registers are then combined to realize output.

that registers – unlike sample surveys – usually cover a limited number of variables [4]. Thus, a single administrative source will rarely suffice to attain the aspired scope and depth of statistics. Consequently, organizational units within SN need to share their statistical registers. This is the main purpose of the SSD.

Figure 1 gives a broad overview of the process underlying the SSD. The core of the SSD is a central data library, maintained and operated by SN's Division of Socioeconomic and Spatial Statistics. The various organizational units of the division are responsible for statistics pertaining to specific themes, e.g. employment, social security, demography, and manage processes in which register data are collected, edited and imputed. As administrative register data are not collected for statistical purposes, these processes are usually quite extensive [5,11,51,53]; extensive processing is required to achieve acceptable quality. Register processing is beyond the scope of the SSD and will therefore not be elaborated upon here, except for the part concerned with consistency for census purposes. This is part of the content-related coordination.

Register processing is followed by the assignment of standardized linkage keys which enable different statistical registers to be combined efficiently and are therefore central in the production of register-based statistics e.g. [36,44,51]. The resulting statistical registers are standardized and then stored in the central data library of the SSD. The corresponding metadata are stored in a central metadata repository. Storing statistical registers in a standardized form in a central library makes it easier for organizational units to share. Organizational units enrich their own statistical registers by joining with statistical registers supplied by others, which enables the realization of statistical output with the required scope and depth. In addition, external scientists are given the opportunity to access and join the statistical registers stored in the SSD for their research purposes. In the remainder of this paper, we shall discuss in more detail each of the processes and elements outlined above.

2.2. Assignment of linkage keys to statistical registers

The central statistical unit types are persons, households, buildings, and organizations (companies and non-profit organizations). In order to link the relevant information on these unit types, all units are identified and assigned a linkage key.

Most administrative registers put at Statistics Netherlands' disposal contain unique personal identifiers, the citizen service (CS) numbers, previously called the social security and fiscal number (SoFi number) [3,33]. Various registers contain other personal identifiers such as date of birth, name and address, without [13] or alongside [3] the citizen services numbers. Although these personal identifiers can be used to link and join different registers, for more efficient register combination as well as for privacy protection, the personal identifiers are replaced by the PIN (person identification number) and AIN (address identification number) linkage keys. Both are anonymous keys which preclude direct identification of persons. The Dutch population register (PR) plays a central role in the assignment of these keys. The PR contains personal identifiers and demographic information for every registered inhabitant of the Netherlands [28]. The cumulative PR data from 1995 onwards are used by SN to maintain a so-called central linkage file of persons (CLFP). In the CLFP, persons are assigned a PIN and addresses an AIN. All registers with data on persons

are linked to the CLFP on the basis of the original personal identifiers. In the case of a link, the PIN and AIN are taken from the CLFP [3,33]. At the same time, the original personal identifiers, except for month and year of birth, are removed from the register.

Obviously, errors in registered personal identifiers as well as duplicate keys may result in missed links as well as erroneous links which in turn may cause bias in derived statistics e.g. [4,22,26,29,53]. Therefore, SN has developed and implemented a method which attempts to maximize linking rates while at the same time minimizing erroneous links [3]. When available, the CS number is used to link files to the CLFP. This number is of high quality in most sources, and results in a linking rate of almost 100 percent [33]. Records without a CS number are linked on the basis of other personal identifiers such as date of birth, sex and address [3,13,24]. In the first step of this process, records are linked if the values of all the personal identifiers are identical. The remaining, unlinked, records are subjected to a second step in which some differences are allowed. The method to assign linkage keys is deterministic. However, because in the second step some differences are allowed in the linkage variables, the deterministic method leads to results that are similar to probabilistic linkage. This prevents the application of computer time intensive probabilistic methods [3]. The resulting linking rate varies depending on the quality of the identifiers. For instance, 87.6 percent of the hospital discharge register records could be uniquely linked [13] while the linkage rate for the Labor Force Survey was nearly 100 percent [33].

Clearly, the household is another important statistical unit in the context of socioeconomic statistics. However, in most European countries it does not exist as such in the available administrative registers e.g. [52,54]. The Netherlands is no exception, the households – and corresponding household identification numbers (HIN) – are therefore established on the basis of information from various registers on family relationships, co-residence and fiscal relationships [46]. The majority of households can be deterministically identified on the basis of information on address and family ties. These are mainly traditional households such as single-person households and families living at the same address. Some of the more challenging households, e.g. unmarried couples, are deterministically identified on the basis of fiscal information. The remaining households are stochastically imputed on the basis of household information collected in the Labour Force Survey (LFS [19]).

A fourth and last important linkage key is the organization identification number (OIN). Various administrative registers contain information about companies or non-profit organizations, the most prominent example being the administrative register on jobs which contains the tax number of the company providing the job. The tax number is replaced by the OIN by linking with the business register which can be regarded as the economic counterpart of the CLFP. The OIN enables the compilation of social statistics which include attributes of companies, such as industrial classification (NACE code) and size class (SC code). Moreover, it enables integration of social and economic statistics [15,47].

2.3. Standardization

SN strives to apply a common architectural framework when designing and redesigning production processes [41]. This framework serves as a guide, thus leaving room for organizational units to tailor their production systems to the specific challenges they face. Consequently, register processing, which precedes the SSD (Fig. 1), is characterized by a low degree of standardization. However, all production systems converge in the SSD, which can be seen as a common system aimed at data sharing and output production. Efficient data sharing requires a high degree of standardization which therefore plays a prominent role in the SSD. A mix of autonomy and integration of production processes also seems to have been fruitful in other national statistical institutions (NSIs) [27,34]. At this stage of the SSD, standardization is mainly related to the format and name of a statistical register and its corresponding, obligatory, metadata files and the data type of linkage keys. The standardization of a statistical register and its accompanying metadata files is checked by a tailor-made software tool to preclude the storing of faultily standardized data in the SSD.

2.4. The central data library of the SSD

The central data library of the SSD constitutes the heart of the SSD as it contains the actual data used by internal as well as external researchers to generate output. The main idea underlying the central data library is efficient data sharing, as well as the coordination thereof. Therefore, we elaborate on some of its key principles below.

Table 1
Some important variables by object type

Object type	Statistical variables	Linkage keys
Person	Reference period, nationality, country of birth, year and month of birth, gender, marital status, position in household, partner (no partner, cohabitation, marriage), educational attainment (Standard Classification of Education), personal income, cause of death, crime reports	PIN HIN AIN PIN father PIN mother PIN partner
Household	Reference period, household type, household composition, number of children in household, household disposable income, household capital	HIN
Building	Taxation value, rented or owner-occupied home, type of home, energy label, regional classifications, geographic coordinates	AIN
Activity: having a job	Reference period, wages, hours worked, temporary or permanent labor contract, employee type, collective labour agreement, company car, discharge reason	PIN OIN
Activity: being self-employed	Reference period, operating profit, industrial classification (NACE)	PIN OIN
Activity: being enrolled in education	Reference period, Standard Classification of Education, year / stage of education	PIN OIN
Activity: receiving social security benefit	Reference period, type of benefit (income support, unemployment, disablement, survivor, benefits from abroad, other), amount received	PIN
Activity: receiving pension	Reference period, type of pension (state or employer pension), amount received	PIN
Activity: registered at employment agency	Reference period, desired working hours	PIN
Activity: registered as medical practitioner	Reference period, medical profession	PIN
Activity: receiving scholarship	Reference period, amount received	PIN
Activity: hospitalizations	Reference period, number of hospitalizations, number of days in hospital	PIN
Activity: owning a vehicle	Reference period, technical specifications of vehicle	PIN
Organization	Reference period, Industrial classification (NACE), size class (SC code)	OIN

2.4.1. Central data library of the SSD: Contents

More than fifty administrative registers underlie the current SSD. Enumerating these registers as well as the contents of the derived statistical registers would provide little insight into the SSD as a register system. Instead, the contents of the central data library of the SSD are explained here by means of a conceptual model (Fig. 2) and a brief overview of the contents (Table 1). In Fig. 2, object types (statistical units) are represented by rectangles and relations between object types by connecting lines. In the SSD, a relation between object types is represented by one of the common linkage keys. For instance, the line to and from the object type "person" represents relations between persons. Persons may be related by being partners (1:1 relation, linkage key PIN) or relatives (1:n relation, linkage key PIN). Similarly, the line between person and building represents the registration of persons at specific dwellings (linkage key AIN). Following Wallgren and Wallgren [51], an object type "activity" is included, where activity should be interpreted very broadly. Examples of activities are "having a job", "being enrolled in an educational program", "receiving a social security benefit". Most activities are related to persons as well as to organizations. A job is obviously related to both the employer and the employee. Similarly, educational enrollment is related to the student as well as to the education institution. Table 1 gives some important variables by object type. This list is not exhaustive.

2.4.2. Central data library of the SSD: Coordination

The need to share data is a major change from traditional survey-based statistics production, which was much more autonomous [41]. Data sharing requires statistical registers to be carefully organized so that they can function as a register system, and therefore coordination is essential [27,44,51]. Lack of coordination may result in various undesirable phenomena, such as identical variables stored under different names, different variables or statistical registers stored under the same name, statistical registers stored in different file formats, linkage keys stored in different data formats, metadata stored in different forms and file formats, etc. Indeed, it is easy to envisage the evolution of an unmanageable collection of data and metadata in the ab-

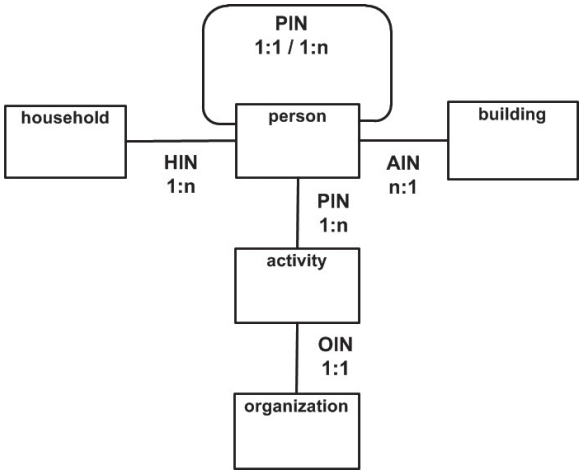

Fig. 2. Conceptual model of the SSD register system. [Rectangles: object types; lines: relations between object types; PIN: person identification number; HIN: household identification number; AIN: address identification number; OIN: organization identification number; the indication x:y denotes the type of relation].

sence of coordination. Moreover, data sharing among organizational units entails increased interdependency as well as the potential for unwanted output overlap. Therefore, being able to monitor the production schedules of other units is of paramount importance. In short, coordination is essential to simplify the combined use of data, to increase consistency between statistical registers, avoid duplicated work, ensure the appropriate application and interpretation of data, and for planning and control. Four types of coordination are distinguished: organizational, technical, content-related and output-related. These will be examined consecutively below.

Organizational coordination

SN's Division of Socioeconomic and Spatial Statistics consists of a number of organizational units. Each unit is responsible for the production of statistical output pertaining to a specific domain, e.g. employment, social security, demography. These units carry out register processing and store the resulting statistical registers in the central data library of the SSD. They are the formal owners of these registers, which means they are accountable for the timely processing as well as the quality of the registers. Several supporting tasks are performed by two central organizational units: one is responsible for assigning linkage keys to statistical registers. To that end, it maintains the CLFP and develops and applies matching algorithms. The other central organizational unit carries out a broad range of activities aimed at the integrity of the SSD and the efficient use of its contents. For instance, it performs micro-integration of different statistical registers, develops and maintains software tools and provides courses on the principles of the SSD. Lastly, two consultation bodies are worth mentioning. First, representatives of all organizational units participate in a consultative body which aims to coordinate the contents and technical aspects of the SSD. Second, a steering committee oversees current and future aspects of the SSD and takes action in the case of conflicts of interest.

Technical coordination

Standardization is the most prominent aspect of technical coordination within the SSD. File formats, data formats of linkage keys, naming conventions, metadata, IT infrastructure and planning tools are all standardized. Technical coordination also aims to prevent redundancy (the same variable in different statistical registers) and ambiguity (same variable under different names). In addition, a key feature of the SSD is an unambiguous link between data and metadata (Fig. 3). Meta-information and its structure is important for the proper processing and understanding of statistical data e.g. [17,44]. The transition to register-based statistics has broadened the demands on metadata as it entails a stronger dependence on external factors such as legislation underlying the administrative registers, variable definitions and data collection methods employed by the register keeper [11,36,51]. The metadata of the SSD are stored in a central metadata repository. Statistical registers are connected one-to-one with their corresponding metadata files, on the basis of the register name. Similarly, variables are related to their metadata on the basis of the variable name.

Content-related coordination

Several processes are directed at the coordination of content. Firstly, when either new statistical registers or modifications of existing registers are developed, the specifications are sent to all organizational units to enable stakeholders to contribute comments that represent their interests. Secondly, a central production schedule is kept within the SSD framework. Organizational units make their own timetables using a standardized planning tool. These timetables are automatically incorporated into a central schedule which can be consulted by all the units. Thirdly, if a historical register is updated frequently in order to produce timely statistics, coordinated versions are identified which are to be used for all statistics with less strict timelines. For instance, the demographic register, which is de-

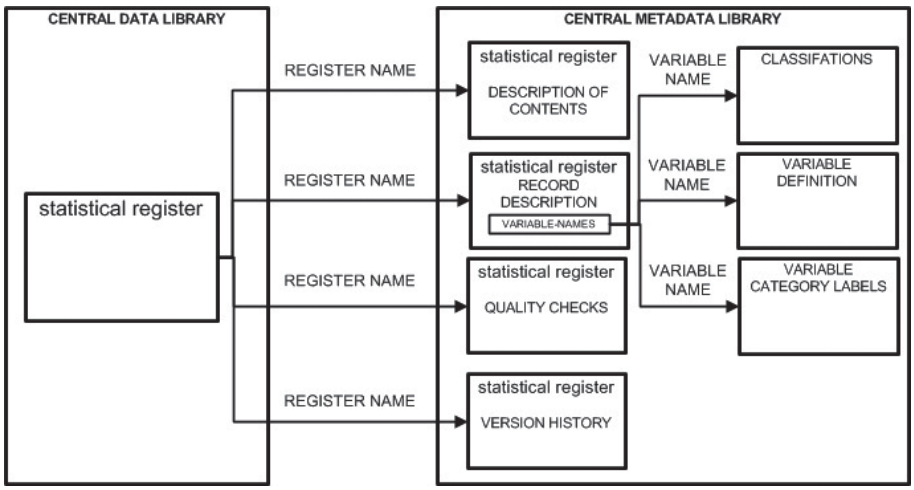

Fig. 3. Relation between data and corresponding metadata on the basis of register names and variable names.

rived from the PR, is a historical register which is updated monthly. The updates include addition of data for the new month as well as the addition and modification of data for previous periods. Therefore, each update also involves a qualitative improvement of "historical" data. Apart from the obvious advantages, this complicates the realization of consistency and reproducibility of statistics. Therefore, a coordinated version is designated each year, which represents a standardized population. Fourthly, classifications are coordinated by storing standardized classifications and groupings thereof in the central metadata repository.

Lastly, micro-integration is carried out to construct a set of mutually consistent as well as longitudinally consistent statistical registers. This register set is produced specifically for Census purposes but is also frequently used by external researchers. The micro-integration framework was developed by Bakker and Daas [6] and Zhang [53], based on the idea that the total survey error approach could also be applied to register data. The framework distinguishes measurement and representation errors. We summarize mainly Bakker [4].

Measurement errors are caused by differences between the administrative and statistical concepts, and by the operationalization and measurement of the administrative concept. The errors can be detected by comparing meta-information and examining inconsistencies in the data. For instance, measurement errors occur if the same variable has different values in different sources, a logical relationship is violated by the data, state and transition figures are inconsistent, impossible or implausible transitions are registered, or the data are inconsistent with some external reference data.

The measurement errors are corrected by a set of edit rules, starting with the conceptual definition of the statistical variable. As measurement may vary between different sources, in the first set of rules, the measured variables are transposed into the statistical ones. This first step is called harmonization. Once harmonization has reduced inconsistencies in the data, the remaining inconsistencies are resolved by choosing the best source for each variable. To choose the best source it is important to know the quality of the variables in the different sources. The quality of a variable in a source can be strong in terms of one aspect, but weak in terms of another. For example, the yearly wages in source A can be of very good quality for government employees, but of fairly poor quality for employees in other economic sectors. If source B is fairly good for all employees, the yearly wages of government employees are derived from source A and those of other employees from source B. If the quality details of the sources is unknown, sometimes the new variable is derived from two or more sources by taking the mean. It is also possible to formulate a decision rule in which the data are required to fulfil a logical relationship. It then depends on the quality of the variables which one is adjusted.

Representation errors originate from differences between the target population of the administrative register and the statistical target population: population elements may be missing in the administrative register, or elements not belonging to the population may still be included in the register, you may have missed links and therefore population elements may be missed, as a result of mislinks elements may be included in the register that do not belong to the population. These er-

ror sources could lead to either undercoverage or overcoverage of the population.

Overcoverage can be corrected by deleting elements that do not belong to the target population. To do this, these elements have to be identified. Undercoverage is more difficult to correct. One way is to combine all kind of incomplete sources, to create a complete list of population elements of the target population. But if population elements are missing in all the combined sources, they will also be missing from the combined file and undercoverage will still occur. Another way to correct for undercoverage is to link administrative register data with survey data that cover the total population. Assuming that the register data cover their part of the population entirely and could be assigned a weight of one, the rest of the population is covered by the weighted records of the survey that could not be linked [24]. It is also possible to assign weights to the records in the combined file or to impute records, if the total population size can be estimated, e.g. with capture-recapture methods. This method is used, for instance, for the Integrated Census in Israel [21].

Output-related coordination

Generation of statistical output requires coordination to avoid duplicated work and inappropriate use of data. If a local organizational unit combines statistical registers for output purposes, it will use its own data as well as data delivered by other units. Subject specialists of those other units should be able to review the input and output specifications so as to ensure that data are used and interpreted appropriately and to check that similar statistical output is not being constructed by other units.

2.4.3. Central data library of the SSD: Privacy protection

In the 1970s there was a growing concern in the Netherlands about the protection of privacy [1,2]. Although the 1971 General Population and Housing Census prompted public debate on the subject [48], this had only a slight effect on the response to that census: the final non-response rate amounted to 0.2 percent. However, the fast growing concern caused a postponement of the 1981 Census – as the non-response rate was expected to rise to as much as 26 percent – and ultimately led to the decision to abandon the traditional Censuses altogether in later decades. The public debate itself started a process of legislation on the subject of protection of privacy. Today the protection of privacy is well-regulated in the Netherlands.

The legal basis of Statistics Netherlands (SN), the Statistics Netherlands Act [37] stipulates that SN must use administrative data from government institutions wherever possible, and grants them authorization to do so. In addition, it authorizes SN to use the Citizen Service (CS) number. Several articles of this official Act are aimed at data security and privacy protection. First of all, the data received by SN may be used solely for statistical purposes. Secondly, SN must put in place technical and organizational provisions against loss or interference with these data, and against unauthorized data access, data alterations and data dissemination. Thirdly, adequate measures must be taken to ensure that publication precludes disclosure of individual data. Lastly, data received by SN may not be passed on to other persons than those charged with carrying out the responsibilities of SN. As an exception to the latter rule, SN is allowed to release micro data to other institutions for the purpose of statistical or scientific research [23]. The law defines which institutions are considered to perform statistical or scientific research, but other organizations and institutions may apply for authorization; the Central Commission for Statistics, the independent supervisory body of SN, must agree with the application and authorize SN to release micro data to the applicant concerned.

In addition to the Statistics Netherlands Act, SN is required to comply with general legislation on privacy protection. In 1988, the Act on Personal Data Registrations (WPR) was adopted as the first legislation regulating the maintenance and use of registers containing personal data. In 2001 the WPR was replaced by the Netherlands Data Protection Act (WBP), the enforcement of which is supervised by the so-called Data Protection Authority (DPA). The WBP states that personal data shall not be processed in a way that is incompatible with the original purpose of the data collection. However, an exception is made for further processing of personal data for historical, statistical or scientific purposes. It is this exception that allows SN to process administrative data as well as to pass on statistical registers, under strict conditions, to other institutions. Furthermore, the WBP states that personal data shall not be retained for a longer period of time than necessary for the realization of the purposes for which the data have been collected. Again, an exception is made for historical, statistical or scientific purposes. Obviously, the afore-mentioned exceptions apply under the explicit condition that adequate measures are taken to safeguard privacy and prevent the use of personal data for other than the stated purposes. Lastly, all process-

ing of personal data must be reported either to the DPA or to a data protection officer appointed within the organization that processes the data. Within SN, such a data protection officer has been appointed to supervise the application of, and compliance with, legislation on privacy protection.

To achieve a sufficient level of data security, measures are taken on the following points:

- As explained previously, personal identifiers are removed from the statistical data and replaced by the common PIN and AIN linkage keys. Although this primarily allows straightforward combination of statistical registers on persons, it additionally constitutes a major privacy protection measure as the linkage keys are anonymous and therefore preclude direct identification.
- The second measure is the restriction of access rights to the central data library. Only staff members who need data from the SSD in order to execute their task are given access to the part of the network on which the SSD is stored. Moreover, access is limited to the required data. As such, access to the data library does not automatically mean access to all data. Finally, those who have access to the SSD do not simultaneously have access to other parts of the network.
- The third measure entails a limitation of e-mail facilities, which minimizes the probability that unsecured data leave SN. The staff who have access to the data library do not have the right to send e-mails with attachments.
- People entering the office buildings of SN are subjected to strict checks. Each staff member has an identity card enabling him or her to enter the building. Visitors have to check in at reception and identify themselves with an official document and must be accompanied by a staff member at all times.

2.4.4. Combining statistical registers and output production

Because statistical registers stored in the SSD are standardized and provided with common linkage keys, combining registers for output purposes is a fairly straightforward procedure. For example, persons can be linked to their households, to the job they are in, and to the dwelling in which they live using the linkage keys HIN, PIN and AIN respectively. Nevertheless, generating the software to carry out the necessary selection and linkage steps can still be time consuming and error sensitive. Moreover, users need to gain knowledge about the contents of the SSD before selections and linking steps can be considered at all. The high degree of standardization of the SSD has enabled the development of software tools which greatly facilitate these steps in the production of output. Two of these tools are of particular importance and will be discussed here.

The first tool generates a table of contents of the SSD library: an inventory of available statistical registers including primary metadata (general description of the register and record description), organized by the responsible organizational unit. Needless to say, the table of contents is a very valuable starting point for users of SSD data.

The second tool is used to generate a required data set. The user specifies the required population and variables. Using these specifications, the tool consults the SSD metadata repository to determine which statistical registers need to be joined on the basis of which linkage keys. It then carries out the actual selections and record linking steps and delivers the desired dataset including a record description and some additional documentation. The analysis of the dataset and the subsequent production of output is done using software that best suits the particular output objectives or best matches the expertise of the user.

3. Examples of output based on the SSD

This section presents four examples of SSD-based output. The first two are illustrations of output based exclusively on integral register data. These are included to demonstrate the possibilities offered by coordinated, integral data from the SSD, namely portraying small groups in society and low-incidence phenomena as well as longitudinal analyses. Although surveys are considerably less suitable for such analyses, they do have their own important merits: a rich variable content tailored to the underlying statistical objectives. Obviously, combining the advantages of survey and register data presents numerous interesting possibilities. The third example was included to demonstrate these. The last example illustrates the use of SSD data by external researchers.

3.1. Example 1: Co-residence of parents and older adult children [35]

Using longitudinal SSD data for 2003–2005 on all adult children aged 30–40 years living in the Nether-

Table 2
Transition to co-residence with parents in 2005 of children aged 30–40 years

	%	abs.
No transition	99.49	1,879,764
Child moved in with parents (s)	0.43	8,124
Parents (s) moved in with child	0.06	1,134
Child and parents (s) moved into new home	0.02	378

lands and their parents, almost two million persons altogether, the authors investigated the extent to which intergenerational co-residence is determined by situations and events associated with the support needs of either generation. They distinguished between four possibilities: no transition, the child moved in with parent(s), parent(s) moved in with the child, child and parent(s) moved into a new home (Table 2). Because moving in with parents or children is a rare event, such an analysis would not be possible using a general survey. Only because the SSD contains data on the entire population, can such rare events be studied. The authors' general conclusion based on cross-sectional data was that support needs of both generations are important, but parents give more support to their children than vice versa. A weak socioeconomic position of both generations is positively correlated to co-residence. Longitudinal analyses show that events like divorce and income loss of the child increase co-residence. The person in need is most likely to move in with the other generation.

3.2. *Example 2: The effect of becoming unemployed on relationship stability [50]*

In this study, married male employees were followed for three years after becoming unemployed in order to establish whether their relationships were dissolved. Of particular interest was the distinction between three documented reasons for dismissal: dismissal because of long term illness, on personal grounds or due to redundancy. In case of redundancy, employees are selected for dismissal through some form of 'last in first out' principle, often per age class. As such, personal performance is not taken into account in contrast to dismissal on personal grounds and because of long-term illness. The results are summarized in Fig. 4. The most notable finding is that dismissal because of redundancy does not significantly enhance the probability of relationship dissolution whereas dismissal because of long term illness, and particularly dismissal on personal grounds, are associated with strongly increased dissolution rates. These results held out when an elaborate set of control variables, such as household income, number of children, duration of the relationship, were added and analyzed multivariately. The results therefore suggest that personal factors underlie both the dismissal and the dissolution of the relationship in the case of employees who were dismissed on personal grounds or because of long-term illness. After all, if dismissal in itself would negatively affect a relationship, dismissal due to redundancy would be expected to increase dissolution rates as well.

Table 3
Dropping out of school versus criminal behavior

	Suspected of a crime in the period 1999–2006
Overall	8.2%
Drop outs	37.6%

3.3. *Example 3: Dropping out of school and criminal behavior [43]*

For this study, data from the secondary education pupil cohort 1999 (VOCL'99) were enriched with data on police arrests and school careers from the SSD. The VOCL'99 is a panel survey which records the school careers of 17,000 pupils who started their secondary education in school year 1999/2000. It additionally documents an extensive set of background information, such as cognitive skills, home situation and social environment. A preliminary study demonstrated that dropping out of school is associated with strongly increased rates of criminal behavior ([42], Table 3). The mechanisms underlying this relationship were investigated by application of an elaborate multivariate model. The results indicated that social bonding, particularly school performance, reduced the risk of criminal behavior whereas early school-leaving and prior criminal behavior strongly increased this risk. The results supported the notion that both social control and self-control explain participation in risky behavior by youths.

3.4. *Example 4: Evaluation of governmental income support policies [18]*

In 1965, income support benefit was introduced in the Dutch social security system. This offered a government-funded minimum income protection. The Dutch income support system was revised recently to place a stronger emphasis on getting people (back) into jobs [49]. In this context, municipal government has been given more responsibilities. This example covers a study aimed at evaluating the income support pol-

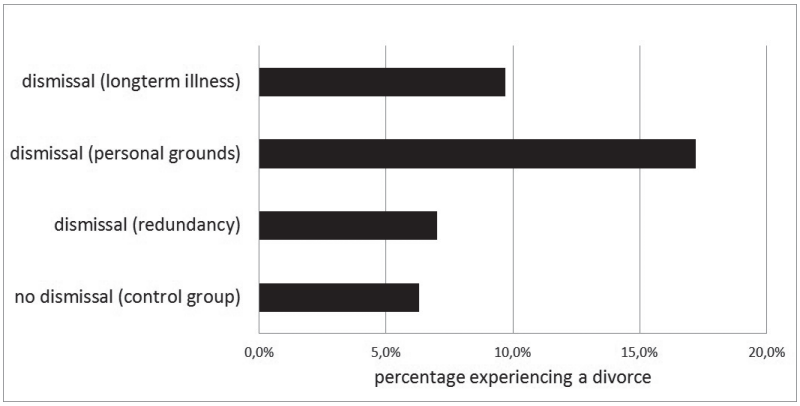

Fig. 4. Rate of relationship dissolution during three years after dismissal and by reason for dismissal.

icy carried out in one municipality, Enschede. One of the questions addressed in this study was to what extent persons whose application for income support was turned down were, in the long-term absorbed into employment and/or shielded from income support. For this purpose, data from the SSD on employment and social security benefits were linked to data held by the municipality. Although a significant part of those whose claim had been turned down found a job within three months (28 percent), in most cases their employment was not enduring. In addition, a significant number were claiming income support later on (25 percent). These, and other, findings were turned into recommendations aimed at improving the municipality's income support policy.

4. Quality issues

A description of the SSD would not be complete without discussing its quality. NSIs generally highlight the following dimensions of output quality [38,44]: *relevance*, *timeliness*, *accuracy and reliability*, *comparability and coherence*, *accessibility and clarity*.

Of these quality dimensions, accuracy and reliability are the most important and methodologically the most challenging ones. Both are measures of uncertainty, where accuracy reflect the systematic error and reliability the random error. Users of statistical data are used to interpreting confidence intervals as a measure for the reliability of survey outcomes. However, no similar measure for combined administrative register data is available. In order to fill this gap, new theory-based research has to be started. Bayes' theory or the super-population theory may open up some avenues for further research.

Another methodological question is how to determine the accuracy of variables from administrative register data. Based on the classical test theory, Bakker [5] determined the quality of some register variables by linking a survey to administrative register data and using structural equation models to compute the indicator validity However, this indicator validity does not distinguish between accuracy and reliability. More research following Saris and Andrews [30] still has to be done to assess the accuracy of administrative register data.

Another possibility to develop a measure for uncertainty is to combine all error sources described by Bakker and Daas [6] and Zhang [53]. However, their framework still has to prove in practice whether it is complete and correct. It should be helpful in the development of work processes in statistical offices for statistics based on administrative register data. In particular, the order of the steps should be proved. Whether the distinguished steps lead to an optimal result in terms of quality and cost is still open to discussion.

Lastly, the SSD data are assumed to cover the entire population, while in fact the definition of the population is restricted to the registered population. Only persons registered in the Population Register are considered to be part of the target population of statistics on persons. However, we know that this does not reflect the actual situation. To estimate the total population size, and therefore the under-coverage of the Population Register, a number of capture-recapture methods could be used. Most of these methods are based on strong assumptions, which are contravened [9,16,20, 45]. More research is needed to optimize these methods and relax the required assumptions.

As for the remaining quality dimensions, a prominent problem that needed to be addressed in the SSD

was that a specific version of a statistical register represents a trade-off between the various quality dimensions. For instance, a register which allows for the production of timely statistics does not simultaneously allow for maximum attainable accuracy and consistency because of various constraints: time constraints associated with the production of timely statistics, budget constraints, and constraints with regard to the availability of other registers which can be used for micro-integration. In a similar vein, comparability and consistency preclude timeliness due to the additional processing time associated with the necessary micro-integration. Moreover, increased relevance may reduce comparability over time. The latter is the case when, for example, an improved definition of a statistical unit cannot be implemented retrospectively due to limitations in those registers that relate to earlier periods. Introducing various versions of a statistical register will do the best justice to all quality dimensions, with each version emphasizing particular quality dimensions or combinations of dimensions. However, this reduces output consistency and poses the problem of the number of versions and the associated costs. To complicate matters, because statistical registers are interdependent in a register system, decisions regarding versions and revisions should be coordinated across the register system. After all, enhancing the accuracy in one particular register may reduce the consistency of the register system. In other words, an integral approach to quality management is necessary. Such an approach requires an organization that supports extensive cooperation between units and, more importantly, a cultural change towards a structure in which cooperation over organizational boundaries as well as a notion of collective ownership of statistical registers is seen as inherent [27,51]. For the SSD, such an approach is still in its infancy. Although an organization and procedures aimed at coordination have been put in place, many details still need to be worked out and, perhaps more importantly, more effort needs to be invested in changing the mindset of employees and management in relation to documentation and data sharing.

5. Conclusion

The development of the SSD has had a significant impact on the way statistics are produced within SN. In the past, organizational units used relatively autonomous production systems. As a result, statistical registers were scattered within SN and were not standardized. This complicated data sharing, output coordination, micro-integration of different registers, as well as the provision of data to external researchers. Today, all production systems converge on the SSD, a central system aimed at data sharing and coordination. The SSD comprises a library of standardized and linked statistical registers, as well as an organization which has been put in place to control various aspects of the system. It has made an invaluable set of registers readily accessible to internal and external researchers and has resulted in many valuable publications and will continue to do so. That said, the development and implementation of the SSD has not been a smooth process. The developers had to overcome many obstacles and still face quite a few more. Perhaps the most prominent are the development of an adequate coordination structure and the realization of a cultural change towards one which embraces collective ownership and data sharing. Within SN, this is an ongoing process. We hope that others will benefit from this article, be they researchers who use data from register-based systems or employees of NSI's which are developing a register-based production system.

The SSD will continue to expand, thus opening up more and more possibilities for statistical research. Firstly, existing registers are updated on a regular basis thus extending time series which in turn enable more extensive longitudinal analyses. Secondly, new registers are frequently added thus expanding the scope of the SSD. Thirdly, survey data are being included so as to facilitate the combination of survey and register data which, as discussed previously, presents numerous interesting possibilities. A case in point is the Labor Force Survey, which was recently added to assess the merits of inclusion of survey data. Finally, a perhaps not too distant prospect is the inclusion of "Big Data", a term that refers to the huge quantity of high frequency digital data captured by digital devices: call logs, mobile-banking transactions, online user-generated content such as blog posts and Tweets, online searches, satellite images, etc. [25]. Two examples serve to illustrate the tremendous volume of this type of data: approximately 80 million traffic loop detection records and around 1 million public social media messages are generated each day in the Netherlands [12]. If certain types of Big Data can be linked to persons, these data potentially constitute a major addition to the SSD content, possibly enabling the formation of snapshots of the well-being of (sub-) populations at high frequency. That said, many challenges pertaining to legislation, privacy protection, financing,

methodology and technology will have to be taken up and brought to a favorable conclusion before such a scenario becomes reality.

References

[1] P.G. Al and J.W. Altena, Data security, privacy and the SSB, *Netherlands Official Statistics* **15** (2000), 47–50.

[2] P. Al and B.F.M. Bakker, Re-engineering Social Statistics by micro-integration of different sources: An introduction, *Netherlands Official Statistics* **15** (2000), 4–6.

[3] K. Arts, B.F.M. Bakker and E. van Lith, Linking administrative registers and household surveys, *Netherlands Official Statistics* **15** (2000), 16–22.

[4] B.F.M. Bakker, Micro-Integration, The Hague/Heerlen: Statistics Netherlands, 2011.

[5] B.F.M. Bakker, Estimating the validity of administrative variables, *Statistica Neerlandica* **66** (2012), 8–17.

[6] B.F.M. Bakker and P. Daas, Some Methodological Issues of Register Based Research, *Statistica Neerlandica* **66** (2012), 2–7.

[7] J. Bethlehem, F. Cobben and B. Schouten, Handbook of nonresponse in household surveys, New Jersey: John Wiley and Sons, Hoboken, 2011.

[8] W.A. Belson, *Validity in Survey Research*, Gower, Brookfield, 1986

[9] Y. Bishop, S. Fienberg and P. Holland, Discrete multivariate analysis, theory and practice, New York: McGraw-Hill, 1975.

[10] N.M. Bradburn, Presidential address: A response to the nonresponse problem, *Public Opinion Quarterly* **56** (1992), 391–397.

[11] P. Daas, S. Ossen, R. Vis-Visschers and J. Arends-Tóth, Checklist for the quality evaluation of administrative data sources. Discussion Paper No. 09042. The Hague/Heerlen: Statistics Netherlands, 2009.

[12] P.J.H. Daas, M.J. Puts, B. Buelens and P.A.M. van den Hurk, Big Data and Official Statistics. Paper presented at the New Techniques and Technologies for Statistics conference, Brussels, 2013.

[13] A. De Bruin, J. Kardaun, F. Gast, E. de Bruin, M. van Sijl and G. Verweij, Record linkage of hospital discharge register with population register: Experiences at Statistics Netherlands, *Statistical Journal of the United Nations ECE* **21** (2004), 23–32.

[14] E. De Leeuw and W. de Heer, Trends in Household Survey Nonresponse: A Longitudinal and International Comparison, in: *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little, eds, New York: Wiley, 2002, pp. 41–54.

[15] P. De Winden, K. Arts and M. Luppes, A proposed model for microintegration of economic and social data, Discussion Paper No. 08005. The Hague/Heerlen: Statistics Netherlands, 2008.

[16] S. Fienberg, The multiple recapture census for closed populations and incomplete 2k contingency tables, *Biometrika* **59** (1972), 591–603.

[17] T. Gelsema, The Organization of Information in a Statistical Office, *Journal of Official Statistics* **28** (2012), 413–440.

[18] J. Gravesteijn, N. de Jong and M. Spijkerman, Instroombeleid Gemeente Enschede. Eindrapport. In opdracht van de Rekenkamercommissie van de gemeente Enschede. SEOR Erasmus School of Economics, Rotterdam, 2011. [in Dutch]

[19] C. Harmsen and A. Israëls, *Register-based household statistics,* Paper presented at the European Population Conference, 26–30 August 2003, Warsaw, Poland.

[20] International Working Group for Disease Monitoring and Forecasting, Capture-recapture and multiple record systems estimation. Part I. History and theoretical development, *American Journal of Epidemiology* **142** (1995), 1059–1068.

[21] C.S. Kamen, The 2008 Israel Integrated Census of Population and Housing. Basic Conception and procedure, Jerusalem: Central Bureau of Statistics, 2005.

[22] G. Kim and R. Chambers, Regression Analysis under Probabilistic Multi-Linkage, *Statistica Neerlandica* **66** (2012), 64–79.

[23] P. Kooiman, J. Nobel and L. Willenborg, Statistical data protection at Statistics Netherlands, *Netherlands Official Statistics* **14** (1999), 21–25.

[24] F. Linder, D. van Roon and B. Bakker, Combining data from administrative sources and sample surveys; The single-variable case. Case study: Educational Attainment, Eurostat, Luxembourg, 2012.

[25] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Hung Byers, Big data: The next frontier for innovation, competition, and productivity, McKinsey & Company, 2011.

[26] J. Neter, E.S. Maynes and R. Ramanathan, The effect of mismatching on the measurement of response error, *Journal of the American Statistical Association* **60** (1965), 1005–1027.

[27] N. Ploug, Recent Developments in the System of Register Based Social Statistics in Denmark. Paper for DGINS conference September 2011 in Wiesbaden.

[28] C.J.M. Prins, Dutch population statistics based on population register data, *Maandstatistiek van de Bevolking* **48** (2000), 9–15.

[29] M. Sadinle and S.E. Fienberg, A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems, Accepted in *Journal of the American Statistical Association*, 2013.

[30] W.E. Saris and F.M. Andrews, Evaluation of Measurement Instruments Using a Structural Modeling Approach, in: *Measurement Errors in Surveys,* P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, eds, New York: John Wiley & Sons, 1991.

[31] E. Schulte Nordholt, Introduction to the Dutch Virtual Census of 2001, in: *The Dutch Virtual Census of 2001,* E. Schulte Nordholt, M. Hartgers and R. Gircour, eds, The Hague/Heerlen: Statistics Netherlands, 2004.

[32] E. Schulte Nordholt, M. Hartgers and R. Gircour, eds, The Dutch Virtual Census of 2001, The Hague/Heerlen: Statistics Netherlands, 2004.

[33] E. Schulte Nordholt and F. Linder, Record linking for Census purposes in the Netherlands, *Statistical Journal of the IAOS* **24** (2007), 163–171.

[34] R. Seljak, *Integrated Statistical Systems and Their Flexibility – How to Find the Balance?* Paper NTTS 2012, Brussels.

[35] A.W.M. Smits, C.H. Mulder and R. Van Gaalen, Parent-child coresidence: Who moves in with whom and for whose needs? *Journal of Marriage and Family* **72** (2010), 1022–1033.

[36] Statistics Finland, Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland, Helsinki: SF, 2004.

[37] Statistics Netherlands, *Statistics Netherlands Act November 2003,* The Hague/Heerlen: Statistics Netherlands, 2004.

[38] Statistics Netherlands, *Statistics Netherlands' Quality Declaration*. The Hague/Heerlen: Statistics Netherlands, 2012.

[39] I. Stoop, *The Hunt for The Last Respondent*, The Hague: SCP, 2005.
[40] I. Stoop, J. Billiet, A. Koch and R. Fitzgerald, Improving survey response. Lessons learned from the European Social Survey, John Wiley & Sons, Chichester, 2010.
[41] P. Struijs, A. Camstra, R. Renssen and B. Braaksma, Redesign of Statistics Production within an Architectural Framework: The Dutch Experience, *Journal of Official Statistics* **29** (2013), 49–71.
[42] T. Traag, O. Marie and R. van der Velden, Risicofactoren voor voortijdig schoolverlaten en jeugdcriminaliteit, *Bevolkingstrends* (2010), 55–60.
[43] T. Traag, Early school-leaving in The Netherlands. A multidisciplinary study of risk and protective factors explaining early school-leaving, Phd-Dissertation, University Maastricht, Maastricht, 2012.
[44] United Nations Economic Commission for Europe, Register based statistics in the Nordic countries Review of best practices with focus on population and social statistics, United Nations, New York and Geneva, 2007.
[45] P.G.M. Van der Heijden, J. Whittaker, M.J.L.F. Cruyff, B.F.M. Bakker and H.N. Van der Vliet, People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates, *The Annals of Applied Statistics* **6** (2012), 831–852.
[46] J. Van der Laan, C. Harmsen and L. Kuijvenhoven, Deriving longitudinally consistent household Statistics from register information, Paper 57^{th} session of the International Statistical Institute, Durban, 2007.
[47] G. Van der Veen, Integration of microdata from business surveys and the social statistical database, Paper DGINS 2007 The Hague/Heerlen: Statistics Netherlands, 2007.
[48] J. Van Maarseveen, The Dutch virtual Census of 2001 compared to previous Censuses, in: The Dutch Virtual Census of 2001, E. Schulte Nordholt, M. Hartgers and R. Gircour, ed., The Hague/Heerlen: Statistics Netherlands, 2004.
[49] W. Van Oorschot, The Dutch welfare state: recent trends and challenges in historical perspective, *European Journal of Social Security* **8** (2006), 57–76.
[50] J. Van Rooijen and R. Van Gaalen, Het effect van ontslag van de man op zijn scheidingskans, *Tijdschrift voor Arbeidsvraagstukken* **29**(4) (2014), 414–425.
[51] A. Wallgren and B. Wallgren, Register-based Statistics: Administrative Data for Statistical Purposes, John Wiley and Sons, New York, 2007.
[52] L.-C. Zhang, A Unit-Error Theory for Register-Based Household Statistics, *Journal of Official Statistics* **27** (2011), 415–432.
[53] L.-C. Zhang, Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica* **66** (2012), 41–63.
[54] L.-C. Zhang and C. Hendriks, Micro-integration of register-based Census data for dwelling and household, Paper presented at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, 2010.