Distr.: General
18 April 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 3: Availability, completeness and quality of data from registers and other sources

The register of addresses and buildings – a combination of different registers

Note by Federal Statistical Office, Germany[1]

Summary

The great challenge posed by the Census 2011 in Germany is combining the various survey components so that unequivocal data records containing the respective census variables can be produced in the end. As data sources are heterogeneous – there are registers, complete enumerations and the sample survey – it is necessary to ensure communication between the census components. To this end a basic register was set up containing all addresses with housing space and occupied living quarters in Germany. In the following the different data sources that were used to build up this basic register are described. The preprocessing of the data including the harmonization and standardisation will be illustrated and the data verification will be explained in this paper.

I. The model of the Census 2011 and the core register

1. With reference day of 9 May 2011, the first population census in more than 20 years is conducted in Germany. The current population data are based on updated results of the

[1] Prepared by Susanne Stiglmayr.

.

Please recycle

last population censuses and, consequently, are getting more and more inaccurate. While in 1987 in the Federal Republic of Germany and in 1981 in the German Democratic Republic, respectively, a complete enumeration of all persons (conventional census) was conducted, the 2011 Census will be complete too, but register-based. The reasons for the methodical change are particularly to achieve the goal of minimising the burden on the population. The use of administrative register allows evaluation of all data without having to contact the citizens directly. Furthermore the new method of the 2011 Census reduces costs. On the other hand a highly developed information technology is necessary for handling and evaluating these very large volumes of data.

2. With a conventional census it is possible to include all the information that is relevant for the census in one questionnaire and, consequently, in one data set. Using the register-based method in Germany, however, the information is available in separate data sets, which are collected separately and which do not have any standard identification number for either persons or addresses. As data sources are heterogeneous – in the end there are registers, complete enumerations and the sample survey – it is necessary to ensure communication between the survey components. For that purpose a basic register was set up containing all addresses with housing space and occupied living quarters in Germany. This register of addresses and buildings (Anschriften- und Gebäuderegister "AGR") works as a connecting element in the entire census model and thus ensures that it will be possible to link the census variables. The functions of the AGR have been described in detail in Section 2 Subsection 2 of the 2011 Census Preparation Act:

The register of addresses and buildings serves

1) to steer the processes pertaining to the census of buildings and housing and the flow control of all primary data surveys forming part of the overall census,

2) to prepare the sample surveys to be conducted as part of the census and is used as their sampling frame,

3) to co-ordinate the census surveys, to combine the data stemming from various sources during the census implementation phase, and to check whether the buildings, dwellings and persons to be covered by the census are complete,

4) to develop a system of spatial analyses and presentations of statistical results, and to create a basis for evaluating census results by small regions.

3. To enable the AGR to fulfil its central role of controlling the 2011 Census, the AGR needed to be extended and used during the preparation phase already. For that purpose, various existing administrative registers were acquired, the data edited, linked at address level in standardised form and aggregated to a complete data stock. The latter was used as a basic register for control and co-ordination purposes when the survey is prepared. The AGR contains all addresses with housing space and occupied living quarters; this means all addresses where there are persons and buildings of relevance for the census. Together with the respective addresses, various characteristics and variables are stored which contain further information on the address.

II. The set up of the register of addresses and buildings

4. For the 2011 Census in Germany, a basic register is needed which combines the different data sources. As there is no unique cross-register identification number, the census uses the address in order to put the data sources in relation to each other. So the steering file of the census – the register of addresses and buildings – must contain all addresses of

occupied accommodations and housing space. Unfortunately such a register does not exist in Germany and has to be set up during the census preparation phase. To set up the AGR, three central registers of official statistics were used. The data source and further conditions are defined in Section 2 of the 2011 Census Preparation Act. The AGR is based on the data sources of the population registers (MR), the Federal Employment Agency (BA), and the geo-referenced address data of the Federation (GAB). Those data stocks were combined in order to completely cover the census-relevant addresses.

5. The population registers (MR) are conducted by the resident's registration offices. In Germany each municipality stored demographic and geographical information of the citizens who take up residence in the municipality. Further explanation you will find in the supporting paper "The population registers in Germany – the main data source in the Census 2011" for this meeting. For collecting all addresses of housing space the attribute "place of usual residence" of the population registers is important.

6. The register of the Federal Employment Agency (BA) contains not only address variables but also employment variables like activity status, occupation and location of place of work. Every employee in Germany is registered at the register of the Federal Employment Agency with his personal data like name, date of birth and address.

7. The third source of the AGR is the geo-referenced address data of the Federation (GAB). In contrast to the first two sources the geo-referenced address data of the Federation doesn't contain any personal information. The register includes every parcel of land including their geographic coordinates. This means every address of a building that is surveyed by the land register authorities of the Länder is included and combined with geographic coordinates. The file includes both occupied and unoccupied accommodations.

8. As the AGR is an aggregated data stock at the address level, the registers used to set up the AGR must have the same level of aggregation and same definition of address. The data of the Federal Employment Agency and the population registers are data on persons, i.e. an address may be available several times sometimes even with different spellings. Therefore, all three registers were first of all aggregated at the level of addresses. In the next step the definition of address has to be harmonised. Each used register has its own interpretation of address. On the one hand it is a result of the purpose of the registers. Primary the registers are not conducted for representing valid addresses. On the other hand the federal structure of Germany maintains the heterogeneous data stock. Decentralised preservation of registers, however, involves complex problems of data quality assurance. There is a lack of standardisation of data between registers and between individual collection authorities in the Länder and municipalities because the registers were not designed to be combined at a federal level. As a consequence, data on address variables are presented in different ways, which sometimes are due to personal preferences of the staff in charge and to differences in update processes. For example all address information at the register of the Federal Employment Agency is stored in only one variable and there in different forms. We defined a standard format of an address at the register of addresses and buildings by following properties:

1. *level of street*
    a) *Official Municipality Code*
    b) *postal code*
    c) *submunicipality*
    d) *street*
2. *level of address*
    a) *house number*

b) house number supplement.

9. Referred to this definition of an address every unique data set gets an own identification number (address-ID). This identification number is used during the whole census in all parts and data stocks of the census.

10. To harmonise data and to assure data quality of the AGR, various steps of data processing were performed in the data stocks of the GAB, MR and BA. The most important steps were the parsing and standardisation of address variables and the referencing of variables i.e. street names with external sources. Parsing here means that address data of specific registers were decomposed into address components. As the same address components were then available in a standard form in all registers, a major requirement for automated comparability across several data sources was met. In this way the addresses of the BA, which were supplied in one single string, could be decomposed in the standard address format. For this purpose, a list of rules was applied which, for example in their simplest form, provides for capital spelling of all street names. After parsing and standardisation, the data in each data stock were aggregated to street level and combined with another data stock through the municipality and street name. The purpose of that action was to confirm the correctness and existence of street names if they occurred in an identical form in at least two data sources. The volume of street names that could not be combined in an automated way was passed on to the statistical offices of the Länder for correction at screen. Once returned, the corrected data were integrated into the data stock in an automated manner, which considerably improved the quality of the data stock. As there are a number of addresses per street, this two-stage procedure – with separate handling at street level – can be applied to correct many addresses. Where the source of error is just a wrong street name, correcting the street name ideally allows combining the number of seemingly different addresses to form one address. In this way, correcting one street name can solve several address problems. Besides the referencing of street names other address variables had to be harmonised. While address identifiers as the official municipality code or the postal code can change over time, these variables had to be unified. Within one data source or between the different data sources the official municipality code or the postal code can come from different territorial classifications. Also they can change over time during the census phase. In both constellations the address identifiers had to be unified at the reference date of the 2011 Census.

11. At least the three prepared data sources MR, BA and GAB were merged into one basic table. The new database contained every address from the three register once a time. The core of the database is the address-ID with the address variables including the geographic coordinates. In the AGR the source of an address (MR, BA and/or GAB) is still observed.

III. The quality assurance of the AGR

12. As a steering file, the AGR thus has a central position in the census model. A uniform stock of addresses makes sure that the standardised notations of the AGR will be used also in other survey components und that due to the structure of the data base, addresses can be maintained only once and centrally in the AGR in the 2011 Census.

13. To verify the quality of the AGR several measures are intended.

14. Using three registers was not only due to the fact that census-relevant information had to be combined from different registers but it had also a function in data quality assurance. As the registers are separate sources, the combination of data could also be used as a validity indicator for an address. It could be assumed that an address occurring in the same version in separate files does exist.

15. As the census covers only accommodations and buildings with housing space, it is relevant for setting up the AGR whether a specific address is to be classified as "address with housing space". If there is no housing space, the address has to be excluded from further census activities. According to the rule, that addresses which are found in at least two of the origin sources (MR, BA and GAB) are addresses with housing space, the addresses found only in one of the three data sources had to be checked by the statistical offices of the Länder. It has to be checked whether or not it is an address with housing space. This verification is constituted in the 2011 Census Act in Section 14. The further process shows the necessity to control the attribute (housing space or not) of an address. Within the meaning of quality assurance the criterion housing space or not had to be checked continuous independent of Section 14 of the 2011 Census Act. The target with or without living space is changing by and by during the census preparation phase and had to be controlled for all addresses over time.

16. Against this background a specific application is designed for updating the AGR. The central application affords the opportunity of editing the address variables of the AGR. As data such as street names can change, continuous adjustment of the register data has to be part of quality assurance. Furthermore the criterion occupied and unoccupied can be changed with the application. The application also provides the feature of adding new addresses to the AGR. As the AGR has to reflect all addresses with housing space and occupied accommodations in Germany, the integration of new occupied accommodations and buildings with housing space is another feature of the application. The Federal Statistic Office as well as the statistical offices of the Länder work on this application to update the AGR.

17. A lot of update-processes of the AGR are being implemented. Therefore, keeping the AGR not only up to date but also complete requires the integration of changes after the basic stock has been set up. Such updating is ensured, first, by integrating several update deliveries from 2008 to 2011 and, second, by linking further survey components. Data of the population registers and the geo-based address data of the Federation are repeatedly delivered to the Federal Statistical Office and are integrated into the data stock. By obtaining new addresses in this way, the completeness of the AGR is ensured. Other gains in information in the AGR are achieved by linking up other survey components. By linking up the register of collective living quarter and institutional household addresses the AGR will be completed. In this way new addresses can be integrated in the AGR. And by linking up the register of the census of buildings and housing the AGR will be improved, too. Linking the owners of buildings to all occupied accommodations and buildings with housing space has two purposes: First, the census of buildings and housing is checked for completeness because an owner has to be found for each census-relevant address contained in the AGR. Second, addresses from the census of buildings and housing can be added to the AGR if they exist and have not been included yet in the stock.

18. At least during the census phase itself the AGR can be improved. The survey components of the census update the stock of address data by integrating their survey results. For example a property owner gets a questionnaire for the census of buildings and housing for a certain address and the property owner declares further properties at the questionnaire. Using this information the AGR can be updated also during the census phase. The results of the sampling survey also improve the AGR. For example an enumerator finds out that there are addresses without any building. Such information is integrated in the AGR. Not necessarily new addresses can be discovered but primarily the criterion occupied versus unoccupied housing space can be changed.

IV. Conclusion

19. For the purpose of combining the different components of the 2011 Census in Germany, a basic register was needed. While such a register does not exist in Germany it came to the decision to set up a core register of addresses. Using this basic register all components of the census of the population and of the census of buildings and housing can be linked together by the address. Combining the data from the different data sources at the level of addresses was one of the biggest challenges of the 2011 Census. The experience acquired with that kind of combining data without numerical identifiers in the 2011 Census leads to concrete proposals for the further development of the new German Census. It is obvious that the German statistical offices would like to take efforts to arrive at a situation where the data on buildings, dwellings and households can be collected on the basis of registers, which would mean rapid and low-cost data collection. The wide experience in setting up the AGR will help us to reach this aim.