The new register-based Census of Germany - a multiple source mixed mode approach

Dr. Sabine Bechtold[1]
[1] Federal Statistical Office, Wiesbaden, Germany, e-mail: sabine.bechtold@destatis.de

Abstract

The 2011 Census in Germany was based on a new data collection method. Some experiences made during the process will be presented. The census model was a multiple source mixed mode method that collected data from administrative registers such as population registers, full enumerations and a sample survey. Multiple sources were necessary to cover up the compulsory census data set of the European Union. The paper will focus on the determination method for the number of inhabitants. To create this census variable the different data sources were used to ensure the data quality. Primarily, this variable was fixed with the information stored in the population registers. These registers are organised in a decentralised way and because of that first a centralised data set with population register data had to be built up and had to be corrected. Another aspect to assure the quality of the results was the use of information from primary surveys. With this information overcoverage and undercoverage within the register data could be measured and corrected statistically. That required that all information based on one of the census components (register, household sample survey and complete enumerations) had to be merged into one central data base based on the data sets from the population registers. The biggest challenge posed to German official statistics by the new census model was the fact that data had to be combined without a uniform personal identification number (ID) and a uniform building ID.

Keywords: Number of inhabitants, administrative registers, quality assurance via sample survey, record linkage

1. The Census Model in Germany

Contrary to the last population census, the 2011 Census was conducted with a new method. To achieve the goal of minimizing the burden on the population, administrative registers were used wherever possible. The register-based census consisted of three major components providing information on the population as well as on buildings and dwellings in Germany. Those components were administrative registers, complete enumerations of population living in residential establishments and collective living quarters (special-facility addresses) and of buildings and dwellings, and a household sample survey. Other variables, which could not directly be covered by the register-based census, were obtained by statistical methods applied in the household generating procedure. The generating process combined information of all components of the census and compiled new variables like the type of household or information about the family structure.

The complex model in Germany could only be realized by using a central data stock that had to be built up. This central data stock allowed the combining and the coordination of the data. On one hand a basic register was set up containing all addresses with housing space and occupied living quarters in Germany. This register of addresses and buildings (AGR) worked as a connecting element in the entire census model and thus ensured that it was possible to link the census variables. On the other hand a reference data stock based on personal data from the population registers was a

central steering file for the organisation and coordination of the different census components.

2. The determination of the variable number of inhabitants

a. The use of administrative registers

The number of inhabitants mainly was derived with the information stored in administrative registers. Data provided by surveys ensured the quality of the variable. In the 2011 Census two main registers were needed in order to implement the model. One register contained data sets on the level of addresses, one on the level of persons. Both together represented the reference data base of the census.

During both the census preparation and implementation phase, addresses served as the basis for coordinating data sources and for establishing links between the survey components. To this end, a central register of addresses and buildings, the AGR, was set up as a data base into which every address was entered only once. To set up the AGR, three central administrative registers of official statistics were acquired, the data edited, linked at address level in standardised form and aggregated to a complete data stock. The AGR was based on the data sources of the population registers, the Federal Employment Agency, and the geo-referenced address data of the Federation. Those data stocks were combined in order to completely cover the census-relevant addresses. As the AGR was an aggregated data stock at the address level, the registers used to set up the AGR had to have the same level of aggregation and same definition of address.

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 were separate sources, the combination of data could also be used as a validity indicator for every address. It could be assumed that an address occurring in the same version in separate files did exist. As the census covered only accommodations and buildings with housing space, it was relevant for setting up the AGR whether a specific address was classified as “address with housing space”. According to the rule, that addresses which were found in at least two of the origin sources were 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 had to be checked whether or not it was an address with housing space. The AGR was used in the census preparation phase for supporting the primary surveys of the Census of Buildings and Housing and for supporting the data collection at special-facility addresses. Also the sample survey needed information based in the AGR as the addresses with housing space contained in the AGR were the sampling frame.

Based on the addresses stored in the AGR personal information from the population registers was added in order to build up a reference data stock containing information both on addresses and on personal data. The census variable “number of inhabitants” was based on the information stored in the population registers as they contained information on every person registered with a place of residence in the Federal Republic of Germany. The population registers were managed by the residents’ registration offices that were legally obligated to transfer the data set to the official statistics for census purposes. The registered place of residence, according to population registration law, was stored there with the values of sole, main or secondary place of residence and, for the register-based census, corresponded to the compulsory EU variable of “place of usual residence”. In addition, that administrative register contained the basic demographic data for every individual (sex, age, marital status, citizenship, place and state of birth). These decentralized registers had to be consolidated by using the information on addresses and personal characteristics of each data set. First the address variables were adjusted with the AGR, second the

personal variables set up a population register linked with the AGR. In this way a temporary centralized data set with population register data for census purposes only was build up for Germany.

To increase the quality of the population register the centralized data stock had to be corrected to determine the number of inhabitants. Correctly representing the variable "place of usual residence" required adjustment of the register data stock by multiple or incorrect entries which were not allowed according to the population registration law but nevertheless happened. For this reason the central data stock of the population registers was being analyzed with the purpose to find incorrect registration. During checks for multiple registration, double counts and registrations that were incorrect in terms of registration law had to be identified and corrected in the data stock. As there was no personal ID stored in the population registers constant variables had to be compared in order to identify incorrect data sets. In the process of finding incorrect data sets the variables "name", "first name", "sex", "date of birth" and "place of birth" were used as mostly fixed demographic characteristics. For checking and correcting the combined data stock the different data sets had to be compared in order to find doublets, triplets etc. and singular data sets. The doublets were defined in a first step as identical in the main demographic characteristics. Because of the registers being organized in a decentralised way without an overall unique ID it was necessary to search for doublets with similar values too, especially in the variables "name", "first name" and "place of birth". In order to clear incorrect entries for persons only registered with one or more secondary places of residence it had to be clarified by survey at what address the relevant person actually had to be counted at the reference date. Also for persons with more than one main or sole place of residence the actual address had to be identified. For this purpose both a survey and an automatic procedure was implemented. The results were integrated into the personal data sets. Combined with the addresses of the AGR the personal data sets represented the reference data stock.

b. The use of surveys

To assure the quality of the number of inhabitants based on the population registers several surveys were implemented. For that purpose complete addresses were questioned in order to identify those persons who had to be counted. The determined inhabitants at an address were compared with the people registered at the address in order to find overcoverage and undercoverage.

Firstly, a complete enumeration was conducted at special-facility addresses such as prisons, psychiatric hospitals, care homes or student residences with a high fluctuation rate. These addresses were marked in the AGR. Due to the fact that, in such facilities, the quality of the population register data was often poor and that there was a systematic error a complete enumeration was implemented. Also because of the sensitivity of the personal data of residents living in specific areas such as prisons or psychiatric hospitals, specific protective measures had been arranged for the survey to ensure that information and data on the residents at those addresses were treated with due care. The information on inhabitants based on the enumeration was compared with the information stored in the temporary centralized data set with population register data and the addresses were corrected exactly. That means that persons could be confirmed, added or deleted at the special-facility address.

The household survey was the main primary survey in order to ensure the data quality of the number of inhabitants by quantifying and correcting overcoverage and undercoverage in the population registers. Another function of the sample survey was to cover variables and population groups that were not contained in registers. The household sample survey was conducted as a random sample and covered nearly 10%

of the population. The addresses classified as address with housing space and occupied living quarters stored in the AGR were used as sampling frame. In the sample design two forms of stratification were realized. On one hand geographical stratums were used. On the other hand the size of the addresses based on the number of inhabitants from the population registers was used as second stratification variable. As there was a complete enumeration in special-facility addresses these addresses were disqualified from the sampling. The addresses that got sampled were marked in the AGR. To improve the quality of the number of inhabitants every person living at the sample unit had to be determined. In that way, the persons living at an address and identified in the household sample survey could be matched with the persons registered at the address as shown in the population registers. That allowed identifying the extent of overcoverage and undercoverage in the population registers and allowed to correct the number of inhabitants statistically. This form of correcting the number of inhabitants was limited to municipalities with 10,000 and more inhabitants. The limitation was due to the fact that the quality of the population registers was higher in small municipalities. Because of the higher rate of overcoverage and undercoverage in municipalities with 10,000 and more inhabitants the increase of the quality of the number of inhabitants by using the results of the sample survey was essential. Another reason for the limitation of the use of the results of overcoverage and undercoverage was the fact that the sample size in small municipalities had to be nearly 100% in order to get reliable sample results. But with the use of the sample results in municipalities with 10,000 and more inhabitants it was possible to correct the population registers statistically by extrapolation. This form of statistical correction allowed us to improve the quality of the results.

The last survey to improve the variable number of inhabitants conducted in the census was a survey that took place in small municipalities (less than 10,000 inhabitants) and was limited on addresses with only one occupied living quarter. If there was a discrepancy between the number of inhabitants stored in the population registers and the number of occupants indicated by the property owners the address was marked in the AGR. It had to be clarified which people had to be counted at that address. A possible correction of the temporary centralized data set with population register data was done only at the questioned address; an extrapolation of the results was not intended. The limitation on small municipalities was due to the fact that in big communities the sample survey was used as corrective element for the temporary centralized data set with population register data.

c. Linking of the data sources

The different types of data sources – registers and survey data – had to be linked in order to generate a central data stock. Firstly, the information stored in the registers had to be linked. On one hand the AGR contained all addresses with housing space and occupied living quarters. Every address had a unique address-ID for identification. On the other hand the centralized data stock of the population registers included a personal ID-number for every personal data set. These ID-numbers were internal numbers generated only for the use in the census. As a linking procedure the data sets from the population register were linked to the addresses by using the address information stored in the registers with the effect that an address-ID had been added to every personal data set. The combination of addresses stored in the AGR and personal data sets stored in the temporary central register built up the reference data stock.

The addresses that had to be collected in the surveys were marked in the AGR. Also the information for the organization and implementation of the surveys such as geographical specifications was provided in the AGR. After the realization of the surveys the personal data sets collected during the several surveys had to be compared with the data sets stored in the reference data stock. For that purpose only the personal

data sets related to the address-ID of the respective survey were taken into account. As there was no central ID-number in Germany, neither for addresses nor for persons, the linking procedure had to be managed by using auxiliary variables; for the record linkage the variables name, first name, sex and date of birth were relevant. The data collection method was conducted by enumerators which had to identify every person living at the address. The enumerators documented the personal data of the respondents for each address in order to compare the information with the personal data sets linked to the corresponding address-ID as soon as possible. This allowed an early controlling system including additional enquiries if there was a great variation. As the enumerators documented the information in writing, some variations in the notation were possible. Because of the diversity and the difference in quality of the data sources of the census both identical and similar values of the auxiliary variables had to be considered during the linking procedure. Every data set of the central register could be confirmed either as up-to-date or as non-active. It was also possible that a missing data set had to be added to the central data stock in order to complete the results of the survey at the questioned address. In that way the overcoverage and undercoverage of the addresses could be determined. The results of the surveys were used to correct the registers in a statistical way only at the surveyed addresses with the exception that results from the household sample survey in municipalities with more than 10,000 inhabitants were used both for a correction of the number of inhabitants at the level of the sampled address and for a statistical correction of the number of inhabitants by extrapolation of the results.

Altogether for determining the number of inhabitants different parts of the census had to be taken into account. Based on the information stored in the population registers several methods were realized in order to assure the quality of this variable. In the following the different steps for deriving the number of inhabitants are listed.

- In a first step the population registers had to be consolidated to generate a temporarily central population register for Germany. This central register contained approximately 86 million data sets.
- The central register had to be corrected for incorrect entries. About 0.5 million data sets had to be checked. Either the type of place of residence had to be corrected or the data set had to be marked as irrelevant. The correction took place at the level of the personal data set.
- A complete enumeration at special-facility addresses was conducted. During this survey information about around 2 million people was collected. The survey contained information for the complete address.
- In municipalities with 10,000 and more inhabitants information about existing occupants was obtained from the household sample survey. About 7.9 million data sets had been exploited. The results of the survey had been compared with the central data stock in order to determine matching data sets, missing data sets and non-active data sets. This information had been extrapolated and corrected the registers in a statistical way.
- In municipalities with less than 10,000 inhabitants a survey based on discrepancies between the population registers and the information obtained from the owners of houses and dwellings was used to assure the quality of the number of inhabitants. The results of the field research were compared with the persons stored in the register and the data sets at the address were corrected. This survey determined information about around 1.4 million inhabitants.

3. Conclusions

The 2011 Census in Germany was a complex procedure requiring coordination and

communication between the individual survey components and data sources. Especially the differences in data quality between the registers and the data determined at surveys made it difficult to combine information. Also, the fact that there were no standard identifiers for persons or dwellings in Germany made data combination difficult. Therefore it was necessary to use auxiliary variables to ensure the linking of the records from the different data sources. First of all it was essential especially for population registers to set up a central data stock containing both addresses and inhabitants, which then had to be coordinated and linked with the other data sources. This central reference data stock was necessary to conduct the complex census model and to consolidate the different data sources.

Because of the complexity of the 2011 Census the time required for the consolidation of the different data sources increased compared to a less complex census model. But to ensure high quality in the census variables it was necessary to conduct the different elements of the census. The trade-off between accurateness of the variables and timeliness of data had to be taken into account and changed during the time. Around the reference date survey information was used to ensure the quality of the census variables. But the realization of the surveys also took time so the completion of the census results depended on the length of the collecting period. During this time the information based on the survey data improved the quality of the variables. Discrepancies between registers and survey data had to be clarified but with an increasing distance from the reference date further methods for obtaining plausible combinations were getting useless. So it had to be decided carefully how long the data collection period especially for additional enquiries should last in order to improve the quality of the number of inhabitants.

Another trade-off of the 2011 Census had to be considered by fixing the method. In order to generate reliable results with the sample survey the sampling fraction had to reach an optimal size. This was necessary to minimize the sampling error. On the other hand the population should be released from the obligation to respond or to be in touch with the census so the sample size had to be kept to a minimum. That means that the trade-off between the accurateness of the results of the sample survey and the burden on the population had to be taken into account.

Referring to the acceptance of the census results it is important to impart the knowledge about the process and the quality of the results. As the 2011 Census was realized with a new method especially the characteristics of the model and the results of this method have to be established. Due to the fact that the number of inhabitants as the main result of the census was obtained by extrapolation the acceptance might be lower because the interpretation of the results is more complex than in a complete enumeration. The reaction of the users and the amount of enquiries will show how the new model with its appraisal of results is accepted in the public.