

30-05-2007

Anonymisation of German Democratic Republic 1971 census data for use as public use micro data file

I. Base material

The 1971 census data are comprised of two sets of data: person data and dwelling data. The person data contain demographic information as well as information about the source of livelihood, education and training, employment and the composition of households. The dwelling data inform about the condition of buildings, occupation and facilities (heating, water supply, toilettes). The base data material comprises 16.4 million persons, 6.2 million households and 6 million dwellings.

II. Linking person and dwelling data

Person and dwelling data were linked through matching on the identification numbers of state (province), district, county, community, census station, census district, census section, building number and dwelling number. After linking the data, a consistency check was done using variables with the same content. Those were “number of principal residents in 1st household” and “number of children under age 17 in 1st household”. Records for vacant dwellings, auxiliary dwellings and dwellings used for other (non-residential) purposes were deleted because there were no person data.

III. Anonymisation methodology

1. Drawing sample

As the first step of the anonymization process, a 60% sample of households will be drawn using the end-digit-procedure. Each person in collective dwellings is regarded as a household. The sampling ensures that a potential data attack cannot be sure whether the sought-for person or household is in the sample.

The original records are sorted by

- state [Land],
- administration district,
- number of persons in household,
- county,
- community,
- census station,
- census district and census section (for private households) or running number of collective dwelling (for collective dwellings)
- running number of building in census section
- running number of dwelling in building (for private households) or running number of person in collective dwelling (for collective dwellings)
- running number of household in dwelling.

Subsequently, each dwelling [in the sorted file] receives a running dwelling number.

For the 25% draw, the last two digits of this dwelling number will be used (=X_i). The sample probability is 25 out of 100 or 1 out of 4. First, a number between 0 and 3 will be chosen randomly as Z. Starting with Z, 25 numbers X_i between 0 and 99 are chosen according to the formula

$$X_i = Z + i*4, \text{ with } I = 0, 1, \dots, 24$$

The dwellings with the last 2 digits combinations of X_i (0 – 99) will be included in the sample file.

2. Deleting regional information

As further anonymization step, all regional identifiers will be deleted with the exception of state [Land]. That includes administration district, county, community, census station, census district, census section, as well as the geographic identifier of the place of work for commuters. To allow identifying and linking uniquely buildings with their dwellings and households, buildings receive running numbers within states. (See non-systematic sorting).

A new variable “size of community” (vp60) will be added to classify communities as follows:

For the states Brandenburg, Mecklenburg-Vorpommern, Sachsen [Saxony], Sachsen-Anhalt [Saxony-Anhalt] and Thüringen [Thuringia], four size classes are used:

- 1 = below 2,000 inhabitants
- 2 = 2,000 to fewer than 10,000 inhabitants
- 3 = 10,000 to fewer than 50,000 inhabitants
- 4 = 50,000 and more inhabitants

For the state Berlin, only two classes are used:

- 5 = under 100,000 inhabitants
- 6 = 100,000 and more inhabitants

3. Deleting some variables

The characteristics “floor space of 3rd kitchen in dwelling” and “floor space of 4th kitchen in dwelling” are populated sparsely and consequently will be removed.

4. Non-systematic sorting

Regional information could be deducted from the order of records in the original files. To protect against that possibility, the data material will be sorted using a non-systematic approach (i.e. it cannot be replicated). Subsequently, non-systematic unique numbers will be assigned to the identification variables of building, dwelling, household, person number.

5. Collapsing of values

Collapsing of values will be undertaken in the following variables:

- vp21 - age:
“100” to “105” will be collapsed to “100 or older”
- vw23 - floor space of rooms in dwelling in 1/10 square meters:
“1700” to “5816” will be collapsed to “1700 or larger”
- vw24 - floor space of kitchens in dwelling in 1/10 square meters:
“448” to “449” will be collapsed to “449”;
“450” to “799” collapsed to “450 or larger”
- vw25 - floor space of 1st kitchen in dwelling in 1/10 square meters:
“360” to “497” will be collapsed to “360 or larger”
- vw26 - floor space of 2nd kitchen in dwelling in 1/10 square meters:
“230” to “497” will be collapsed to “230 or larger”
- vw41 - number of secondary residents in 1st household:
“12” to “24” will be collapsed to “12 or more”
- vw43 - floor space of rooms of 1st household in 1/10 square meters:
“1599” to “4599” will be collapsed to “1600 or larger”
- vw47 - number of secondary residents in 2nd household:
“7” to “8” will be collapsed to “7 or more”
- vw48 - number of rooms of 2nd household:
“7” to “9” will be collapsed to “7 or more”
- vw49 - floor space of rooms of 2nd household in 1/10 square meters:
“800” to “1800” will be collapsed to “800 or larger”

6. Removal of cases

Four households will be removed because they have characteristics which identify them as unique, or as one of two, cases in the total data material.