Developments in Survey Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 15, Ljubljana: FDV, 2000

Nonresponse and Socio-Demographic Characteristics of Enumeration Areas

Metka Zaletel[1] and Vasja Vehovar[2]

Abstract

The data from national registers (Central Register of Population, Tax Register, Register of Territorial Units, Register of Housing Units), Census '91 data and some other sources (e.g., Telephone Directory) were merged at the level of enumeration areas for the whole territory of Slovenia. There are about 9,000 enumeration areas with around 65 households each. The centroids of the building and the altitudes were also included as there exists a strong correlation between interviewing costs and altitude of the responding household. In addition to Census and administrative data the survey data from telephone and face-to-face surveys were also included, especially non-response, refusal and non-contacted rates, and the travel expenses of the interviewers. All these data were used to build the model of detailed geo-demographic stratification of the country. The model enables us to conduct more efficient sample designs, to minimize costs of surveys and to reduce non-response rates by adjusting interviewers' training for "more difficult" areas.

1 Introduction

Data from the national registers basically serve for administrative purposes. Their use for statistical purposes is often not enough explored. In the paper, we show a step which was undertaken at the Statistical Office of the Republic of Slovenia to incorporate data from the register in the efficient way to draw the samples of resident persons and households.

[1] Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia; Metka.Zaletel@gov.si

[2] Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia; Vasja.Vehovar@uni-lj.si

In the past, we have already tried to use the register data for the analysis of nonresponse. First we have concentrated on matching the register data and the survey data (with emphasis on the refusals and other nonresponse) on a personal level (Vehovar and Zaletel, 1995). The results were encouraging and proved that the register data should be exploited much more to help us to design and conduct our surveys more efficiently. Main results from the matching project on the personal level were:

- higher nonresponse rate among of single households, households living in big apartment buildings, households living in urban areas;
- no influence of household size on refusal rate;
- no influence of education on nonresponse and refusal rate.

Certain characteristics of nonrespondents were shown to be the same as in other countries, e.g., nonrespondents are more likely to be older, living in single households and in urban areas (Groves and Couper, 1993); on the other hand, some characteristics were country specific (interaction of response, urbanicity and education). Part of our research was also the calculation of the nonresponse bias, which was found to be very high (Rbias=5%) for the income variable, but surprisingly low for the unemployment rate.

Some of the above results were found also in studies performed in other countries: age, household size, urbanicity, type of dwelling (located in a big apartment building) tend to be fairly strong correlates of non-response in most of the countries (Bros, 1995).

Almost all samples of official surveys in Slovenia have the same sampling design - they are two-stage stratified samples and primary sampling units are usually enumeration areas. The post-survey adjustment for non-response is also quite similar for most of the surveys: weights are calculated at the level of primary sampling units. If the adjustment is done at the level of enumeration areas, perhaps we can also predict the non-3response at the level of enumeration areas. Another motivation for this idea are certainly the results from some of the countries (e.g., King, 1996; Groves and Couper, 1993) where the division of the country into small areas according to the socio-economic variables was made in advance. Then it was proven that the non-response rates vary across socio-demographic types of areas. We decided to generalise the idea: to build the socio-demographic types of enumeration areas according to the non-response rates achieved in some of the official surveys. This model would enable us to predict the non-response rates for similar surveys in the future.

There are about 14,000 enumeration areas (EA) in Slovenia with 45 households each on average. Unfortunately, some of the EAs are very small (with less than 30 inhabitants each) or even empty, especially in remote areas. This fact caused a lot of problems in the process of sample designing and selection. In 1996, we merged all small EAs with their larger neighbours. We ended up with 9,872 clusters of enumeration areas (CEA) with an average of 65 households. The problem of small

EAs vanished almost completely. Since 1996, the primary sampling units in the majority of official surveys are CEAs.

The Central Register of Population itself has already been serving as the primary sampling frame source, yet not all information was effectively used. Further, the data from other registers were also not exploited.

Our first step was to use the data from the Register of Territorial Units. Together with the Geographical Information System (GIS), the centroids of all sampling units were calculated as well as the average distance from the centre of municipalities. This enables the calculation of the expected travel costs of the interviewer. The average altitude was also attached to every sampling unit. The data from the following sources have also been incorporated: Census '91 data, Central Register of Population (CRP), Database on Employed Persons in Republic of Slovenia (DEP), Telephone Database (TD).

The next step in constructing the database was to attach the non-response data from the official surveys, conducted in the past at the Statistical Office of the Republic of Slovenia. The above constructed information system is basically GIS - System. However, its richness is extremely helpful in constructing the optimal stratification, finding designs with minimal field costs and adopting the frame to non-response problems.

In Section 3 the available data is introduced. In Section 4 we explain the analysis of given data in four steps. Finally the results of analysis and their advantages are shown.

2 Data used to build geo-demographic stratification

Our main sources of data used to build geo-demographic stratification were of course administrative data sources. At the same time we also used survey data to evaluate each step of the analysis of administrative data and to judge the importance of variables involved in the models. In this section we first describe all available administrative data sources, then we describe surveys used for evaluation, and, finally, we introduce variables used in later modelling.

2.1 Administrative data sources

All the major administrative data sources available at the Statistical Office of the Republic of Slovenia and some other institutions were used:

- Central Register of Population (CRP),
- Census '91 database,
- Database on Employed Persons in the Republic of Slovenia (DEP),
- Register of Territorial Units (RTU),
- Telephone Database (TD).

At this stage of research, the Taxation Register (kept by the Ministry of Finance) has not been included in the estimations, but when the TR is available, the model will be re-estimated.

A very important point, which needs to be stressed here, concerns the time distance from the Census '91. All data from the Census are obviously now 6 years old, but we took from the Census mostly data about the dwellings and migrations. The Slovenian population is very stable and only about 2% of population has been moving per year till now. In fact, most of those 2% are migrations within the same towns or villages. The situation about dwellings has not changed much in Slovenia since 1991 since not a lot of new dwellings have been build in-between. We can assume that the Census data are good enough for our purposes.

2.2 Surveys

We included the following surveys:

- Labour Force Survey 1994, 1995, 1996 (LFS): this survey was conducted annually in May every year. Sample sizes were approximately 8,000 households per year. The whole field work organisation was very similar from year to year: five follow-ups, advance letters, about 140 free-lance interviewers, face-to-face surveys in PAPI mode. Average length of an interview was 18 minutes. The non-response rates were as follows: 8.9% in 1994, 9.0% in 1995 and 10.1% in 1996.
- Household Budget Survey 1993, 1994, 1995, 1996 (HBS): this survey was also conducted annually in December every year. Sample size in 1993 was 4,500 households, while sample sizes in 1994 - 1996 were about 1,400 households. The field work organisation was similar to that of the LFS, except for the number of interviewers. In 1993 there were 109 free-lance interviewers. In later surveys about 30 interviewers were involved. Average length of interview was about 90 minutes. The non-response rates were as follows: 19.7% in 1993, 17.8% in 1994, 18.0% in 1995 and 34.6% in 1996. The tremendous increase in the nonresponse rate in 1996 is due to the expected change of the total design of HBS in 1997 (introduction of diaries, change of questionnaire, new sample design); consequently, staff didn't take much care to properly conduct the HBS96 but they put a lot of effort in preparation for the new survey.
- Household Survey on Energy and Fuel Consumption (HSEFC): the survey was conducted for the first time in Slovenia in May 1997. The sample size was 5,000 households. Sample design for one half of the sample was stratified simple random sampling and for another half was two-stage stratified sampling. Therefore only the results for the second half were used in the analysis presented in this paper. The field work was not organised by the Statistical Office of the Republic of Slovenia as in other surveys, but the organisation of the field work was very similar. The number of interviewers was about 100.

Average length of an interview was 23 minutes. The non-response rate was 17,9%.

2.3 Variables

First of all we defined five sets of variables, concerning (1) persons (e.g., proportion of children under 15 years), (2) dwellings (e.g., proportion of privately owned dwellings), (3) households (e.g., proportion of farming households), (4) settlements (e.g., type of settlement) and (5) clusters of enumeration areas (e.g., density of population). All variables are defined in Table 1. Then we re-calculated all these variables at the level of clusters of enumeration areas. We started the estimation of the model with the variables presented in Table 1:

Table 1: Variables, used in the analysis.

		data source				
set	variable	CRP	Census	DEP	RTU	TD
1	proportion of children under 15 years	■				
1	proportion of persons over 65 years	■				
1	proportion of employed persons			■		
1	proportion of persons with higher education		■			
2	proportion of privately owned dwellings		■			
2	proportion of dwellings in apartment buildings		■			
2	proportion of weekend or summer houses		■			
3	proportion of farming households		■			
4	proportion of migration for school or work out of the settlement		■			
4	if the settlement is a centre of municipality or not				■	
4	type of settlement				■	
5	density of population				■	
5	air distance from the centre of municipality		■		■	
5	telephone coverage					■

The first idea how to use surveys which were conducted in the past was to take the response rate at the level of enumeration areas. After merging all the data we realised that there are some problems with data from the Household Budget Survey 1993. We were able to define the initial sample size and the responses for each CEA, but that was not the case for the ineligible persons. In every survey we usually experience about 5% of ineligible households because of some differences between de-iure and de-facto addresses of those persons. After some investigation of the problem we concluded that this problem is equally spread all over the country and that the results for response rate and the completion rate (i.e. the

number of responses divided by the number of initial sample) are the same. Therefore we simplified the problem and took the completion rate at the level of CEA.

3 Analysis of available data

As it was explained before, analysis of all available data was performed in a few steps. First, completion rates were computed at the level of CEA for each of the variables from independent sources separately to determine their importance and role in the future designs of surveys. Second, simple linear regression was run to determine the level of importance of each of the variables. Then, correspondence analysis was performed to show the quality of categorisation of available variables. According to the results of all these analyses, all possible types of CEA were the input data for cluster analysis where we clustered several types of CEA to get the final geo-demographic stratification.

3.1 First step - computation of completion rates

Let us first observe a few figures presenting the dependence of CEA completion rates on selected variables. We calculated general completion rates regardless of the survey.

The general finding of this step was that there existed a dependence of completion rate on most of the analysed variables. There is one very important question appearing: are the presented results survey dependent?

Therefore the completion rates for each of the surveys were computed according to all variables from administrative data sources. Below, some of the results are presented.

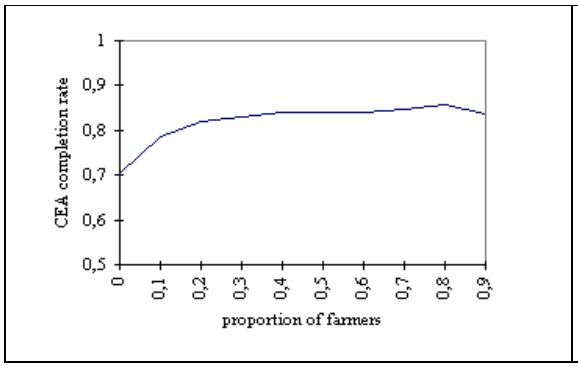

Figure 1: CEA completion rate according to the proportion of farmers.

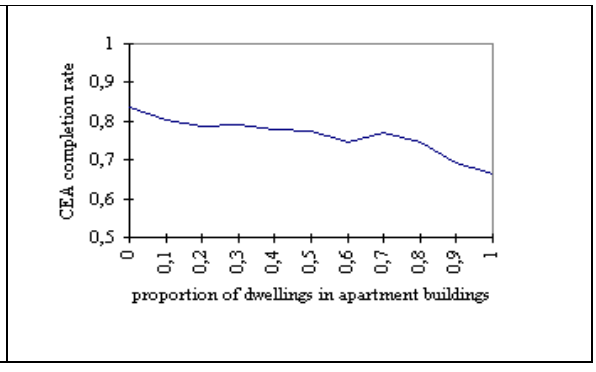

Figure 2: CEA completion rate according to the proportion of dwellings in apartment buildings.

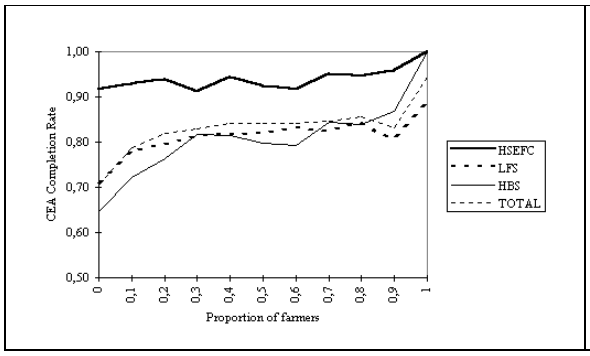

Figure 3: CEA completion rate for three different surveys according to the proportion of farmers.

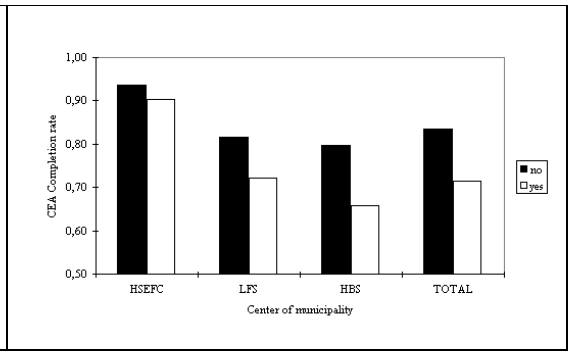

Figure 4: CEA completion rate for three different surveys according to location of CEA (centre of municipality).

We notice that the HSEFC is behaving very differently in comparison with the other two surveys which are very similar. The same picture would be given with other variables which are not shown here. Even before the estimation of the model we can expect that we have to estimate separate models for each of the surveys included. At the same time we can say that the model for the HSEFC will not explain a lot of variability in completion rates. One possible explanation is that the completion rate achieved in the HSEFC was very high. But let us have a look first at the estimation of the models.

3.2 Second step - linear regression

The second step was simple linear regression where we wanted to find predictors from available data sources for the completion rate.

The estimation of the regression model has shown what we expected and predicted according to the results of the previous section: the results cannot be generalised independent of the survey topic. Another result seen from the figures was proved: available variables do not explain the variability in completion rates for the HSEFC at all.

In Table 2 we labelled only the variables which were significant at least in one of the three regression models. The level of significance is 0.05. Other variables described before are not significant at given level.

Short interpretation of the above results would tell us that large proportion of employed persons, persons with higher education and dwellings in the apartment buildings is associated with a lower completion rate; on the other hand, a large proportion of farming households is associated with a higher completion rate. The results are showing that noncontacts are the most important component of total nonresponse in our surveys.

We can see that more or less the same variables are significant in the models for the LFS and the HBS. Only one variable is significant for the HSEFC, but even this one does not explain any variability of completion rates. Some of the "demographical" variables are much stronger in the HBS model; on the other hand, "urbanisation" variables are much stronger in the LFS model.

Table 2: Regression coefficients for significant variables.

set	variable	HBS	LFS	HSEFC
1	proportion of employed persons	-0.19		
1	proportion of persons with higher education	-0.20	-0.04	
2	proportion of dwellings in apartment buildings	-0.07	-0.11	
3	proportion of farming households		0.02	
4	proportion of migration for school or work out of the settlement	0.03		0.12
4	if the settlement is a centre of municipality or not	0.06	0.03	
4	type of settlement	-0.02	-0.02	
	Intercept	1.02	0.86	0.87

3.3 Third step - correspondence analysis

Almost all used variables were continuous. For simplification and their easier use in the future we decided to define for each of the variables some categories as described in Table 3:

Table 3: Definition of break points for selected variables.

variable	break point	
proportion of children under 15 years	20%	
proportion of persons over 65 years	20%	
proportion of employed persons	40%	
proportion of persons with higher education	40%	
proportion of privately owned dwellings	50%	
proportion of dwellings in apartment buildings	50%	
proportion of farming households	30%	
proportion of migration for school or work out of the settlement	30%	
density of population	1000 pers./km^2	
air distance from the centre of municipality	5	10

Variables not listed in the table were discrete by their definition or we left them out of the analysis.

The correspondence analysis was computed first of all for three surveys together, then for LFS and HBS together, and finally for HSEFC only. There are

two aims of these analyses: first, to evaluate the categorisation of variables described before, and second, to compare all three analyses.

Basic results which were produced with correspondence analysis and the comparison of three analyses, are the following: dimension 1 in all three results contains the "urbanisation component" of available variables, dimension 2 in all three results contains more "demographic component". In the case of completion rate in all three results, the first dimension (or the urbanisation dimension) is much stronger than the demographic dimension. Two variables concerning dwellings (proportion of privately owned dwellings and proportion of dwellings in apartment buildings) are in all results approximately the same. In further analysis, we will use only one of them. The effect of demographic variables in the case of the HSEFC is minor in comparison with the LFS and the HBS. Categorisation of variables seems to be reasonable, so further analysis will be run under this assumption.

3.4 Fourth step - cluster analysis

According to the results of the correspondence analysis, we defined all possible types of CEAs. The typology was made according to all available (now discrete) variables. In the first step, we defined them only for "used" clusters. All together, there are all together 382 different types of CEAs; 150 of them occurred only once. This kind of typology is much too detailed for future use, so we decided to analyse all different types with the hierarchical cluster analysis to merge them into reasonable clusters.

4 Final geo-demographic stratification and its advantages

The final geo-demographic stratification was obtained after cluster analysis of different types of clusters of enumeration areas. Forty final types were created on the basis of available variables and the cluster analysis. Types are described with basic properties of CEAs merged together, for example: "families with children, higher education, living in suburban areas around two largest cities".

First of all, final geo-demographic stratification was evaluated according to available data from the surveys. We are interested if completion rates differ across types of CEAs. In the figure one can observe completion rates across types.

We can conclude that this way of stratification is effective also in the prediction of completion rates. Achieved types are very homogenous in distances of CEAs from the centres of municipalities, so the stratification is also effective in the prediction of costs.

5 Conclusions

In the paper, we show the modelling of detailed geo-demographic stratification in the case of Slovenia. Data from different administrative sources were merged and then analysed according to the results from three official surveys which have been carried out during the last four years at the Statistical Office of the Republic of Slovenia. The aim of building such a detailed stratification was to be able to draw more efficient sample designs, to predict response rates better and to organise field work of future surveys much more efficiently. With correspondence and cluster analysis of all possible different types of clusters of enumeration areas we constructed forty different types of primary sampling units. In the case of completion rates we proved that this kind of stratification is effective.

The future work would be directed into the similar analysis but separately for noncontact rates and refusal rates. As we can conclude from regression coefficients (Section 3.2), noncontacts are very important component of nonresponse and can be explained very well with some of our variables.

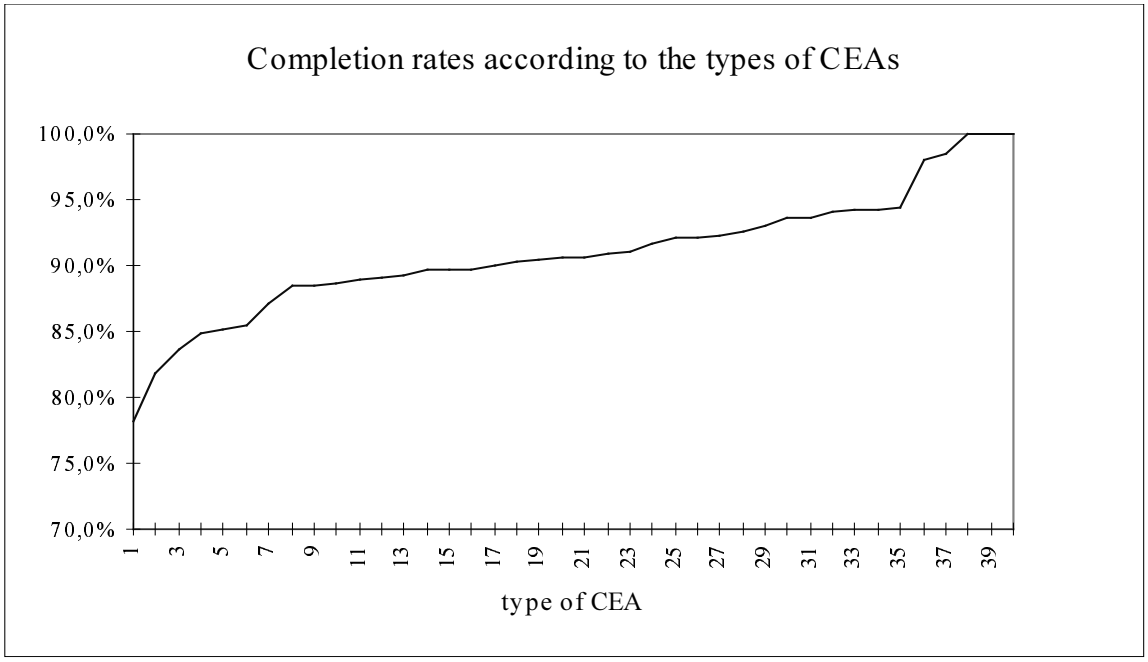

Figure 5: Completion rates according to the types of CEAs.

References

[1] Bros, L., de Leeuw, E., Hox, J., and Kurver, G. (1995): Nonrespondents in a Mail Survey: Who Are They? In S. Laaksonen (Ed.), *International Perspectives on Nonresponse*. Statistics Finland.

[2] Groves, R.M. (1989): *Survey Errors and Survey Costs*. Wiley.

[3] Groves, R.M. and Couper, M.P. (1993): *Correlates of Nonresponse in Personal Visit Survey.*

[4] King, J. (1996): *The use of geo-demographic coding schemes for understanding household non-response.* 7th International Workshop on Household Surveys Non-response, Rome, October 1996.

[5] Openshaw, S., Blake, M., and Wymer, C.: *Using Neurocomputing Methods to Classify Britain's Residential Areas.*
http://www.geog.leeds.ac.uk /staff/ m.blake /gisruk/gisruk5.html.

[6] Openshaw, S. and Blake, M.: *Selecting Variables for Small Area Classifications of 1991 UK Census Data*.
http://www.geog.leeds.ac.uk/staff/m.blake/v-sel/v-sel.html

[7] Vehovar, V. (1995): Field Substitutions in Slovene Public Opinion Survey. In A. Ferligoj and A. Kramberger (Eds.), *Contributions to Methodology and Statistics. Metodološki zvezki,* **10**. Ljubljana: FDV, 39-66.

[8] Vehovar, V. and Zaletel, M. (1995): The Matching Project in Slovenia - Who are the Nonrespondents? In S. Laaksonen (Ed.), *International Perspectives on Nonresponse*. Statistics Finland.

[9] Vehovar, V. and Zaletel, M. (1996): *Does the confidentiality concern increase the non-response rate?* 7th International Workshop on Household Surveys Non-response. Rome, October 1996.