Making sense of nonresponse with the newest Belgian census data?

A test case on the Belgian ESS6 data

Celine Wuyts, KU Leuven

Anina Vercruyssen

Geert Loosveldt

Introduction

While data linkage of survey response outcomes to individual-level auxiliary data from external sources has contributed to a better understanding of nonresponse, obtaining and linking such data is demanding, costly and often obstructed by privacy regulations. Administrative or census data aggregated to some regional level are usually more readily available and can naturally be assigned to both respondents and non-respondents. There are two ways in which such area characteristics may be related to individuals' survey response outcomes. Firstly, as a result of aggregation of individual characteristics that relate to response outcomes. For example, if there are more people with children in an area, a person living in that area has a higher likelihood of having children and may therefore be more easily contacted and more cooperative. In addition, area characteristics may give an indication of the social environment or (local) culture which may influence people who live there, even those that do not exhibit this particular individual characteristic (Groves and Couper, 1998; Johnson, 2006; 2010). For example, the presence of children in an area might indicate higher social cohesion, which may increase people's willingness to participate in a survey even if they do not have children themselves. All in all, it looks like aggregated administrative data are informative and useful to study and address survey nonresponse. In line with previous research in which area characteristics are linked to individual survey outcomes, this study investigates the potential value of the newly available and easily obtainable municipality-level data from the Belgian census 2011 for making sense of nonresponse in ESS round 6 in Belgium.

The individual, the environment and survey nonresponse

There is ample evidence that both accessibility and reluctance to cooperate are related to observable characteristics of sample units, even if those are only rough proxies of people's socio-psychological dispositions (Groves, Cialdini & Couper, 1992) and accessible at-home patterns. Such observable characteristics include age, gender, nationality, socio-economic status, family structure and housing characteristics (see e.g. Groves & Couper, 1998; Stoop, 2005; Durrant & Steele, 2009). Survey response outcomes have also been linked to observable characteristics of areas. One such classic correlate is one that is both intuitive and supported by consistent empirical evidence, namely urbanicity. People in more urbanized areas tend to be underrepresented because they are harder to reach and/or more reluctant to cooperate in surveys (Smith, 1983; Groves & Couper, 1998; Johnson et al., 2010). Related to this effect of urbanicity is the finding that people living in areas with many multi-unit buildings tend to be harder to reach, which may be explained by the presence of physical impediments such as intercoms (Groves & Couper, 1998). In addition, people in affluent areas, as measured by the proportion of people in managerial and professional occupations, appear to be both less likely to be contacted and less likely to cooperate (Johnson et al., 2006). More likely to cooperate, on the other hand, are people in residentially stable areas, as measured by fewer address changes (Johnson et al., 2006), and in more socially cohesive areas, as measured by the presence of children (Groves & Couper, 1998). In line with this research on individual and area characteristics, we examine to which extent census data aggregated to the municipality level help to explain individuals' nonresponse probabilities.

Data and methods

We use individual-level response outcome data from the sixth round of the European Social Survey (European Social Survey, 2012) in Belgium (ESS6-BE), and municipality-level data from the Belgian Census 2011 (Statistics Belgium, 2014). Final disposition codes for all sample units (3 267) were categorized into respondents (57.2%), noncontact (6.4%), refusal (23.8%), not able and other nonresponse (7.2%), and ineligible (2.6%). Based on this categorization, three binary indicators of the final response outcome were constructed: (a) whether or not an eligible sample unit is a noncontact case, (b) whether or not a contacted eligible sample unit is a refusal case, and, for overall nonresponse irrespective of the source, (c) whether or not an eligible sample unit is a completed interview case.

Due to the sampling design, sample units are clustered within 221 municipalities (Tirry & Loosveldt, 2013). The Belgian Census 2011, for the first time constructed by linking various administrative sources, was completed at the end of 2014 and resulted in a large amount of freely available data on socio-demographics, employment, education and housing of Belgian municipalities. To this list of municipality data we, added population density (log transformed) based on the population counts of Statistics Belgium on 1 January 2012.

For each of the three binary dependent variables, we estimate a logistic multilevel model with sample units nested in municipalities. A large number of potential municipality-level explanatory variables are available but high correlations exist among many of them. The choice of explanatory variables is therefore somewhat of a challenge, which may be addressed by carefully selecting the most promising municipality characteristics guided by the available literature (e.g. in Groves & Couper, 1998; Johnson et al., 2006). In this study, we selected the following municipality characteristics: population density as an indicator of urbanicity, the proportion of homemakers as an indicator of time spent at home, the proportion of multi-unit buildings as an indicator of likely presence of physical impediments, the proportion of changed address as an indicator of residential instability, the proportion of couples with young children as an indicator of social cohesion and time spent at home. The probability of noncontact is expected to be higher in municipalities that are dense, where fewer people are homemakers, where more buildings are multi-unit structures, and where there are more address changes. The probability of refusal is expected to be higher in municipalities that are dense, where there are more address changes and fewer couples with young children.

In addition to the theory-guided variable selection approach, we attempt an alternative approach, namely variable extraction. By applying an exploratory factor analysis, we try to uncover the main dimensions underlying municipality variability. The extracted factors are uncorrelated and are used as explanatory variables in further analysis.

In all our models, we control for age (included as a six-category factor) and gender, which are available from the sampling frame for all sample units.

Results

In the first step of the model building, only the control variables, age and gender of the sample unit, are included (Table 1). We find support for nonlinear associations between age and nonresponse. Elderly people are more likely to be contacted while young adults (25 to 35 years) are less likely to be contacted. Refusals, on the other hand, are less likely for the youngest age group.

Most of the selected municipality characteristics have significant effects on noncontact in the expected directions when added separately to a model which controls for sample unit age and gender (Table 2). People living in a municipality with a high population density are less likely to be contacted and so are people living in municipalities with a high number of multi-unit buildings, in municipalities with many people changing address, or with few 'couple with children' households. For the number of homemakers, our results contradict our expectations. We assumed that homemakers, who are responsible for managing the household, would spend more time at home. People living in municipalities with many homemakers would then, on average, spend more time at home and would be easier to reach. However, we find that people who live in municipalities with many homemakers are less likely to be contacted instead of more likely. Being a homemaker is possibly a flawed indicator of spending a lot of time at home at the individual level. The traditional picture of a 'homemaker' who spends most of his or her time at home with the children may not correspond to the actual reality.

However, when multiple municipality characteristics are added to a model simultaneously, all of the effects become insignificant (not tabulated). We expected that some of the municipality characteristics correlate strongly among themselves, e.g. population density and the presence of multi-unit buildings (r = 0.71). However, we also found significant correlations that we did not expect, such as the correlation between address changes and the presence of homemakers (r = 0.76). The five selected municipality characteristics turn out to be too strongly correlated (all bivariate correlations exceed 0.5) and the individual effects too hard to disentangle. As a result, AIC-based backward selection results in a noncontact model which only includes sample unit age and municipality population density. The other municipality characteristics are too strongly correlated with population density to improve upon the model fit.

The exploratory factor analysis based on 49 municipality characteristics including urbanicity, and composition in terms of age, gender, nationality, marital status, family structure, occupation, type and age of housing, yielded five municipality dimensions, namely "urbanization with low social cohesion" (population density, address changes, singles, rented houses), "presence of elderly and elderly couples" (elderly, married, retired, couples without children), "presence of elderly singles and single parents" (elderly, single parents, retired, widowed and divorced), "entrepreneurship" (self-employed, employers, higher education), and "presence of highly educated" (higher education, students). These factors are jointly added as explanatory variables (Table 3). Based on this analysis, we find that the probability of noncontact is significantly higher for people living in urban municipalities with low social cohesion. There is also some indication that reaching people is harder in municipalities with a high degree of entrepreneurship. In those municipalities, there are many self-employed and employers, who may be away from home more often. Alternatively, entrepreneurship may be related to affluence and, in turn, to the presence of particular physical impediments such as locked gates and fences.

For survey nonresponse overall, we find very similar effects of the municipality characteristics (separately) and municipality dimensions as for noncontact whereas none of the selected municipality characteristics are significantly related to the probability of refusal. Nor can the dimension reduction approach provide any more insight.

Overall survey nonresponse is more likely in municipalities where there are many multi-unit buildings, when there are many address changes, or when there are few couples with young children. In terms of municipality dimensions, we find that survey nonresponse is more likely in municipalities that are more urbanized and less socially cohesive, and where the extent of entrepreneurship is higher.

Table 1: Estimated coefficients of sample unit age and gender

	(a) noncontact			**(b) refusal**			**(c) nonresponse**		
	est.	**s.e.**		**est.**	**s.e.**		**est.**	**s.e.**	
Respondent-level									
female	-0.009	0.172		0.139	0.086		-0.095	0.074	
age									
15-25 years (ref.)									
25-35 years	0.511	0.262	°	0.526	0.166	**	0.512	0.138	***
35-45 years	-0.008	0.282		0.493	0.161	**	0.300	0.136	*
45-55 years	-0.267	0.293		0.442	0.158	**	0.196	0.134	
55-65 years	-0.504	0.324		0.390	0.166	*	0.269	0.139	°
65 years and over	-1.336	0.369	***	0.587	0.153	***	0.458	0.128	***
N	3184			2713			3184		

Table 2: Estimated coefficients of municipality characteristics (seperately, after controlling for individual age and gender)

	(a) noncontact			**(b) refusal**			**(c) nonresponse**		
	est.	**s.e.**		**est.**	**s.e.**		**est.**	**s.e.**	
Municipality-level									
population density	0.314	0.087	***	0.012	0.048		-0.072	0.040	°
% homemakers	8.043	2.846	**	1.488	1.624		-3.416	1.346	*
% multi-unit buildings	1.423	0.465	**	0.139	0.271		-0.500	0.223	*
% address changed	11.970	3.593	***	2.207	2.037		-4.590	1.683	**
% couples with young children	-4.945	2.083	*	-1.567	1.060		2.332	0.890	**
N	3184			2713			3184		

Table 3: Estimated coefficients of municipality dimensions (multivariate, after controlling for sample age and gender)

	(a) noncontact			**(b) refusal**			**(c) nonresponse**		
	est.	**s.e.**		**est.**	**s.e.**		**est.**	**s.e.**	
Municipality-level									
“urbanization and low social cohesion”	0.320	0.091	***	0.055	0.055		0.126	0.045	**
“elderly and elderly couples”	-0.235	0.112	*	-0.003	0.058		-0.026	0.049	
“elderly singles and single parents”	0.094	0.104		0.048	0.057		0.046	0.048	
“entre-preneurship”	0.194	0.099	°	0.082	0.060		0.114	0.049	*
“highly educated”	0.058	0.100		0.042	0.056		0.031	0.047	
N	3184			2713			3184		

*** p < 0.001; ** p < 0.01; * p < 0.05; ° p < 0.10

Conclusion

In this study we explored to which extent easily obtainable municipality characteristics from the Belgian Census 2011 can be used to explain survey nonresponse and its two main components, noncontact and refusal, in ESS Belgium. We found that several characteristics at the municipality level (population density, presence of multi-unit buildings, extent of address changes, presence of couples with young children) can help to explain noncontact. Thus, the new data confirm old findings.

Multivariate analyses suffer from relevant municipality characteristics being too strongly correlated. Thus, the main challenge in this type of exercise appears to be the selection of area-level variables. Some indicators may not adequately represent what we expect it to. The number of homemakers in an area appears to be an inadequate indicator of the probability a sample unit living in that area spends a lot of time at home. In addition, there may be multiple potential indicators that relate to the same thing. We decided on the proportion of couples with children as an indicator of social cohesion, but we may have used other indicators such as the proportion of single-person households or the proportion of people with a non-Belgian nationality instead. How to decide on which characteristic(s) to retain? An alternative to theory-guided variable selection is the extraction of the dimensions underlying the various municipality characteristics by factor analysis. The advantage of such a dimension reduction approach is that the resulting factors are uncorrelated but still capture a large share of the variation in the area data. While there is some subjectivity in the selection of municipality characteristics, a similar argument can be made against dimension reduction. Extracted factors usually do not correspond well with predefined constructs. Thus, the interpretation of the extracted factors is often not very straightforward. Johnson et al. (2006) is right in saying that, in the context of area characteristics, "data reduction techniques based on factor or cluster analysis present their own problems". Still, this study showed that factor analysis can be a useful approach to capture most of the municipalities' variability in just a few uncorrelated dimensions. We find five municipality dimensions, among which two are related to survey nonresponse.

In this study, both approaches suggest that population density is the main municipality-level correlate of noncontact. The attempt to identify municipality-level correlates of refusal, however, was not as fruitful. Whether on the basis of the original municipality characteristics or on the basis of the extracted municipality dimensions, the municipality data could not explain the probability of refusal. In so far the municipality level census data provide insight, they do so for noncontact cases only. This may come as no surprise, as the decision to cooperate strongly depends on individual characteristics and the interaction of the individual with the interviewer. Still, this is a rather disappointing conclusion, since noncontact remains a relatively small component of survey nonresponse in ESS Belgium (noncontact rate ESS6-BE 6.8%). In addition, noncontact is already partially addressed by requiring at least four contact attempts, with at least one in the weekend and at least one in the evening. Taking into account the results of this study, the allocation of fieldwork efforts may be targeted more efficiently, by requiring less contact attempts in rural areas and more in urban areas. However, given the small number of noncontact cases, such a strategy may not have a large impact on overall nonresponse. Even though the problem of reluctance to participate is more pressing, municipality characteristics unfortunately do not improve our understanding of it.

These results bring us to the question whether municipality characteristics such as population density or aggregated census data can be used to construct better poststratification weights. For example, respondents living in the most urban environments may be weighted more while respondents living in the most rural environments are given less weight. Assuming that the environment in which respondents live affects the attitudes and behaviors that are asked about in the survey, poststratification weights based on area characteristics may reduce nonresponse bias.

References

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(2), 361–381.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475. http://doi.org/10.1086/269338

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Johnson, T. P., Cho, Y. I., Campbell, R. T., & Holbrook, A. L. (2006). Using community-level correlates to evaluate nonresponse effects in a telephone survey. *Public Opinion Quarterly*, *70*(5), 704–719.

Johnson, T. P., Lee, G., & Cho, Y. I. (2010). Examining the association between cultural environments and survey nonresponse. *Survey Practice*, *3*(3).

Stoop, I. A. L. (2005). *The hunt for the last respondent: nonresponse in sample surveys*. The Hague: Social and Cultural Planning Office of the Netherlands.