RISQ

Representativity Indicators for Survey Quality

Partial Indicators for Representative Response

Work Package 5
Deliverable 4
Version 2[1]

Natalie Shlomo & Chris Skinner
University of Southampton, United Kingdom

Barry Schouten, Thaya Carolina, Mattijn Morren
Centraal Bureau voor de Statistiek, Netherlands

20 May 2009
Revised: 31 August, 2009

SEVENTH FRAMEWORK PROGRAMME

The RISQ Project is financed by the 7th Framework Programme (FP7) of the European Union. Cooperation Programma, Socio-economic Sciences and the Humanities, Provision for Underlying Statistics

[1] Version 2 of deliverable 4 was made in order to address comments raised by Robert Groves (Survey Research Centre, University of Michigan) in his review of the first version (May 31, 2009)

Partial Indicators for Representative Response

1. Introduction

The project RISQ (Representativity Indicators for Survey Quality), funded by the European 7th Framework Programme (FP7), is a joint effort of the NSI's of Norway, the Netherlands and Slovenia, and the Universities of Leuven and Southampton to develop quality indicators for survey response. These indicators measure the degree to which the group of respondents of a survey resembles the complete sample. When this is the case, the response is called representative. In survey practice, response rates are almost always computed. However, an indication of the contrast between respondents and the full sample is seldom given explicitly since information is needed on characteristics of households or enterprises that did not respond to the survey. Nonetheless, when information is available that is auxiliary to the survey one can indirectly measure part of the contrast. It is the objective of the RISQ project to translate auxiliary information to Representativity Indicators, to develop these quality indicators, to explore their characteristics and to show how to implement and use them in a practical data collection environment.

It is by now a well-established finding in the survey literature that survey response rates as single indicators provide insufficient information about the quality of estimates based upon respondent data. Non-response bias arises from a contrast between respondents and non-respondents on survey items. The response rate, however, sets only a bound to the maximal contrast; an increase in response rate may well go together with an increase in bias. There is a need for indicators that complement the response rate and measure the contrast between non-respondents and respondents. Since we shall consider that a key purpose of an indicator will be to support comparisons of surveys as a whole, we shall choose to define our indicators in such a way that they are not dependent upon specific survey items. Nevertheless, it is important to recognize that the definition of non-response bias is dependent upon one or more survey items. Some discussion of the relation between our indicator and non-response bias is given in Schouten, Cobben and Bethlehem (2009).

Two additional caveats are needed. Without information that is auxiliary to a survey it is not possible to make a statement about the representativity of survey response. For this reason, differences between indicator values across different surveys only have a meaning when they are based on the same set of auxiliary information. This implies that different surveys need to share some subset of auxiliary variables. The auxiliary information may either be available through direct linkage to administrative data, frame data or registers, or by means of population statistics. The other caveat concerns the sample size. Like survey statistics themselves any indicator for representativeness is a random variable with a precision depending on the sample size. Small samples do not allow for strong conclusions about the representativity of the survey response. Indicators based on auxiliary information are also subject to the effects of measurement error and coverage error in this information.

The RISQ project distinguishes R-indicators and partial R-indicators. R-indicators provide a single value between zero and one that measures closeness to representative response. Representativity is defined in terms of the response propensities of different sample units given their values on a specified set of auxiliary variables. Response is said to be representative if all the response propensities in the sample are equal (and none are equal to zero). Our definitions of R-indicators will be most effective in capturing non-response bias in a survey estimate when the auxiliary variables are, in combination, strong predictors of the survey item(s) upon which the estimate is based.

Partial R-indicators will be defined in terms of a single specified auxiliary variable and in terms of the categories of this variable when it is categorical. They will be designed to measure the impact of the specified variable on deviations from representative response. We shall also make a distinction between unconditional and conditional partial R-indicators.

The definitions we shall present of partial R-indicators will be designed to supplement R-indicators and to be used in conjunction with R-indicators.

The first RISQ paper (deliverable 2.1, Shlomo et al 2008) describes the statistical properties of two potential R-indicators: the indicator R proposed by Schouten, Cobben and Bethlehem (2009) and the variable selection measure q^2 proposed by Särndal and Lundström (2008). In that paper, we assumed a fixed set of auxiliary variables known at the sample level and compared different surveys based on datasets assembled from the participating countries in the RISQ research project. The paper covered definitions and theoretical properties of both R-indicators and a report on the empirical results of a simulation study as well as estimates from the country datasets.

The second RISQ paper (deliverable 3, Schouten et al. 2009) investigated the dependence of the R-indicators on the selected set of auxiliary variables and compared models with a fixed set of variables to models where we employ variable selection. We also examined different models for estimating the response probabilities. See Cobben and Schouten (2005, 2007) and Särndal and Lundström (2008) for more discussion on the motivation and potential uses of these indicators.

The aim of this paper is to define partial indicators and discuss their statistical properties and demonstrate their uses. Partial indicators evaluate the contribution of single auxiliary variables to a lack of representative response. The paper is self-contained. We will provide background from earlier deliverables.

Partial R-indicators may be used in different settings. We recognize the use in:

- *The comparison of different surveys*. In this setting partial R-indicators are supplementary to R-indicators. Models to describe response are simple and employ general auxiliary variables only.
- *The comparison of a survey in time*. In this setting partial R-indicators are again supplementary to R-indicators. However, models may be more complex, e.g. define multiple model equations or levels, and may employ paradata additionally to standard auxiliary variables.
- *The monitoring of data collection*. In this setting partial R-indicators assist in identifying groups that are underrepresented and may support decisions in responsive or adaptive designs or a change in future survey designs. Response models may identify different non-response types and data collection stages that produce missing data. Models may employ paradata additionally to standard auxiliary variables.

In this paper we will restrain ourselves to the first two types of use; comparing surveys and comparing a survey in time. In the RISQ project *Indicators and Data Collection Monitoring* (WP6), we will discuss the use of partial R-indicators during data collection. In order to enable a comparison we have selected household and business data sets from the five countries. Furthermore, we employ simulated data to investigate the properties of indicators. In this paper we restrict ourselves to the bias-adjusted indicator R of Shlomo et al. (2008) although the partial indicators defined in this paper can easily be extended to Särndal and Lundström's q^2 .

In Section 2 we define the partial indicators and discuss their properties. Section 3 contains a simulation study and Section 4 results of partial indicators on the country datasets. Section 5 summarizes these results and Section 6 concludes with a discussion and future work.

2. Partial indicators

Partial indicators for representative response complement representativity indicators or R-indicators (see Shlomo et al 2009, Schouten, Cobben and Bethlehem 2009). Both types of indicators are based on definitions of representative response. We, therefore, in section

2.1 start by defining what we mean by representative response. From there we move to general properties for partial indicators in section 2.2. In section 2.3 we introduce partial indicators. The proposed partial indicators are defined for categorical auxiliary characteristics of sample units. In section 2.4 we briefly discuss the extension of the indicators to continuous variables. Finally, in section 2.5 we discuss basic statistical properties like bias and precision of the partial indicators.

2.1 Definition of representative response and R-indicators

We use the notation and definition of response propensities as set out in the previous RISQ deliverables (Shlomo et al. 2008, Schouten et al. 2009). We let U denote the set of units in the population and s the set of units in the sample. We define a response indicator variable R_i which takes the value 1 if unit i in the population responds and the value 0 otherwise. The *response propensity* is defined as the conditional expectation of R_i given the vector of values x_i of the vector X of auxiliary variables:

$$\rho_X(x_i) = E(R_i = 1 \mid X = x_i) = P(R_i = 1 \mid X = x_i)$$

We assume that the values x_i are known for all sample units, i.e. for both respondents and non-respondents, and can include both specified variables and survey fieldwork conditions. Thus, X may include variables such as mode of data collection, whether there has been an advance contact, the number of callbacks, reissuance constraints etc. The response propensity is thus defined conditional on design choices which have been previously made at a particular point in time and the propensity might change over time for a given unit if new design choices are introduced. In addition to defining $\rho_X(x_i)$ as the response propensity of population unit i having value x_i on auxiliary vector X, we define $\rho_{X,Z}(x_i, z_i)$ as the response propensity of a population unit having scores x_i on X and z_i on Z.

We propose two definitions for representativeness of survey response: representative response and conditional representative response.

Definition: A response to a survey is representative with respect to X when response propensities are constant for X, i.e. $\rho_X(x_i)$ takes the same fixed value for all units i in the sample.

Definition: A response to a survey is conditional representative with respect to X given Z when conditional response propensities given Z are constant for X, i.e. $\rho_{X,Z}(x_i, z_i) = \rho_Z(z_i)$ for all units i in the sample.

Given these definitions, we want indicators for representative response to be distance measures that attain a value zero when the definition is true. Many choices of distance measures are available in the mathematical literature. The most obvious is the Euclidean distance measure. For this reason we relate deviations from (conditional) representativity on the Euclidian distance between two vectors of response propensities ρ_1 and ρ_2:

$$d(\rho_1, \rho_2) = \sqrt{\frac{1}{N} \sum_U (\rho_{1,i} - \rho_{2,i})^2} \tag{1}$$

The population variance of the response propensities ρ_X is defined as

$$S^2(\rho_X) = \frac{1}{N-1} \sum_U (\rho_X - \bar{\rho}_X)^2 ,$$

with $\bar{\rho}_X = N^{-1} \sum_U \rho_X(x_i)$

Similarly $S^2(\rho_{X,Z})$ is the population variance of the response propensities $\rho_{X,Z}$. We define R-indicators as:

$$R(\rho_X) = 1 - 2S(\rho_X) \text{ and } R(\rho_{X,Z}) = 1 - 2S(\rho_{X,Z})$$

The estimation of the propensities is typically based on a logistic regression model: $\log[\rho_X / (1 - \rho_X)] = x'\beta$ where β is a vector of unknown parameters to be estimated, and x may involve the transformation of the original auxiliary variables (e.g. by including interaction terms) for the purpose of model specification. The estimator of the response propensity is:

$$\hat{\rho}_X = \frac{\exp(x'\hat{\beta})}{\exp(x'\hat{\beta}) + 1}$$

where $\hat{\beta}$ is the estimator of β based on the model.

The estimator of the variance of the response propensities equals:

$$\hat{S}^2(\hat{\rho}_X) = \frac{1}{N-1} \sum_s d_i (\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_X)^2$$

where $d_i = \pi_i^{-1}$ is the design weight or inclusion weight and $\hat{\bar{\rho}}_X = \frac{1}{N} \sum_s d_i \hat{\rho}_X(x_i)$.

We estimate the R-indicator by:

$$\hat{R}(\hat{\rho}_X) = 1 - 2\hat{S}(\hat{\rho}_X)$$

Similarly, we define

$$\hat{\rho}_{X,Z} = \frac{\exp((x,z)^t \hat{\beta})}{\exp((x,z)^t \hat{\beta}) + 1} \text{ and } \hat{R}(\hat{\rho}_{XZ}) = 1 - 2\hat{S}(\hat{\rho}_{XZ})$$

with $\hat{S}^2(\hat{\rho}_{XZ}) = \frac{1}{N-1} \sum_s d_i (\hat{\rho}_{XZ}(x_i) - \hat{\bar{\rho}}_{XZ})^2$ and $\hat{\bar{\rho}}_{XZ} = \frac{1}{N} \sum_s d_i \hat{\rho}_{XZ}(x_i)$.

2.2 Properties of partial indicators

In section 2.1 we defined (unconditional) representative response and conditional representative response. The partial indicators that we propose measure the distance to both types of representative response for single auxiliary variables that can be linked to the survey sample and that reflect relevant characteristics of the population of interest. As such, partial indicators supplement R-indicators and can be used in conjunction with those R-indicators. Recall that R-indicators provide an overall measure of the representativeness of a survey response.

In section 2.3 we define partial indicators using standard distance measures. However, before we give definitions, we enumerate a number of additional properties that make the indicators useful for practical settings.

We would like partial indicators to have the following properties:

1) independence of the method of estimating the response propensities;
2) absence of a reference category for the auxiliary variable under investigation;
3) bounded values, i.e. they attain values in range [-1,1] or [0,1].

Additionally partial indicators should satisfy either

4) applicability to any auxiliary variable (including ones not used in modelling the propensities), where the value of the measure is not dependent on the values of other auxiliary variables;

or

5) adjustment for multivariate relations (so that the value depends on the values of other auxiliary variables).

Properties 4 and 5 cannot be requested simultaneously. A partial indicator that satisfies property 4 will not adjust for the relation between the specified auxiliary variable and other auxiliary variables. Partial indicators that satisfy property 5 will only be applicable to auxiliary variables in the model for response propensities.

We define two types of partial indicators. Unconditional partial indicators measure the contribution of single variables to a lack of representative response. Conditional partial indicators measure the contribution of single variables to a lack of representative response *given* other variables, i.e. with respect to conditional representative response. Unconditional partial indicators are designed typically for comparisons of different surveys or surveys in time. Conditional partial indicators are especially suited for data collection monitoring.

2.3 Definition of Partial Indicators

As previously mentioned, in this section we define unconditional and conditional partial indicators. First, we introduce some basic notation.

Let Z be a categorical variable with categories $k = 1,2,\ldots,K$ for which we would like to evaluate the partial indicator. Partial indicators are denoted by $P(Z,\rho_X)$ for the overall influence of variable Z and $P(Z,k,\rho_X)$ for the influence of single categories k of Z. In both cases indicators are computed given response propensities modelled by X.

The partial indicators that we propose are all based on variances of response propensities or components of these variances: the between and the within variance given a stratification defined by the auxiliary variables. Let $S_w^2(\rho_X \mid W)$ and $S_b^2(\rho_X \mid W)$ be, respectively, the within and between variance given a stratification based on a categorical variable W having categories $l = 1,2,\ldots,L$, i.e.

$$S_w^2(\rho_X \mid W) = \frac{1}{N-1}\sum_{l=1}^{L}\sum_{i\in U_l}(\rho_X(x_i) - \bar{\rho}_{X,l})^2 \tag{2}$$

$$S_b^2(\rho_X \mid W) = \frac{1}{N-1}\sum_{l=1}^{L} N_l(\bar{\rho}_{X,l} - \bar{\rho}_X)^2 \cong \sum_{l=1}^{L}\frac{N_l}{N}(\bar{\rho}_{X,l} - \bar{\rho}_X)^2 \;, \tag{3}$$

where U_l is the set of population units in stratum l, N_l is the size of stratum l, and $\bar{\rho}_{X,l}$ is the average response propensity in stratum l. Furthermore, we may denote the within and between variance attributable to a single category l of W by $S_w^2(\rho_X \mid W = l)$ and $S_B^2(\rho_X \mid W = l)$ respectively, and write

$$S_w^2(\rho_X \mid W = l) = \frac{1}{N-1}\sum_{i\in U_l}(\rho_X(x_i) - \bar{\rho}_{X,l})^2 \tag{4}$$

$$S_b^2(\rho_X \mid W = l) = \frac{N_l}{N}(\bar{\rho}_{X,l} - \bar{\rho}_X)^2 \,. \tag{5}$$

Obvious estimators for the within and between variances are weighted sample variances of estimated propensities, i.e.

$$\hat{S}_w^2(\hat{\rho}_X \mid W) = \frac{1}{N-1}\sum_{l=1}^{L}\sum_{i\in s_l} d_i(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,l})^2 \tag{6}$$

$$\hat{S}_b^2(\hat{\rho}_X \mid W) = \sum_{l=1}^{L}\frac{\hat{N}_l}{N}(\hat{\bar{\rho}}_{X,l} - \hat{\bar{\rho}}_X)^2 \tag{7}$$

$$\hat{S}_w^2(\hat{\rho}_X \mid W = l) = \frac{1}{N-1}\sum_{i\in s_l} d_i(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,l})^2 \tag{8}$$

$$\hat{S}_b^2(\hat{\rho}_X \mid W = l) = \frac{\hat{N}_l}{N}(\hat{\bar{\rho}}_{X,l} - \hat{\bar{\rho}}_X)^2 \tag{9}$$

where s_l is the set of sample units in stratum l, and $\hat{N}_l = \sum_{s_l} d_i$ is the estimated population size of that stratum.

2.3.1 Unconditional Partial Indicators

Unconditional partial indicators measure the distance to representative response for single auxiliary variables. We propose two closely related variants of unconditional indicators based on the between variance given a stratification with categories of Z. We use the between standard deviation as the partial indicator to obtain the interpretation of a Euclidian distance metric between two vectors of response propensities.

$$P_1(Z,k,\rho_X) = \sqrt{S_b^2(\rho_X \mid Z=k)} = S_b(\rho_X \mid Z=k) \quad (10)$$

$$P_2(Z,k,\rho_X) = S_b(\rho_X \mid Z=k)\frac{(\bar{\rho}_{X,k} - \bar{\rho}_X)}{|\bar{\rho}_{X,k} - \bar{\rho}_X|} = \sqrt{\frac{N_k}{N}}(\bar{\rho}_{X,k} - \bar{\rho}_X) \quad (11)$$

when Z is not used to model response propensities, and

$$P_1(Z,k,\rho_{X,Z}) = \sqrt{S_b^2(\rho_{X,Z} \mid Z=k)} = S_b(\rho_{X,Z} \mid Z=k) \quad (12)$$

$$P_2(Z,k,\rho_{X,Z}) = S_b(\rho_{X,Z} \mid Z=k)\frac{(\bar{\rho}_{X,Z,k} - \bar{\rho}_X)}{|\bar{\rho}_{X,Z,k} - \bar{\rho}_X|} = \sqrt{\frac{N_k}{N}}(\bar{\rho}_{X,Z,k} - \bar{\rho}_{X,Z}) \quad (13)$$

when Z is used to model response propensities. The two indicators are strongly related, $P_1 = | P_2 |$. It can easily be seen that $P_1 \in [0,1]$, $P_2 \in [-1,1]$. Partial indicator P_1 can also be computed on the variable level:

$$P_1(Z,\rho_X) = S_b(\rho_X \mid Z) \text{ or } P_1(Z,\rho_{X,Z}) = S_b(\rho_{X,Z} \mid Z)$$

P_1 and P_2 are in fact simple indicators and can easily be computed from stratum means and overall means of response propensities. They get meaning when they are compared to the full variance $S^2(\rho_X)$ which contains also the within variance. Estimators $\hat{P}_1$ and $\hat{P}_2$ for P_1 and P_2 are obtained by replacing the propensities with estimated propensities, replacing N_k by $\hat{N}_k = \sum_{s_k} d_i$ the estimated population size of stratum *k*, and the population variances by design-weighted sample variances.

2.3.2 Conditional Partial Indicators

Conditional partial indicators measure the distance to conditional representative response. For conditional partial indicators, Z is necessarily included in the model for response propensities. Let δ_k be the 0-1 dummy variable that is equal to 1 if $Z = k$ and 0 otherwise.

We propose two conditional partial indicators. The first indicator is based on the within standard deviation given a stratification with categories of X (assuming X is categorical)

$$P_3(Z,\rho_{X,Z}) = \sqrt{S_w^2(\rho_{X,Z} \mid X)} = S_W(\rho_{X,Z} \mid X) \tag{14}$$

$$P_3(Z,k,\rho_{X,Z}) = \sqrt{\frac{1}{N-1}\sum_{l=1}^{L}\sum_{U_l}\delta_{k,i}(\rho_{X,Z}(x_i,z_i) - \bar{\rho}_{X,Z,l})^2} \tag{15}$$

with $\bar{\rho}_{X,Z,l}$ the average of response propensities $\rho_{X,Z}$ in stratum l of X , and $P_3(Z,k,\rho_{X,Z})$ is the within standard deviation restricted to population units in stratum k . Note that this partial indicator stratifies on X while the partial indicators in (10) to (13) stratify on Z.

For the second indicator, we first define for the variable as a whole

$$P_4(Z,\rho_{X,Z}) = R(\rho_X) - R(\rho_{X,Z}) = 2(S(\rho_{X,Z}) - S(\rho_X)) \tag{16}$$

which is the difference in the R-indicator for the model including and excluding Z . Subsequently, we define

$$P_4(Z,k,\rho_{X,Z}) = P_4(\delta_k,\rho_{X,Z}) = R(\rho_X) - R(\rho_{X,\delta_k}) = 2(S(\rho_{X,\delta_k}) - S(\rho_X)) \tag{17}$$

Again it can be shown that $P_3 \in [0,1]$ and $P_4 \in [0,1]$.

An estimator $\hat{P}_3$ for P_3 is calculated by replacing propensities with estimated propensities and population variances by design-weighted sample variances. P_4 is estimated by differencing estimates of R-indicators based on estimated propensities.

The two indicators have some similarity as:

$$\begin{aligned} P_3(Z,\rho_{X,Z}) &= \sqrt{S_w^2(\rho_{X,Z} \mid X)} = \sqrt{S_w^2(\rho_{X,Z} \mid X) + S_b^2(\rho_{X,Z} \mid X) - S_b^2(\rho_{X,Z} \mid X)} \\ &= \sqrt{S^2(\rho_{X,Z}) - S_b^2(\rho_{X,Z} \mid X)} \\ &= \sqrt{S^2(\rho_{X,Z}) - S_b^2(\rho_{X,Z} \mid X) + S_b^2(\rho_X \mid X) - S_b^2(\rho_X \mid X) - S_w^2(\rho_X \mid X)} \\ &= \sqrt{S^2(\rho_{X,Z}) - S^2(\rho_X) + S_b^2(\rho_{X,Z} \mid X) - S_b^2(\rho_X \mid X)}, \end{aligned} \tag{18}$$

as by definition $S_w^2(\rho_X \mid X) = 0$. Since $S_b^2(\rho_{X,Z} \mid X) - S_b^2(\rho_X \mid X)$ may be expected to be small, $P_3(Z,\rho_{X,Z})$ is approximately equal to the difference in the variances of the response propensities for ρ_X and for $\rho_{X,Z}$. Note that this is similar to the $P_4(Z,\rho_{X,Z})$ partial indicator in (16) with the exception that $P_4(Z,\rho_{X,Z})$ is the difference in standard deviations times 2. It can be shown that

$$P_4(Z,\rho_{X,Z}) = 2\sqrt{P_3^2(Z,\rho_{X,Z}) + P_1^2(X,\rho_{X,Z})} - 2P_1(X,\rho_X)$$

Hence, the two indicators are proportional in size.

An interesting property of $P_4(Z,\rho_{X,Z})$ is that it can be expressed in terms of the maximal absolute bias. Recall from Schouten, et al. (2009) that the standardized bias with respect to auxiliary information only is $B_m(X)=\frac{1-R(\rho_X)}{2\rho}$ where ρ represents the survey response rate. B_m represents the maximal absolute bias under the scenario that non-response correlates maximally to the selected auxiliary variables. From here we obtain:

$$\Delta B_m(Z \mid X)=B_m(XZ)-B_m(X)=\frac{R(\rho_X)-R(\rho_{XZ})}{2\rho}=\frac{P_4(Z,\rho_{XZ})}{2\rho} \qquad (19)$$

Similarly, for a category k of Z according to the notation above:

$$\Delta B_m(Z,k \mid X)=B_m(XZ_k)-B_m(X)=\frac{R(\rho_X)-R(\rho_{X,\delta_k})}{2\rho}=\frac{P_4(Z,k,\rho_{XZ})}{2\rho} \qquad (20)$$

from the model used to estimate the response propensities.

2.4 Partial Indicators for Continuous Variables

So far we have assumed that auxiliary variables are categorical and as a consequence allow for a stratification of the population. In many practical survey settings, however, some of the auxiliary variables are continuous or discrete, e.g. age or income. In such cases one may categorize the variables by defining classes. It may also be desirable to measure the impact of such variables directly.

$P_4(Z,\rho_{X,Z})$ is the only partial indicator that is well-defined for continuous X and/or Z. $P_1(Z,\rho_{X,Z})$, $P_1(Z,k,\rho_{X,Z})$, $P_2(Z,k,\rho_{X,Z})$ and $P_4(Z,k,\rho_{X,Z})$ are defined for continuous X, but Z must be categorical.

Analogues of $P_3(Z,\rho_{X,Z})$ and $P_3(Z,k,\rho_{X,Z})$ where X is continuous, are

$$P_3(Z,\rho_{X,Z})=\sqrt{\frac{1}{N-1}\sum\nolimits_U(\rho_{X,Z}(x_i,z_i)-\rho_X(x_i))^2} \qquad (21)$$

$$P_3(Z,k,\rho_{X,Z})=\sqrt{\frac{1}{N-1}\sum\nolimits_U \delta_{k,i}(\rho_{X,Z}(x_i,z_i)-\rho_X(x_i))^2} \qquad (22)$$

For continuous Z, one may derive analogues of conditional indicators $P(Z,z,\rho_{X,Z})$, where z is a continuous value, by plotting or even regressing $\rho_{X,Z}(x_i,z_i)-\rho_X(x_i)$ or $(\rho_{X,Z}(x_i,z_i)-\rho_X(x_i))^2$ against z_i. For unconditional indicators it does not make much sense to extend to continuous variables Z as they may not be included in the model.

`

2.5 Statistical Properties of Partial indicators

2.5.1 Bias Adjustments

As shown in Shlomo et al. (2008), estimated R-indicators have a sample size dependent bias. When the sample size decreases, the bias is bigger. For this reason a bias adjustment was proposed for $\hat{R}(\hat{\rho}_X)$. Based on various simulations it turned out that the proposed adjustment is effective in removing the bias.

When the sampling design is a simple random sample without replacement the adjusted R-indicator has the form

$$\hat{R}_B(\hat{\rho}_X) = 1 - 2\sqrt{(1 + \frac{1}{n} - \frac{1}{N})\hat{S}^2(\hat{\rho}_X) - \frac{1}{n}\sum_{i \in s} z_i^T \left[\sum_{j \in s} z_j x_j^T\right]^{-1} z_i} \qquad (23)$$

with $z_i = \nabla h(x_i^T \hat{\beta}) x_i$ and h the link function in the model for response propensities. For linear regression h is the linear function and for a logistic regression it is the logit function.

For stratified simple random samples without replacement, the adjusted estimator is

$$\hat{R}_B(\hat{\rho}_X) = 1 - 2\sqrt{\hat{S}^2(\hat{\rho}_X) + \sum_{h=1}^{H} \frac{N_h^2}{N^2}(\frac{1}{n_h} - \frac{1}{N_h})\hat{S}_h^2(\hat{\rho}_X) - \frac{1}{N}\sum_{i \in s} \frac{N_{h(i)}}{n_{h(i)}} z_i^T \left[\sum_{j \in s} z_j x_j^T\right]^{-1} z_i} \qquad (24)$$

where $h = 1, 2, \ldots, H$ denote the strata, n_h is the (fixed) stratum sample size, N_h is the population stratum size, $h(i)$ is the stratum to which unit i belongs, and

$$\hat{S}_h^2(\hat{\rho}_X) = \frac{1}{n_h - 1}\sum_{s_h} (\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,h})^2 , \qquad \hat{\bar{\rho}}_{X,h} = \frac{1}{n_h}\sum_{s_h} \hat{\rho}_X(x_i)$$

with s_h the sampled units in stratum h.

From the observation that the R-indicator is biased, we can conclude directly that all proposed partial indicators are biased as well as they are either based on the same variance or components of that variance. Hence, a bias adjustment is needed to avoid false conclusions about the impact of single variables. We adopt a simple, pragmatic approach to adjust the bias.

The R-indicators (23) and (24) are based on a bias adjusted variance of the response propensities. The partial indicator P_1 is based on the between variance given the stratifying variable *Z*. By calculating the complementary within variance given the stratifying variable *Z*, we implement a heuristic of pro-rating the bias correction of the R-indicator between the decomposed variance components. Similarly, the partial indicator P_3 is based on the within variance given the stratifying variable *X*. We calculate the complementary between variance given the stratifying variable X and pro-rate the bias correction between the decomposed variance components. Thus we obtain bias corrections for both partial indicators P_1 and P_3.

The partial indicator P_4 is adjusted by differencing bias adjusted R-indicators.

2.5.2 Confidence Intervals of Partial indicators

The other important property is the standard error. Since partial indicators are random variables, they will have a certain precision that depends on the size of the sample. Hence, we need to evaluate their values in terms of confidence intervals.

In this paper we, resort to resampling methods for the estimation of confidence intervals (Efron and Tibshirani, 1993). We recompute partial indicators P_1, P_2, P_3 and P_4 for M bootstrap samples $m = 1,2,\ldots,M$ and form a $100(1-\alpha)\%$ confidence interval estimate by ordering the estimates for the different bootstrap replicates and define the confidence interval in terms of the $\alpha/2$ and $1-\alpha/2$ quantiles. In this paper we use $M = 1000$ and $\alpha = 0.05$, i.e. we omit the smallest 25 estimates and the 25 largest estimates.

In the appendix, we provide analytical expressions of approximations to the variance of $P_1(Z,k,\rho_X)$ in (10) and $P_3(Z,k,\rho_{X,Z})$ in (15). The approximations will be evaluated and implemented in future research through Work Packages 6 and 7.

3. Application to Simulated Datasets

In this section we investigate the properties of the partial indicators proposed in Section 2.3 using simulated survey data. The goal of the simulation is to analyze the effectiveness of the proposed bias adjustments and the dependence of confidence intervals on the sample size. In other words, can we remove the bias and to what extent does sample size influence the conclusions that can be drawn from the indicators?

For the simulation study, we use a dataset from the 1995 Israel Census Sample of Individuals aged 15 and over (N=753,711). Population response propensities were calculated using a 2-step process:

1. Probabilities of response were defined according to variables: child indicator, income from earnings groups, age group, sex, number of persons and locality type.

2. Using the response indicator as the dependent variable, we fit a logistic regression model on the population using the above explanatory variables. The predictions from this model serve as the 'true' response propensities for our simulation.

The overall non-response rate generated in the simulated dataset was 22%. Table 3.1 presents the non-response rates for the different variables used in the logistic model for defining the population response propensities. High non-response rates in categories are likely to cause the sub-group in the population to be under-represented in the partial indicators.

Table 3.1: Percentage of non-response generated in simulated dataset according to auxiliary variables

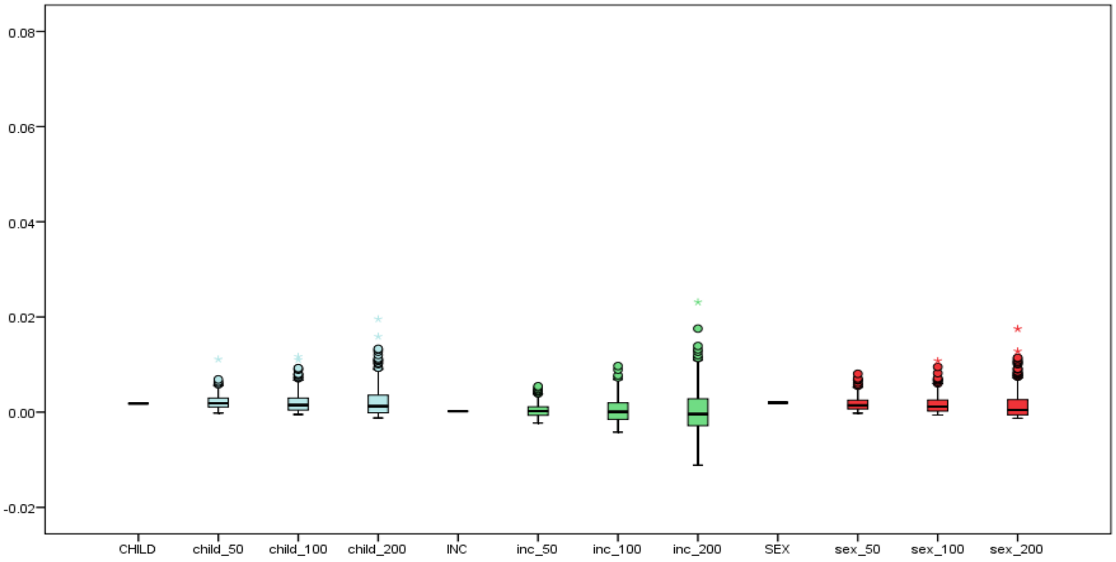
Variable	Category	Percentage non-response
Sex	Male	11
	Female	11
Children	None	18
	1+	4
Type of Locality	3 largest cities	7
	Jewish	12
	Non-Jewish	3
Age group	Young	5
	Middle	15
	Elderly	2
Persons in Household	1-2	12
	3-4	7
	Over 5	3
Income Groups	none	9
	low	7
	high	6

From this population, we drew 400 samples under three sample fractions: 1:50 (sample size is 15,074), 1:100 (sample size is 7,537) and 1:200 (sample size is 3,768) using simple random sampling. We present the results through a series of box plots in Figures 3.1 to 3.10. Box plots show the mean, median and the spread of the distribution for each of the R-indicators across the 400 simulations. In each Figure, the variables are labelled according to the name of the variable (or category). Each variable has 4 box plots associated with it. The first is the 'true' value in the population (denoted by a straight line) followed by box plots based on the repeated samples according to the sampling fraction: '50' for the 1:50 sample; '100' for the 1:100 sample; '200' for the 1:200 sample.

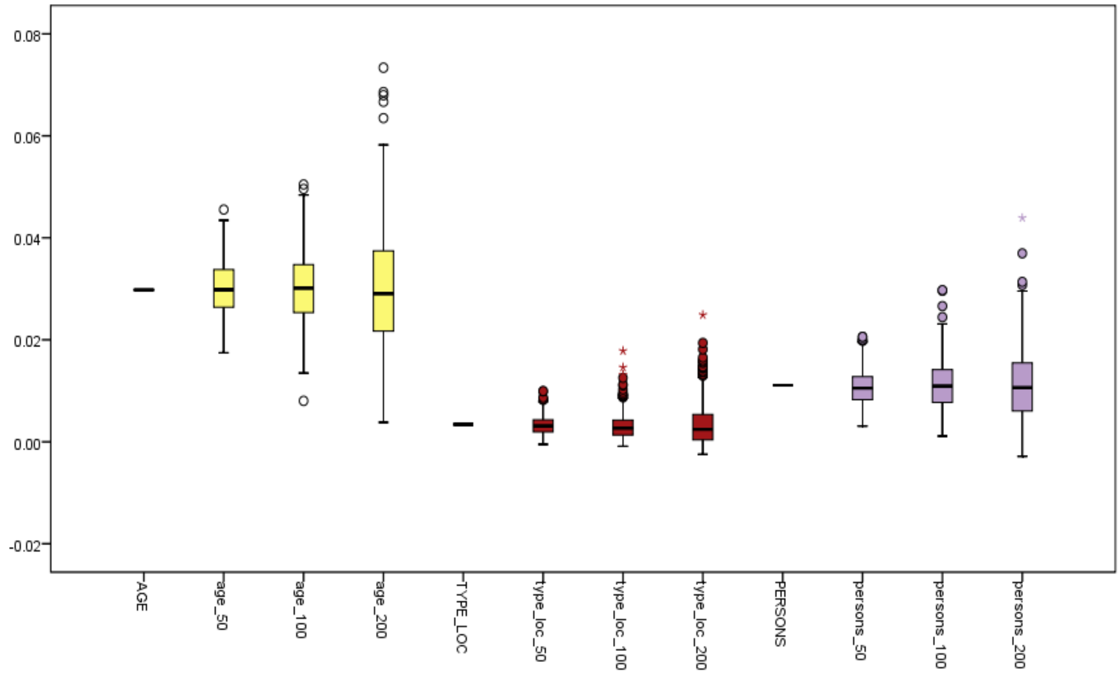
3.1 The conditional partial indicators

The partial indicator P_4 is the difference between the R-indicator based on estimated propensities from a smaller model *X* and the R-indicator based on estimated propensities from a larger model *X,Z*. A high partial indicator P_4 means that more bias is explained by the particular variable *Z*. Figure 3.1 show similar results and good estimation of the partial indicator P_4 compared to the 'true' population partial indicator. The larger sample size in Figure 3.1 results in smaller inter-quartile ranges across all variables. Although P_4 is small in this simulation (between 0 and 0.05), it is clear that Age Group contributes the most to explaining the bias or lack of representativity of the sample. Note that we would expect that the less variables are present in the model, the less is the variance of the response propensities and hence a higher R-indicator. Therefore, we expect positive partial indicators P_4. Some of the samples drawn in the simulation, however, resulted in negative values for P_4 as can be seen, for example, in the Income Group variable.

Figures 3.1a: Partial Indicator $P_4(Z,\rho_{X,Z})$ (Difference in R-indicator after excluding designated variable from auxiliary variable set)



Figures 3.1b: Partial Indicator $P_4(Z, \rho_{X,Z})$ (Difference in R-indicator after excluding designated variable from auxiliary variable set)



**Population P_4: Child 0.0018, Income Group 0.0002, Sex 0.0020, Age Group 0.0298, Type of Locality 0.0034, Persons 0.0111*

The interpretation of P_3 (similarly to P_4) in Figure 3.2 is how much of the variation is left in the cells defined by variables X after removing variable Z. A high P_3 means that there is more variation within the cell after removing Z and hence less representativity. The range of the partial indicator P_3 compared to P_4 is also small (between 0 and 0.05) and we see higher values for Age Group, Type of Locality and Number of Persons. It is interesting to note that although we expect P_3 to behave similarly to P_4 from (18), the variables Type of Locality and Number of Persons have higher partial indicators P_3 compared to P_4. This is likely due to the fact that P_3 is based on a logistic model for estimating response propensities which included an interaction term Type of Locality*Number of Persons, whereas P_4 is based on a logistic model with main effects only, each time dropping one variable from the model. P_3 is on average estimated accurately compared to the 'true' population values except for the Income Group variable. It seems that the estimate for P_3 is overestimating the contribution to the lack of representativity for Income from Earnings Group. This may be because of the highly skewed distribution of income from earnings with over half of the persons in the dataset having no earnings from work. The smaller sample size (1:200 sampling fraction) results in under-estimation of the contribution to the lack of representativity compared to the other sample sizes.

Figure 3.2a: Partial Indicator $P_3(Z,\rho_{X,Z})$ (Within variance of cross- classified variables after removing the designated variable)

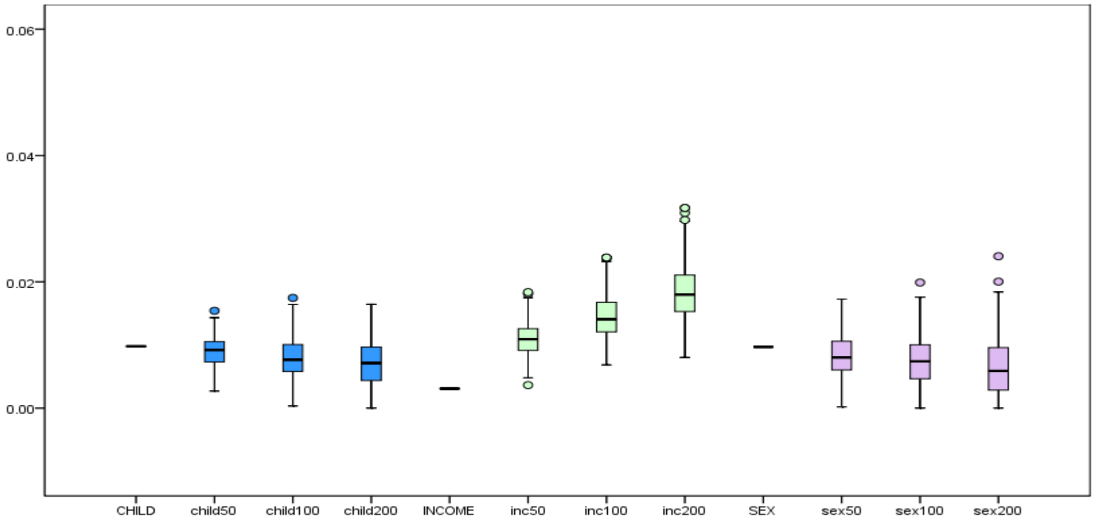
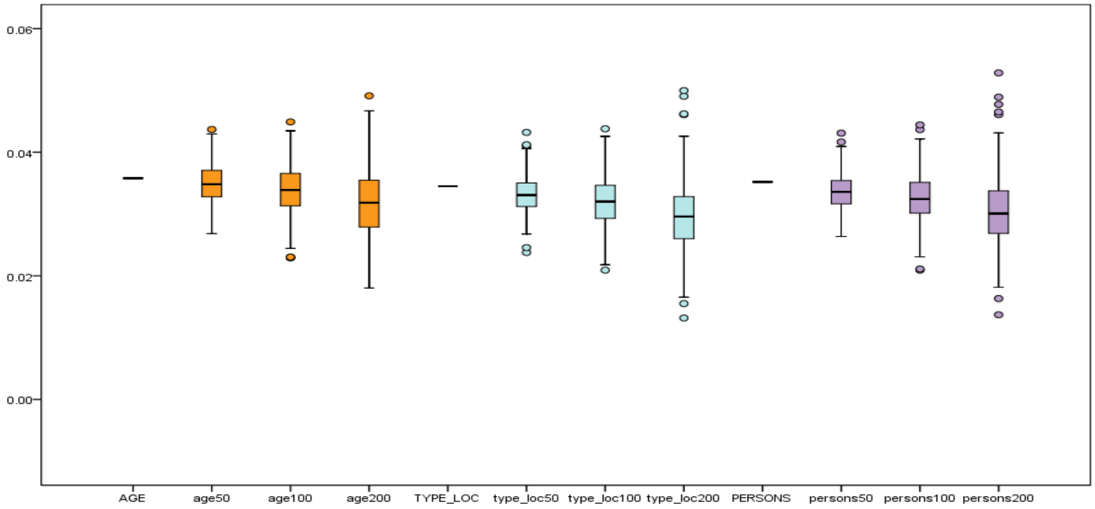


Figure 3.2b: Partial Indicator $P_3(Z,\rho_{X,Z})$ (Within variance of cross- classified variables after removing the designated variable)



* *Population* P_3**:** *Child Indicator 0.0098, Income Group 0.0031, Sex 0.0097, Age Group 0.0358, Type of Locality 0.0345, Persons 0.0352*

3.2 The unconditional partial indicators

P_1 is not conditional on other variables. It is based on the between variance across categories of a single variable *Z*, i.e. the larger the between variance the larger the differences in representativity across categories. The range of P_1 is similar to the previous indicators P_3 and P_4. In Figure 3.3, we see evidence in P_1 of overestimating the representativity of the Income Group variable as was seen for P_3. The variables: Education, Region, Ethnicity and Marital Status were not in the original logistic model used to estimate the response propensities. Nevertheless, we are able to estimate the representativity of these variables. Similar to the conditional partial indicators, Age Group is explaining the most bias or lack of representativity.

In Figures 3.4 to Figures 3.10, the partial indicator $P_2(Z,k,\rho_X)$ in (11) and (13) is depicted for each separate variable. Recall that P_2 is bounded by [-1,1] and hence negative values of P_2 indicate underrepresentation while positive values indicate overrepresentation.

Figure 3.3a: Partial Indicator $P_1(Z,\rho_X)$ (Between variance across categories of designated variable)

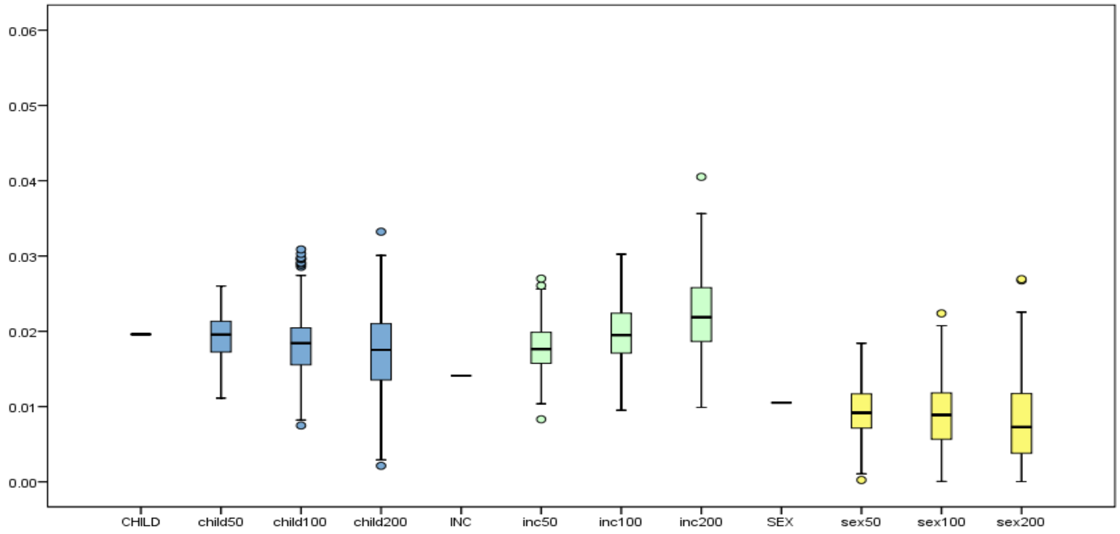


Figure 3.3b: Partial Indicator $P_1(Z,\rho_X)$ (Between variance across categories of designated variable)

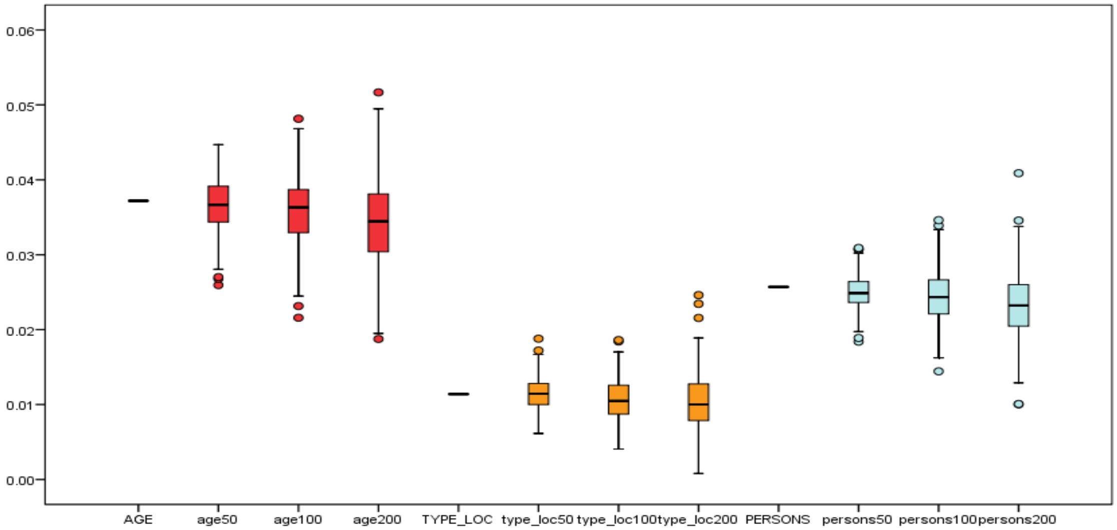
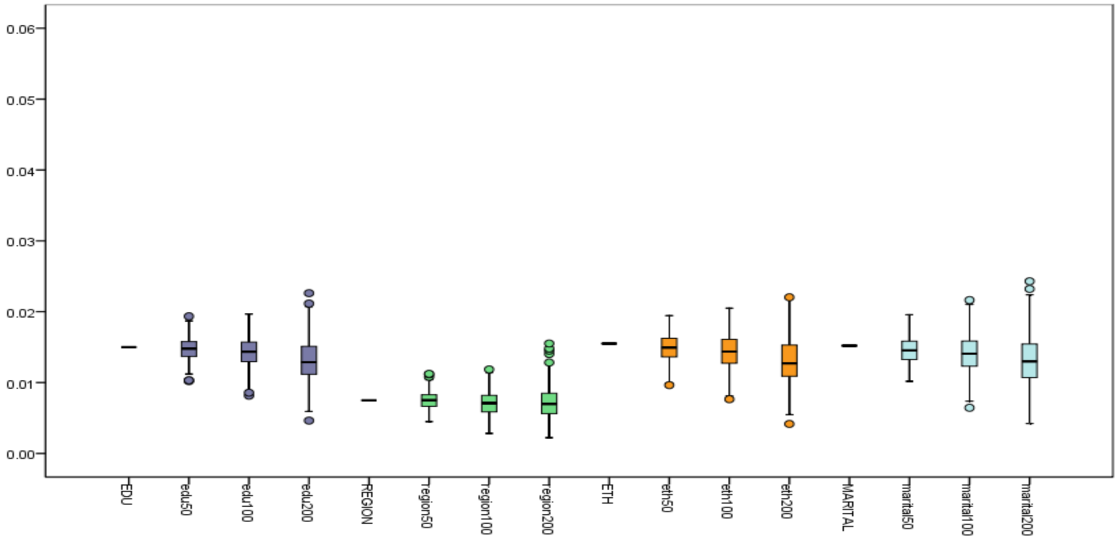
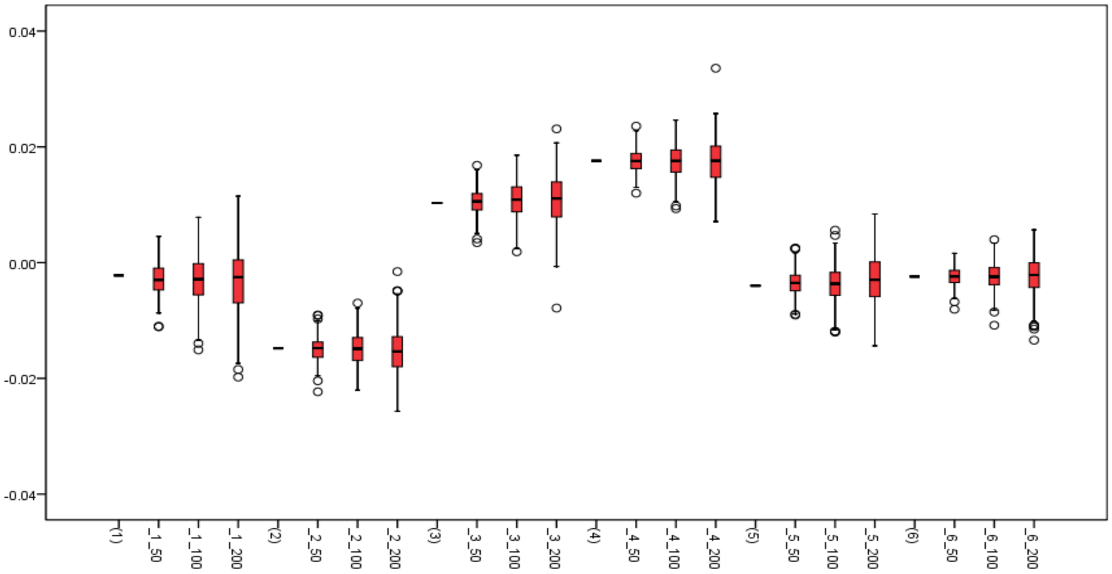


Figure 3.3c: Partial Indicator $P_1(Z,\rho_X)$ (Between variance across categories of designated variable)



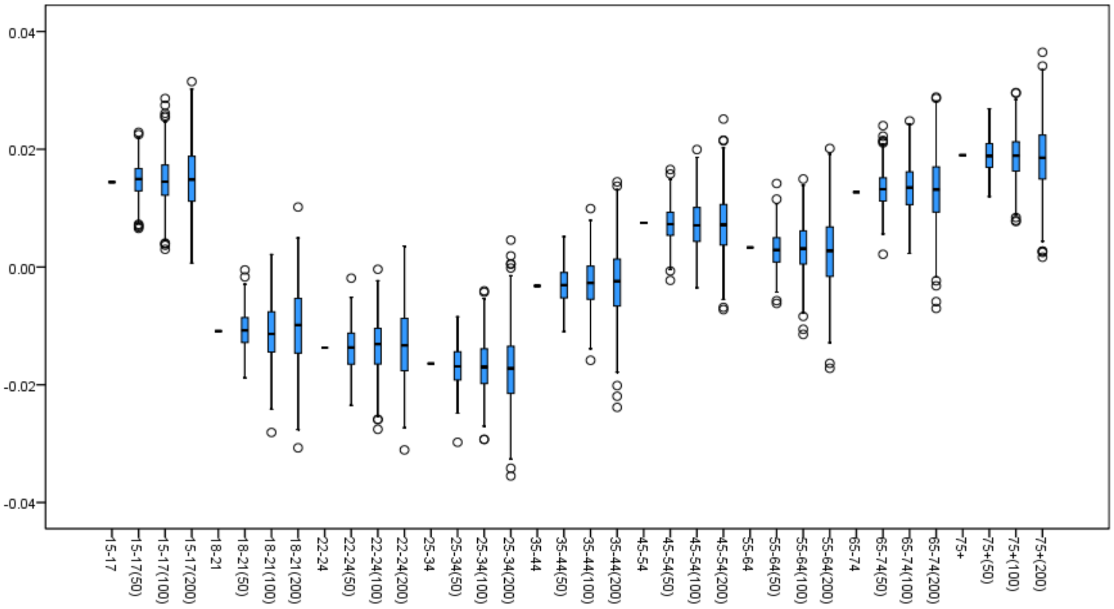
* *Population* P_1 **:** *Child Indicator 0.0196, Income Group 0.014, Sex 0.0105, Age Group 0.0372, Type of Locality 0.0114, Persons 0.0257, Education 0.0150, Region 0.0075, Ethnicity 0.0155, Marital Status 0.0152*

Figure 3.4: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Number of Persons



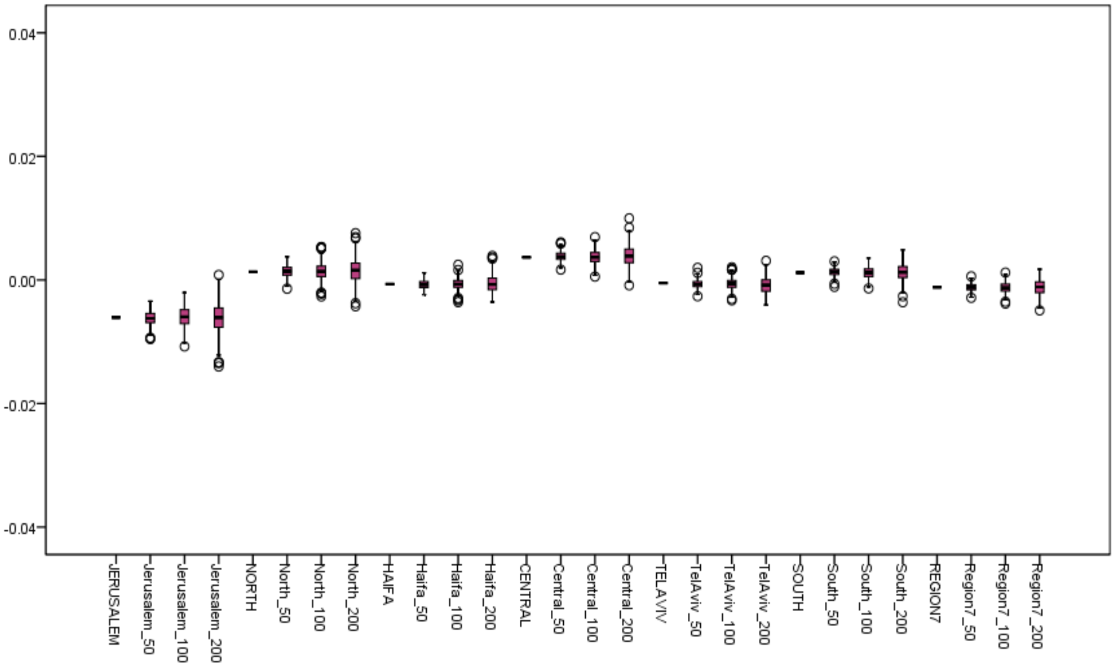
* *Population P_2: 1 Person -0.0022, 2 Persons -0.0148, 3 Persons 0.0103, 4 Persons 0.0176, 5 Persons -0.0040, 6+ Persons -0.0024*

Figure 3.5: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Age Group



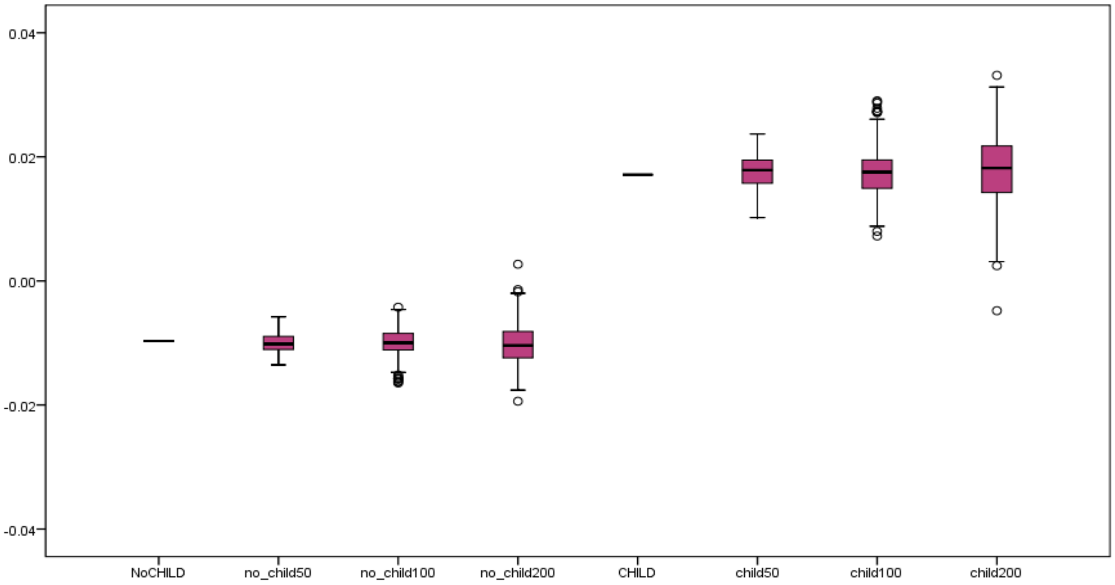
* *Population P_2: 15-17 0.0144, 18-21 -0.0109, 22-24 -0.0137, 25-34 -0.0164, 35-44 -0.0032, 45-54 0.0075, 55-64 0.0033, 65-74 0. 0127, 75+ 0.0190*

Figure 3.6: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Region



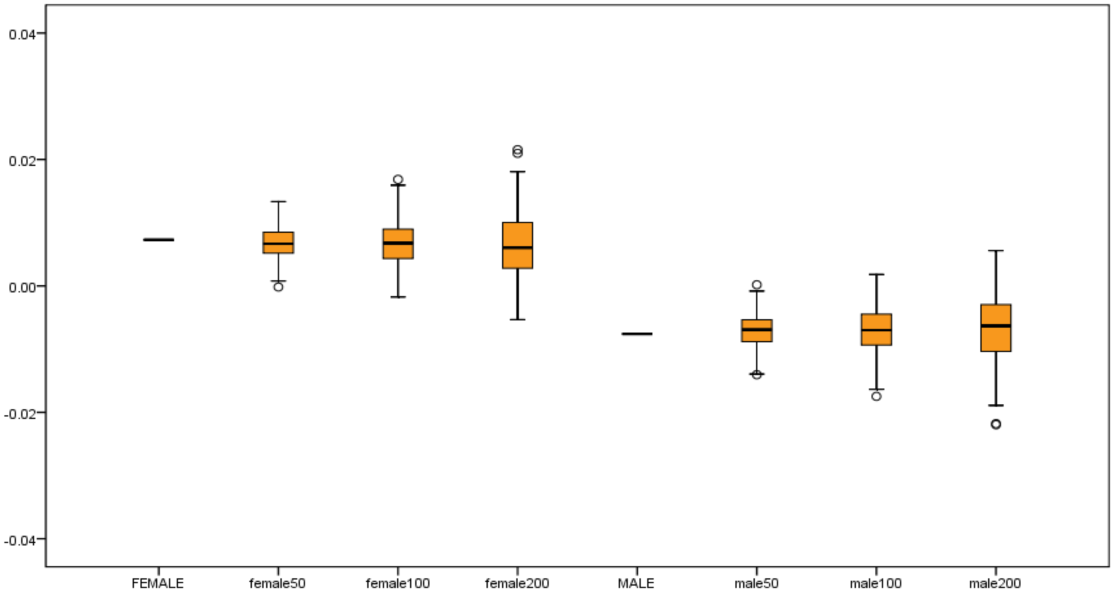
** Population P_2: Jerusalem -0.00607, North 0.0013, Haifa -0.00066, Central 0.0036, TelAviv -0.00049, South 0.00117, Region7 -0.00120*

Figure 3.7: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Child Indicator



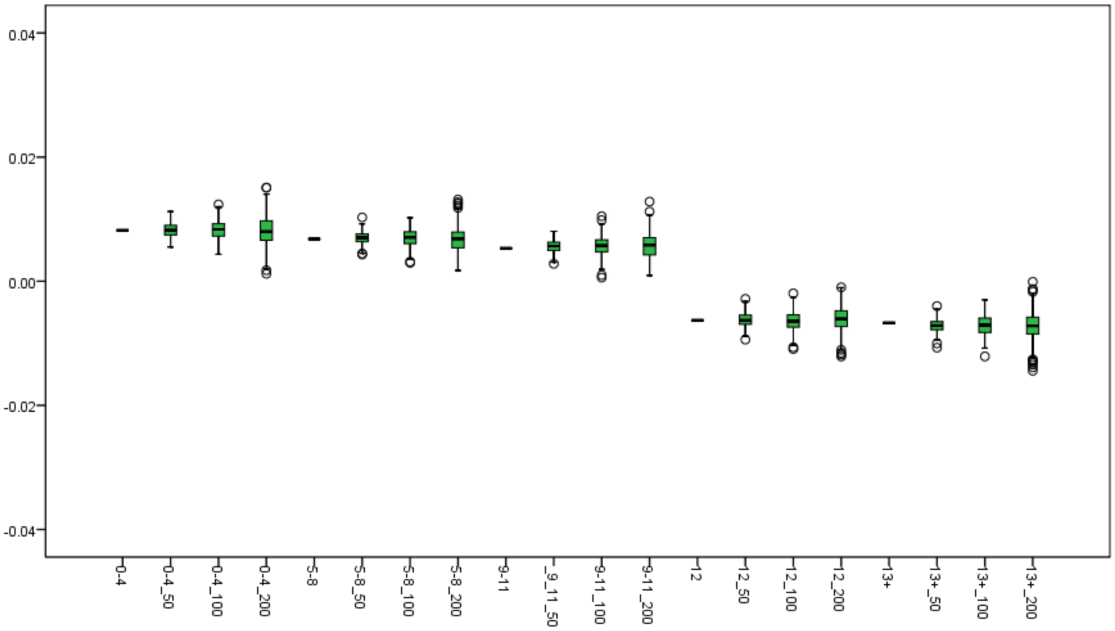
** Population P_2: No Child: -0.0097 Child: 0.0171*

Figure 3.8: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Gender



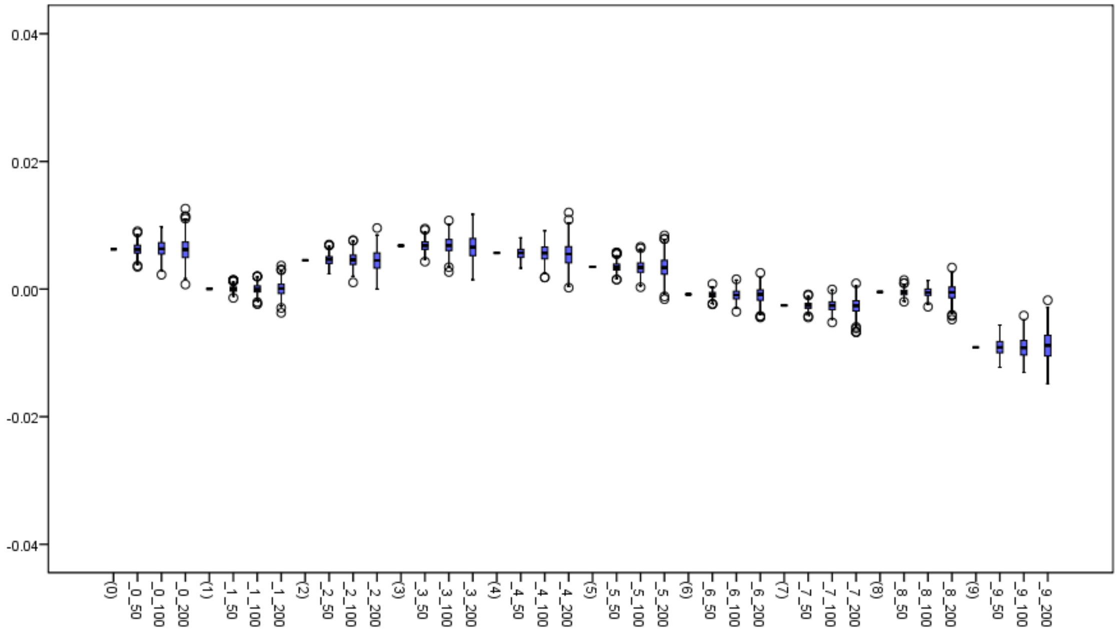
** Population P_2: Males: -0.0076, Females: 0.0073*

Figure 3.9: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Years of Study



** Population P_2: 0-4 0.0082, 5-8 0.0068, 9-11 0.0053, 12 -0.0063, 13+ -0.0067*

Figure 3.10: Partial Indicator $P_2(Z,k,\rho_X)$ for Categories of Ethnicity



* *Population* P_2**:** *0 0.00623, 1 0.00002, 2 0.00450, 3 0.00679, 4 0.00567, 5 0.00347, 6 -0.00084, 7 -0.00255, 8 -0.000455, 9 -0.00913*

Figures 3.4 to 3.10 show the use of the partial indicator $P_2(Z,k,\rho_X)$ to identify categories of variables that are underrepresented (below zero) and overrepresented (above zero). Used in conjunction with the R-indicator, these partial indicators assist in the individual analysis of representativity and can be especially useful for field work monitoring and for localizing sub-groups for targeted data collection. Examples of underrepresented groups in this simulation are: household sizes of 1 or 2, no children, males, ages 18-34, over 12 years of education, Jerusalem and ethnic group 9 (native born).

The figures also show that the bias adjustment on the partial indicators is effective at eliminating the bias due to sample size. The average values of the partial indicators across the repeated samples are approximately the same for the different sample sizes. In addition, confidence intervals are narrower for the larger sample size with less outliers. The sample sizes assessed in the simulation present consistent conclusions with respect to the representativity of the variables and their categories.

4. Application to Real Survey Data

In this section, we apply the unconditional and conditional partial indicators to country data sets participating in the RISQ project: Belgium, Norway, Netherlands, Slovenia and UK. The data sets are documented and described in RISQ (2008). Data set documentations are available at www.R-indicator.eu .

The following is a brief description of each of the datasets:

Household data:
Dutch Consumer Sentiments survey 2005 (CSS-CBS)
The Consumer Sentiments Survey is a continuous survey of households with questions about general economic development, and the financial situation of the household. The survey is meant to provide insight into short term economic development, and early indicators of differences in consumer trends. The number of cases in the file is 17,908. The response rate was 66.9%.

The Dutch Health Survey 2005 (HS-CBS)
The Dutch Health Survey is a continuous survey of individuals with questions about health, life style and use of medical care. It consists of three questionnaires; a CAPI base module, a CAPI topical module about health and a supplementary paper questionnaire. The number of cases in the file is 15,411. The response rate was 67.3%.

UK 2001 Labour Force Survey (LFS-UK)
A part of the UK 2001 Census Link Study, we evaluate the Labour Force Survey from May-June 2001, including all households that had a successful link with the Census data. The number of households in the dataset is 7,830 and the response rate about 80%.

Norwegian European Social Survey 2006 (ESS-NO)
ESS is a biennial multi-country survey of individuals covering over 30 nations. It is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. The data set only contains the survey data of Norway. The number of cases in the file is 2,673. The response rate was 65.5%.

Norwegian Survey of Level of Living 2004 (LLS-NO)
The survey of living conditions has two main purposes. One is to throw light on the main aspects of the living conditions in general and for various groups of people. Another purpose is to monitor development in living conditions, both level and distribution. Over a three-year period the cross-sectional survey of living conditions will cover all main

areas of the living conditions. The survey topics change during a three-year cycle. Housing conditions, participation in organisations, leisure activities, offences and fear of crime were topics in 2004. It is a survey of individuals. The number of cases in the file is 4,837. The response rate was 69.1%.

Belgium European Social Survey 2006 (ESS-BE)
As described for the Norwegian dataset, the ESS is an EU harmonized social survey. The data set contains the survey data of Belgium. The number of cases in the file is 2,927. The response rate was 61.4%.

Slovenian Labour Force Survey 2007 (LFS-SLO)
The Slovenian Labour Force Survey is an EU harmonized rotating panel survey conducted continuously through the year. The data contains employment related characteristics and demographic characteristics of all individuals 15 years or older living in selected households. The number of households varies between 7,010 and 7,160 households which is around 16,900 responding individuals. The response rate is around 80%.

Business data:
Slovenian Survey on usage of information-communication technologies (ICT) in enterprises 2007 (ICT-SLO)
The Slovenian survey is an EU harmonized annual survey on the usage of ICT and provides information on whether the enterprises use computers, the internet, electronic commerce and other ICTs. The number of cases in the file is 1,998. The response rate is 87.6%.

Dutch Short Term Statistic on Industry 2007 (STS-IND-CBS)
The Dutch Short Term Statistics on Industry is a monthly survey for Eurostat. It measures turnover for businesses in The Netherlands. The number of cases in the file is 64,413. The response rate was 92.5%

Dutch Short Term Statistic on Retail 2007 (STS-RET-CBS)
The Dutch Short Term Statistics on Retail is a monthly survey for Eurostat. It measures turnover for businesses in The Netherlands. The number of cases in the file is 93,799. The response rate was 92.3%.

In most cases, we calculate partial indicators for two response models: a logistic model based on a small auxiliary variable set and a logistic model based on an extended auxiliary variable set. The small auxiliary variable set consists of variables that are shared by all countries and, hence R-indicators can be compared across countries. The extended auxiliary variable sets include country specific variables. With the evaluation we have two goals in mind:

1. compare partial indicators across countries, and
2. investigate the impact of the model (small versus extended).

In the following the results from the collection of country datasets are presented.

4.1 Household Data

A. CSS-CBS dataset: Sample size = 17,908 overall response rate: 66.9% **Small auxiliary variable set**: AgeGroup*MaritalStatus (14), Gender (3), Urbanicity (5) R-indicator: 0.833 (CI: 0.818-0.848)
Extended auxiliary variable set: AgeGroup*MaritalStatus (14 categories), Gender (3), Urbanicity (5), HouseValue (9), Ethnicity (5), Type of Household (7), Job (2)
R-indicator: 0.821 (CI: 0.807-0.834)

Table 4.1: CSS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Categories of Gender and Urbanicity

Category	Small Variable Set			Extended Variable Set		
	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
			Gender			
Males	-0.0310	-0.0334	-0.0243	-0.0304	-0.0372	-0.0237
Females	-0.0373	-0.0395	-0.0306	-0.0358	-0.0422	-0.0298
Mixed	0.0339	0.0324	0.0381	0.0327	0.0289	0.0367
			Urbanicity			
Very Strong	-0.0162	-0.0184	-0.0098	-0.0162	-0.0226	-0.0100
Strong	0.0028	0.0007	0.0088	0.0031	-0.0029	0.0092
Moderate	0.0058	0.0035	0.0123	0.0060	-0.0004	0.0120
Little	0.0057	-0.0009	0.0121	0.0055	-0.0005	0.0118
Not	0.0033	-0.0031	0.0095	0.0030	-0.0036	0.0092

Table 4.1 shows the results for the partial indicator $P_2(Z,k,\rho_{X,Z})$ for the categories of the variables Gender and Urbanicity based on data from the Dutch CSS. The same is illustrated in Figure 4.1. The total sample size equals 17,908 whereas the overall response rate fort this particular survey was equal to 66.9%. $P_2(Z,k,\rho_{X,Z})$ is calculated for a small auxiliary variables set, containing the variables AgeGroup*MaritalStatus (14 categories), Gender (3), Urbanicity (5) and an extended containing in addition the variables HouseValue (9 categories), Ethnicity (5), Type of Household (7), Job (2). The R-indicator equals 0.833 for the small set and 0.821 for the extended variable set. Recall

that negative values of P_2 indicate categories that are underrepresented while positive values of P_2 indicate overrepresentation.

Figure 4.1a: CSS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Gender, for small (sm) and extended (ex) auxiliary sets

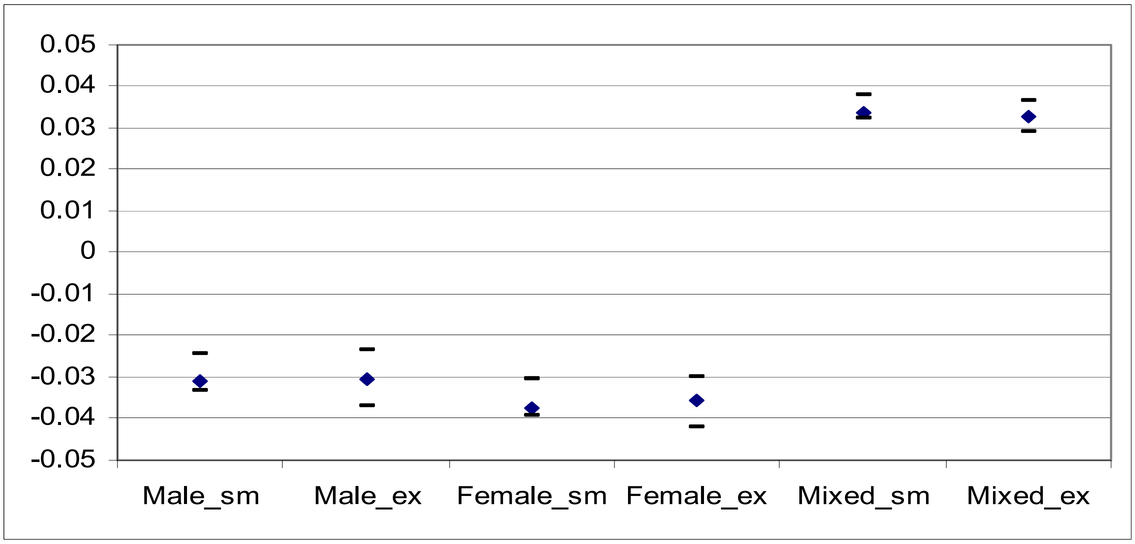


Figure 4.1b: CSS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Urbanicity, for small (sm) and extended (ex) auxiliary sets

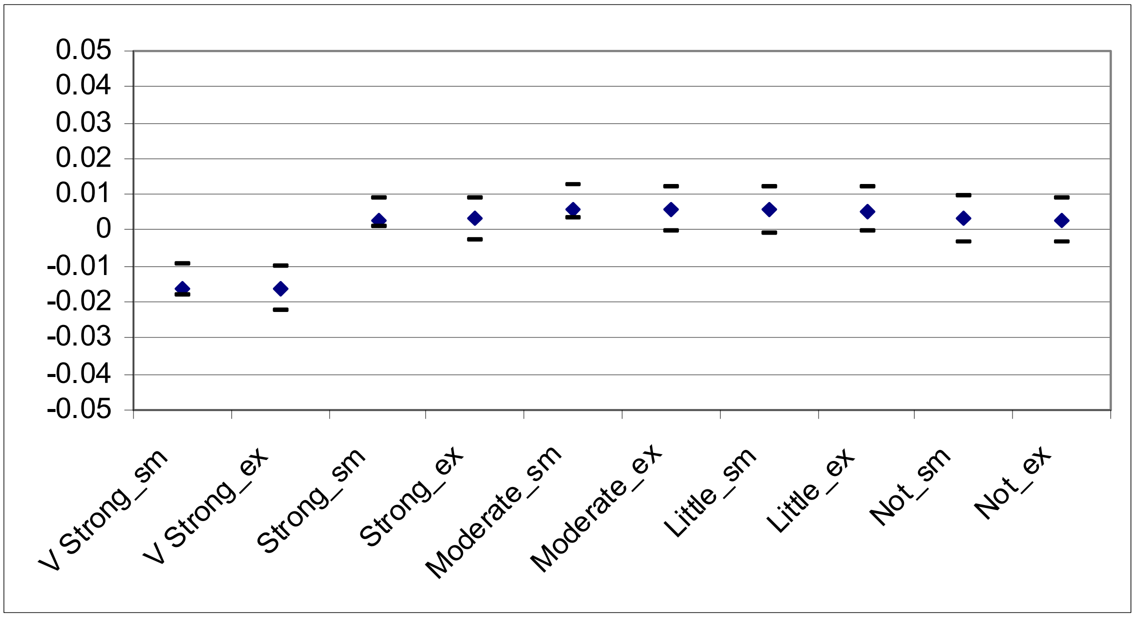
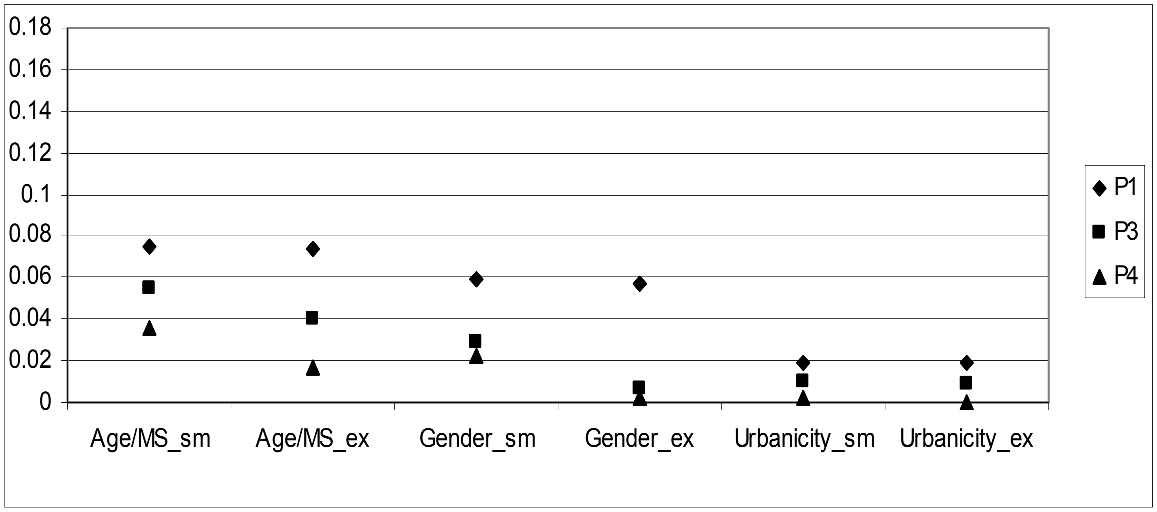


Table 4.2: CSS-CBS: partial indicators $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by variables

Partial Indicator	Small Variable Set			Extended Variable Set		
	Age Group, Marital Status	Gender	Urbanicity	Age Group, Marital Status	Gender	Urbanicity
$P_1(Z,\rho_{X,Z})$	0.0754 (0.0729-0.0830)	0.0592 (0.0566-0.0664)	0.0186 (0.0117-0.0257)	0.0733 (0.0662-0.0807)	0.0572 (0.0505-0.0640)	0.0186 (0.0119-0.0255)
$P_3(Z,\rho_{X,Z})$	0.0545 (0.0521-0.0616)	0.0293 (0.0270-0.0366)	0.0104 (0.0040-0.0169)	0.0408 (0.0349-0.0472)	0.0070 (0.0025-0.0117)	0.0088 (0.0030-0.0148)
$P_4(Z,\rho_{X,Z})$	0.0361 (0.0200-0.0345)	0.0227 (0.0148-0.0307)	0.0019 (-0.0007-0.0046)	0.0169 (0.0109-0.0229)	0.0020 (-0.0001-0.0041)	0.0004 (-0.0011-0.0019)

Figure 4.2: CSS-CBS: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ for small (sm) and extended (ex) auxiliary sets for Age/MaritalStatus, Gender and Urbanicity



Similar analysis is shown in Table 4.2 and Figure 4.2 for the unconditional partial indicator $P_1(Z,\rho_{X,Z})$ and the conditional partial indicators $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$. High values for P_3 and P_4 signify a higher contribution of the specific category to the lack of representativity.

B. HS-CBS dataset: Sample size=15,411, overall response rate: 67.3%
Small auxiliary variable set: AgeGroup*MaritalStatus (15), Gender (2), Urbanicity (5)
R-indicator: 0.832 (0.819-0.847)
Extended auxiliary variable set: AgeGroup*MaritalStatus (15), Gender (2), Urbanicity (5), HouseValue (10), Ethnicity (6), TypeofHousehold (8), Job (2)
R-indicator: 0.808 (0.794-0.823)

Table 4.3: HS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Categories of Gender and Urbanicity

Category	Small Variable Set			Extended Variable Set		
	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Gender						
Males	-0.0025	-0.0079	0.0025	-0.0026	-0.0078	0.0026
Females	0.0025	-0.0024	0.0079	0.0026	-0.0025	0.0077
Urbanicity						
Very Strong	-0.0333	-0.0401	-0.0265	-0.0332	-0.0403	-0.0266
Strong	-0.0144	-0.0209	-0.0079	-0.0144	-0.0211	-0.0073
Moderate	0.0100	0.0035	0.0164	0.0096	0.0029	0.0157
Little	0.0191	0.0126	0.0254	0.0195	0.0130	0.0261
Not	0.0206	0.0143	0.0271	0.0205	0.0141	0.0268

Figure 4.3a: HS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Gender for small (sm) and extended (ex) variable sets

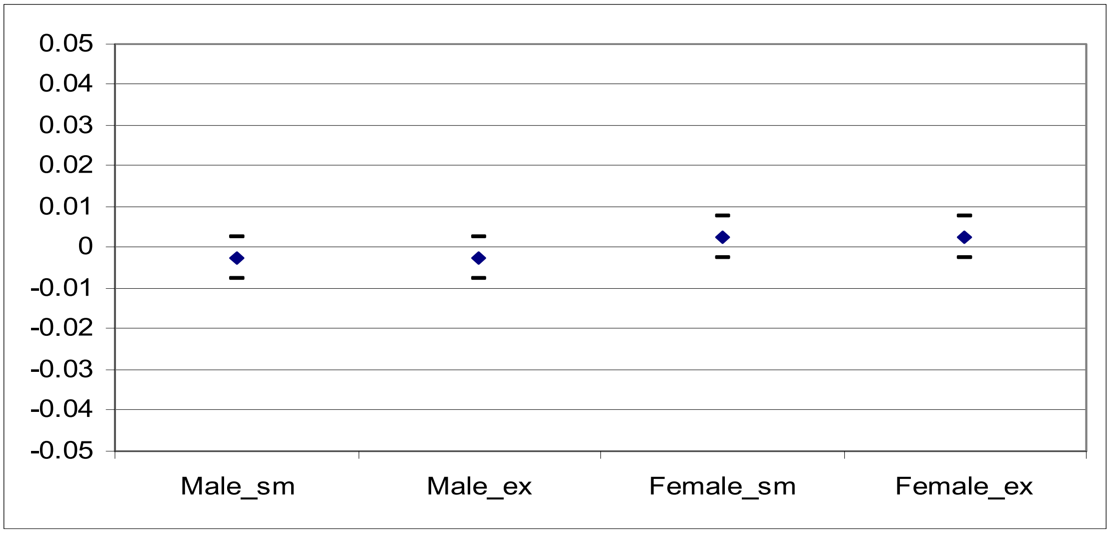


Figure 4.3b: HS-CBS: $P_2(Z,k,\rho_{X,Z})$ for Urbanicity for small (sm) and extended (ex) variable sets

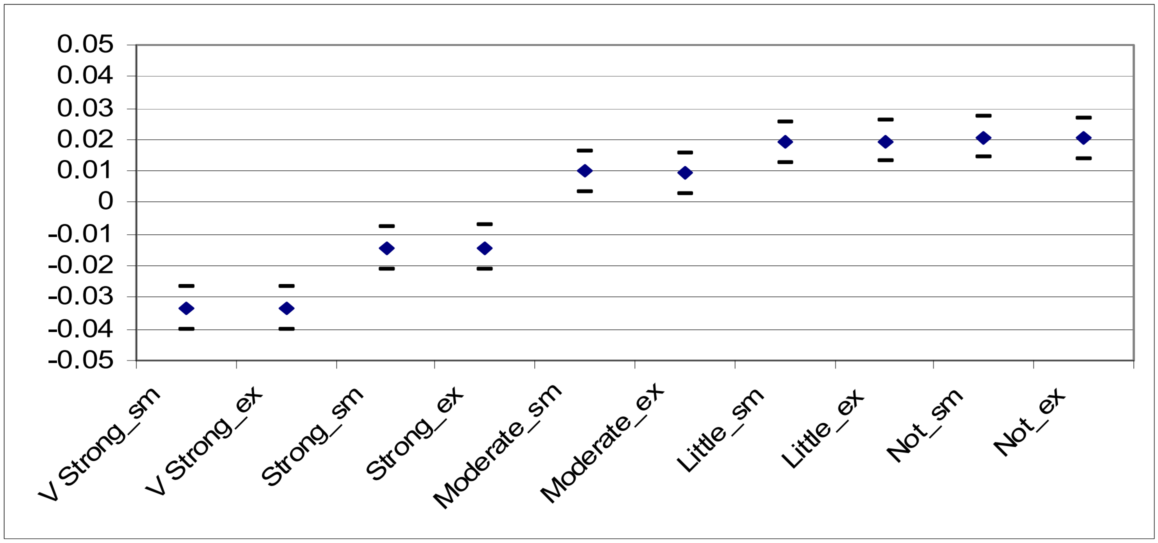
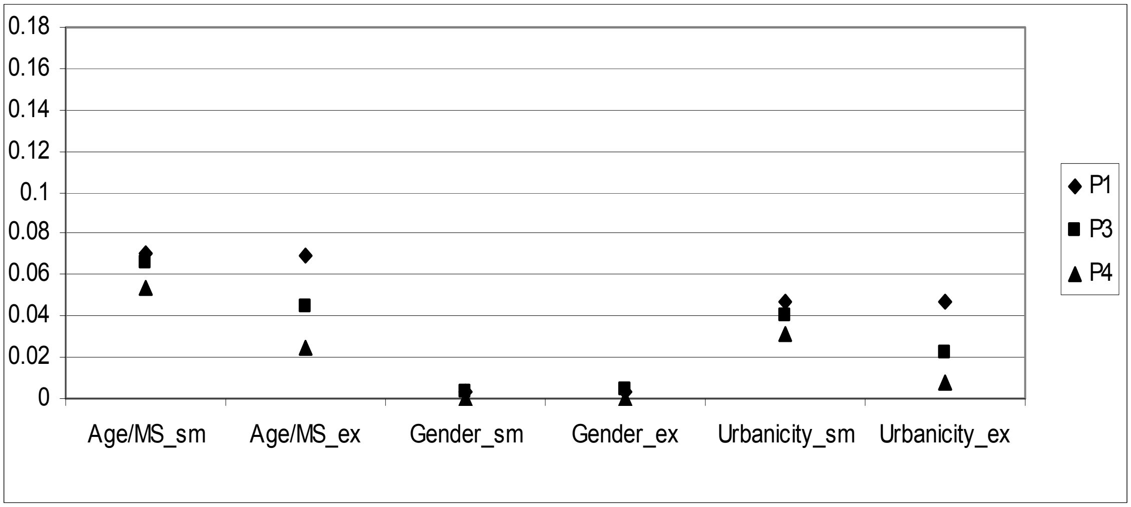


Table 4.4: HS-CBS: partial indicators $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age/Marital Status, Gender and Urbanicity

Partial Indicator	Small Variable Set			Extended Variable Set		
	Age Group, Marital Status	Gender	Urbanicity	Age Group, Marital Status	Gender	Urbanicity
$P_1(Z,\rho_{X,Z})$	0.0705 (0.0635-0.0774)	0.0036 (-0.0008-0.0105)	0.0470 (0.0391-0.0546)	0.0689 (0.0619-0.0759)	0.0037 (-0.0067-0.0140)	0.0469 (0.0392-0.0543)
$P_3(Z,\rho_{X,Z})$	0.0663 (0.0591-0.0733)	0.0035 (-0.0008-0.0104)	0.0400 (0.0325-0.0473)	0.0449 (0.038-0.0509)	0.0043 (0.0001-0.0098)	0.0220 (0.0160-0.0279)
$P_4(Z,\rho_{X,Z})$	0.0536 (0.0424-0.0647)	0.0000 (-0.0007-0.0008)	0.0309 (0.0208-0.0411)	0.0247 (0.0165-0.0328)	0.0004 (-0.0008-0.0016)	0.0074 (0.0030-0.0119)

Figure 4.4: HS-CBS: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ for small (sm) and extended (ex) variable sets by Age/Marital Status, Gender and Urbanicity



C. LFS-UK dataset: Sample size=7,830 overall non-response rate: 82%
Auxiliary variable set: Sex (2), AgeGroup (8), RegionUrban (19)
R-indicator: 0.9282 (STD: 0.0181)

Table 4.5: LFS-UK: $P_2(Z,k,\rho_{X,Z})$ for Age Group and Gender

Category	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Age Group			
16-20	-0.0009	-0.0091	0.0078
21-30	0.0053	-0.0025	0.0136
31-40	-0.0046	-0.0122	0.0030
41-50	0.0025	-0.0057	0.0105
51-60	-0.0006	-0.0084	0.0079
61-70	-0.0000	-0.0077	0.0082
71-80	0.0017	-0.0070	0.0105
81 and over	-0.00377	-0.0124	0.0047
Gender			
Males	-0.0043	-0.0095	0.0013
Females	0.0053	-0.0016	0.0117

Figure 4.5a: LFS-UK: $P_2(Z,k,\rho_{X,Z})$ **for Gender**

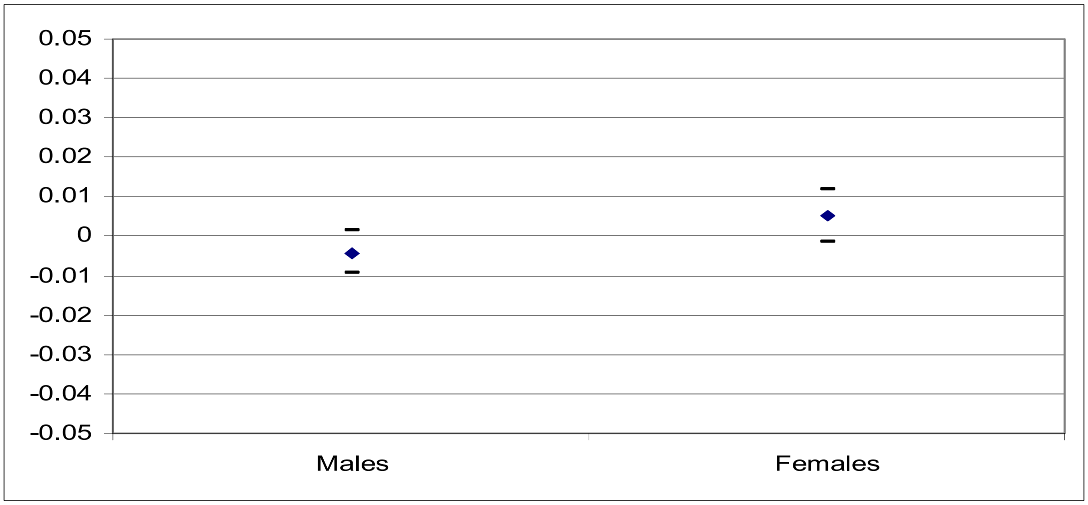


Figure 4.5b: LFS-UK: $P_2(Z,k,\rho_{X,Z})$ **for Age Group**

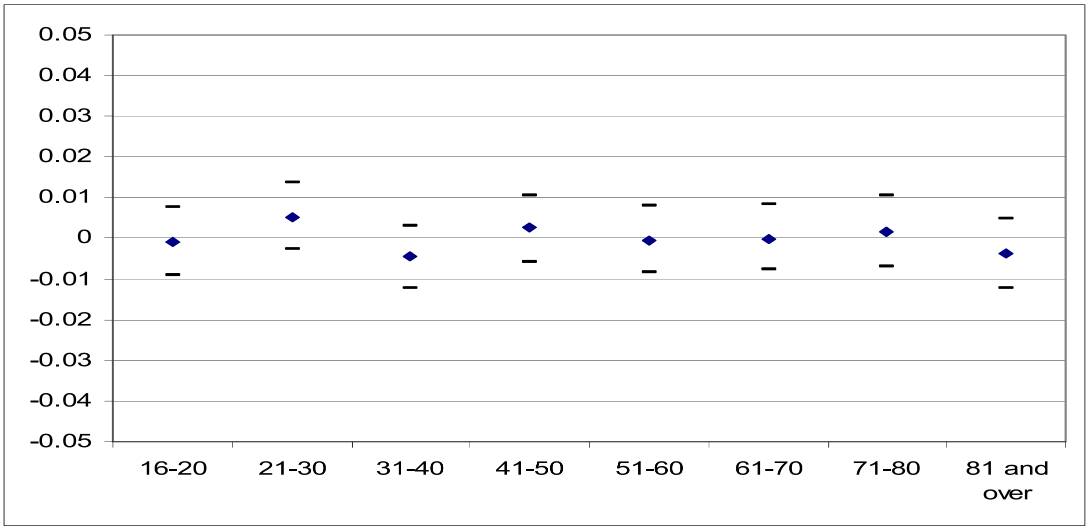


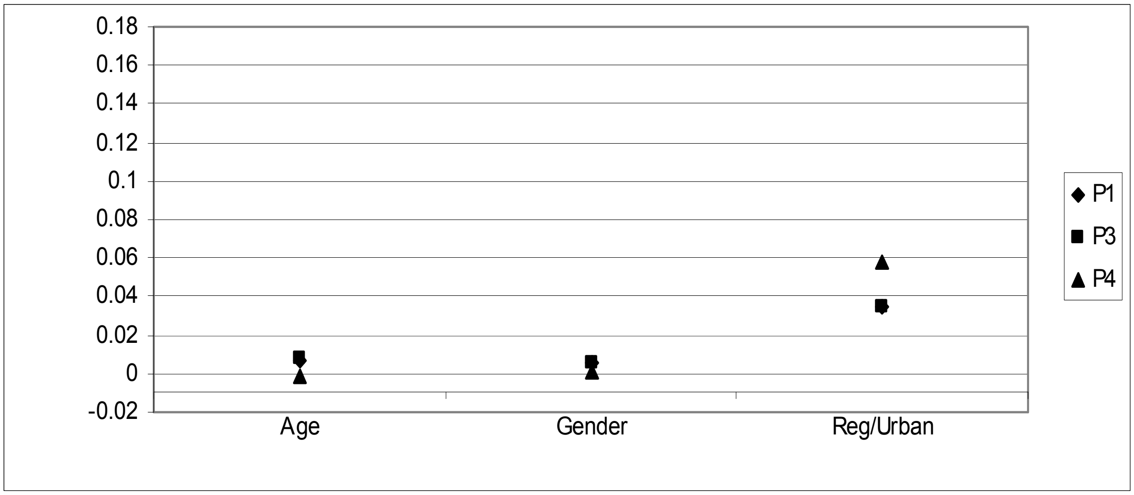
Table 4.5 and 4.6 in conjunction with Figures 4.5 and 4.6 show the results for the conditional and unconditional partial indicators for the UK dataset based on the Labour Force Survey 2001. Note that for all categories of the variables Gender and Age Group there are very small absolute values for P_2. In addition, based on the confidence intervals for the values of P_2, no category of the variables Age Group and Gender seems to be significantly over- or underrepresented. This is in line with the results obtained for the partial indicators P_1, P_3 and P_4, which are depicted in Figure 4.6. Note that all three partial indicators are very close to zero for both variables Age Group and Gender.

Accordingly, these variables do not seem to give any contribution to the lack of representativity. On the other hand, Region or Urbanicity seem to explain most of the bias.

Table 4.6: LFS-UK: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and RegionUrban

Partial Indicator	Age Group	Gender	RegionUrban
$P_1(Z,\rho_{X,Z})$	0.0072 (0.0063-0.0193)	0.0057 (0.0003-0.0137)	0.0346 (0.0305-0.0499)
$P_3(Z,\rho_{X,Z})$	0.0077 (0.0069-0.0195)	0.0058 (0.0004-0.0136)	0.0346 (0.0306-0.0499)
$P_4(Z,\rho_{X,Z})$	-0.0016 (-0.0019-0.0080)	0.0008 (-0.0005-0.0051)	0.0582 (0.0443-0.0859)

Figure 4.6: LFS-UK: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$, $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and RegionUrban



D. ESS-NO dataset: Sample size=2,673 overall response rate: 65.5%
Auxiliary variable set: Age Group (15), Gender (2), RegionUrban (30)
R-Indicator: 0.8828 (STD: 0.0203)

Table 4.7: ESS-NO: $P_2(Z,k,\rho_{X,Z})$ for Age Group and Gender

Category	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Age Group			
20-24	-0.0010	-0.0271	0.0078
25-29	0.0015	-0.0143	0.0179
30-34	0.0009	-0.0151	0.0179
35-39	0.0100	-0.0066	0.0275
40-44	0.0065	-0.0105	0.0235
45-49	0.0117	-0.0037	0.0285
50-54	0.0118	-0.0053	0.0274
55-59	0.0039	-0.0148	0.0204
60-64	0.0200	0.0027	0.0346
65-69	-0.0072	-0.0255	0.0093
70-74	-0.0050	-0.0264	0.0099
75-79	-0.0156	-0.0331	0.0024
80-84	-0.0071	-0.0259	0.0090
85-89	-0.0352	-0.0532	-0.0155
90+	-0.0203	-0.0385	0.0004
Gender			
Males	0.0209	0.0066	0.0338
Females	-0.0203	-0.0330	-0.0063

Figure 4.7a: ESS-NO: $P_2(Z,k,\rho_{X,Z})$ for Gender

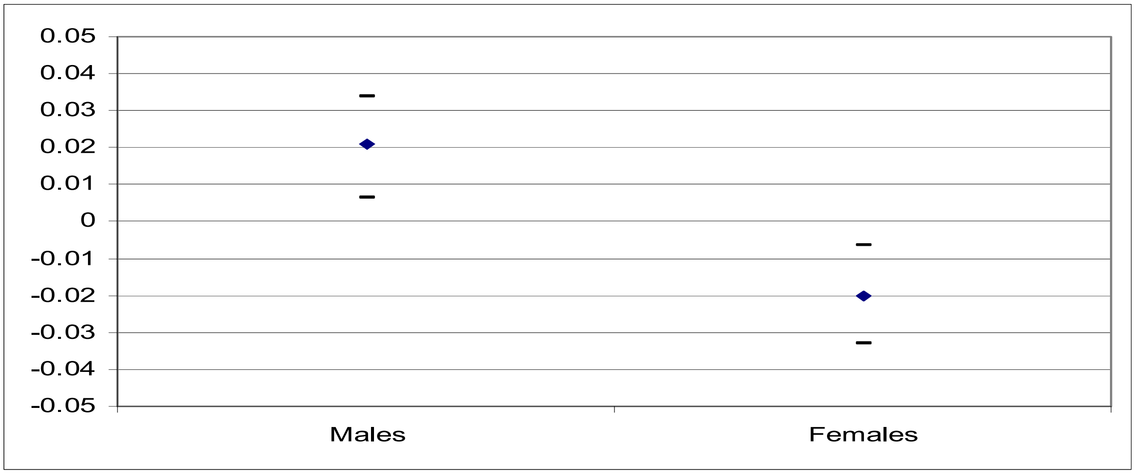


Figure 4.7b: ESS-NO: $P_2(Z,k,\rho_{X,Z})$ for Age Group

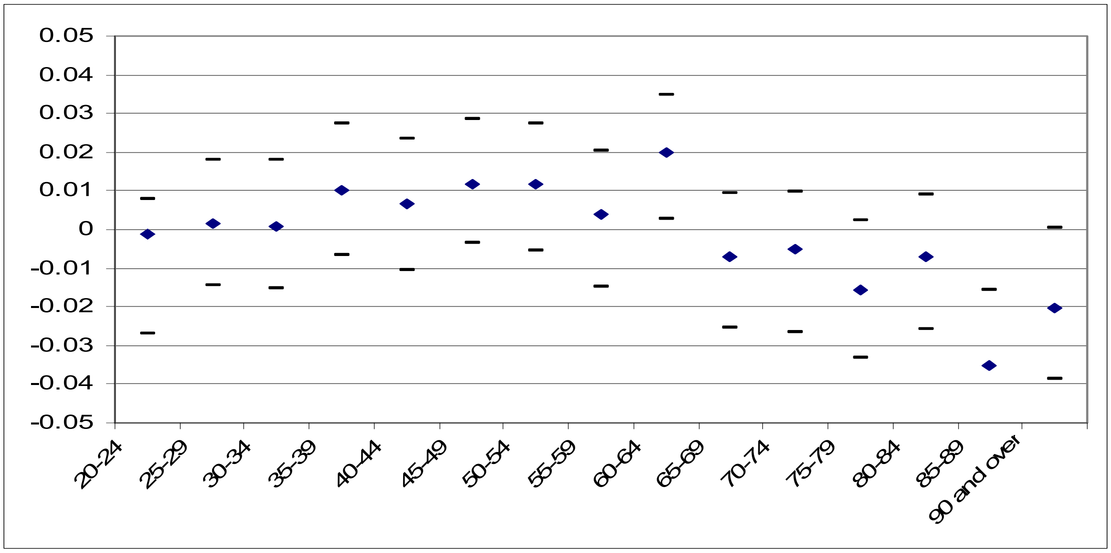
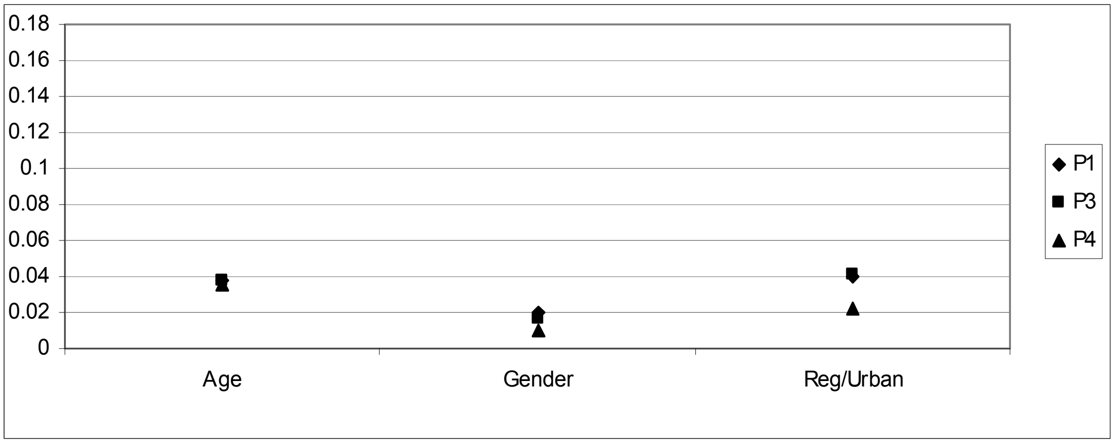


Table 4.8: ESS-NO: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and Region/Urban

Partial Indicator	Age Group	Gender	RegionUrban
$P_1(Z,\rho_{X,Z})$	0.0378 (0.0363-0.0702)	0.0203 (0.0072-0.0395)	0.0404 (0.0444-0.0773)
$P_3(Z,\rho_{X,Z})$	0.0376 (0.0359-0.0682)	0.0165 (0.0046-0.0320)	0.0411 (0.0452-0.0777)
$P_4(Z,\rho_{X,Z})$	0.0359 (0.0160-0.0707)	0.0104 (-0.0004-0.0220)	0.0219 (0.0181-0.0820)

Figure 4.8: ESS-NO: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and Region/Urban



E. LLS-NO dataset: Sample size= 4,837 overall response rate: 69.1% **Auxiliary variable set**: Age Group (11), Gender (2), Urbanicity (37) R-indicator: 0.8722 (STD: 0.0138)

Table 4.9: LLS-NO: $P_2(Z,k,\rho_{X,Z})$ for Categories of Age Group and Gender

Category	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Age Group			
-19	0.0001	-0.0123	0.0112
20-24	-0.0037	-0.0164	0.0081
25-29	-0.0035	-0.0172	0.0080
30-34	0.0025	-0.0106	0.0142
35-39	0.0173	0.0060	0.0281
40-44	0.0155	0.0041	0.0268
45-49	0.0135	0.0014	0.0251
50-54	0.0091	-0.0029	0.0231
55-59	-0.0096	-0.0222	0.0027
60-64	0.0004	-0.0123	0.0129
65+	-0.0315	-0.0438	-0.0205
Gender			
Males	0.0026	-0.0056	0.0115
Females	-0.0026	-0.0115	0.0057

Figure 4.9a: LLS-NO: $P_2(Z,k,\rho_{X,Z})$ for Gender

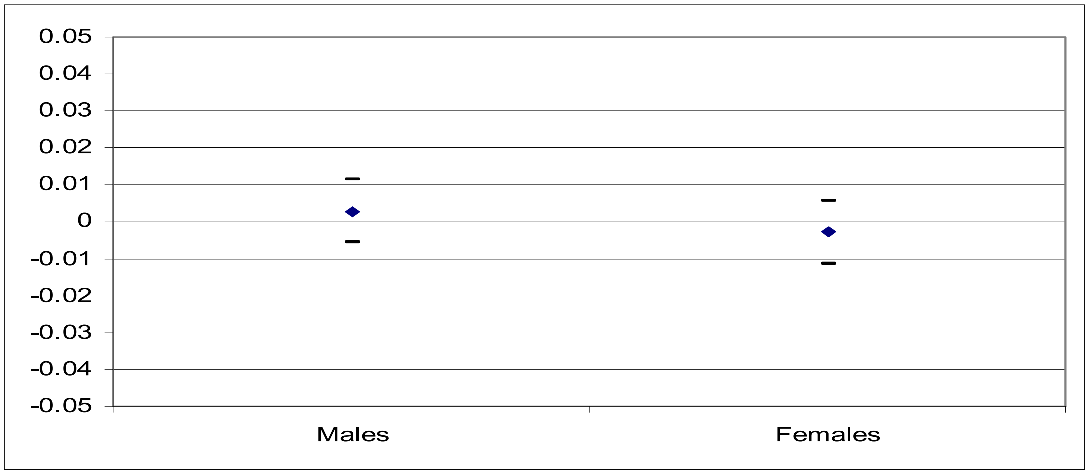


Figure 4.9b: LLS-NO: $P_2(Z, k, \rho_{X,Z})$ for Age Group

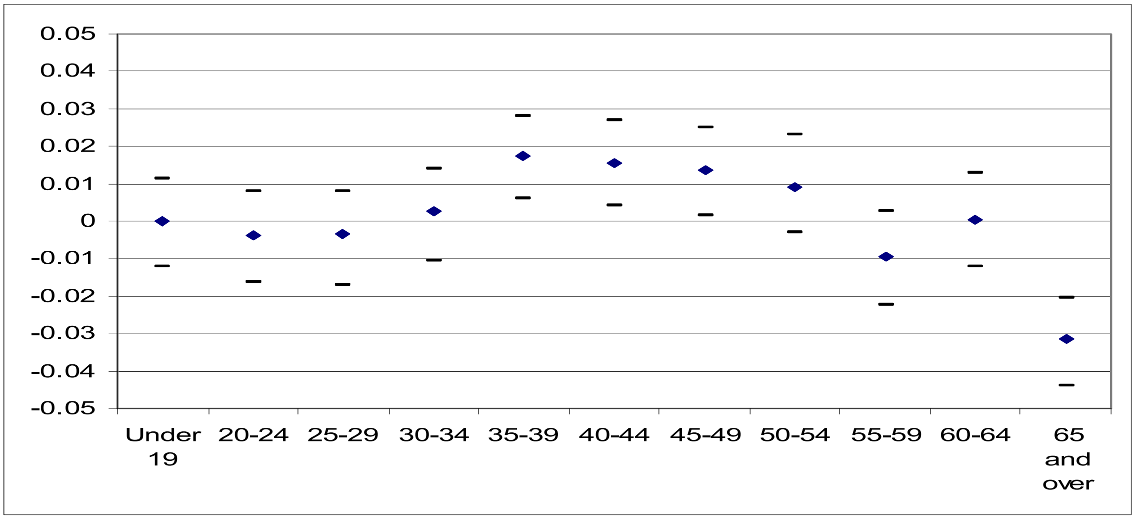
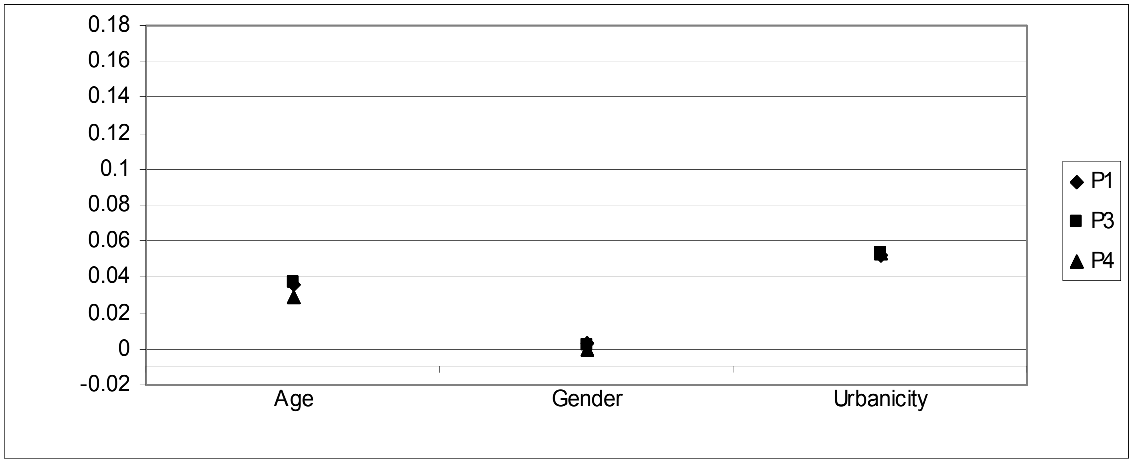


Table 4.10: LLS-NO: $P_1(Z, \rho_{X,Z})$, $P_3(Z, \rho_{X,Z})$ and $P_4(Z, \rho_{X,Z})$ by AgeGroup, Gender and Urbanicity

Partial Indicator	Age Group	Gender	Urbanicity
$P_1(Z, \rho_{X,Z})$	0.0359 (0.0300-0.0551)	0.0030 (0.0002-0.0145)	0.0520 (0.0510-0.0763)
$P_3(Z, \rho_{X,Z})$	0.0367 (0.0307-0.0551)	0.0022 (0.0002-0.0123)	0.0528 (0.0519-0.0772)
$P_4(Z, \rho_{X,Z})$	0.0283 (0.0128-0.0473)	-0.0006 (-0.0006-0.0022)	0.0528 (0.0454-0.0983)

Figure 4.10: LLS-NO $P_1(Z, \rho_{X,Z})$, $P_3(Z, \rho_{X,Z})$ and $P_4(Z, \rho_{X,Z})$ by Age Group, Gender and Urbanicity



F. ESS-BE dataset: Sample size=2,927 overall response rate: 61.4% **Small auxiliary variable set**: Sex (2), Age Group (4), Region (3) R-indicator: 0.8053 **Extended auxiliary variable set:** Gender (2), Age Group (4), Region (3), Apartment (2), Urban (2), Foreign (2)
R-indicator: 0.7982

Table 4.11: ESS-BE: $P_2(Z,k,\rho_{X,Z})$ for Categories of Age Group, Gender and Region

Category	Small Variable Set			Extended Variable Set		
	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Age Group						
-20	0.0204	0.0054	0.0352	0.0215	0.0061	0.0387
24-40	0.0016	-0.0128	0.0148	0.0017	-0.0141	0.0175
40-60	0.0069	-0.0072	0.0197	0.0073	-0.0038	0.0208
60+	-0.0211	-0.0347	-0.0072	-0.0223	-0.0410	-0.0082
Gender						
Males	-0.0090	-0.0209	0.0038	-0.0094	-0.0231	0.0062
Females	0.0086	-0.0036	0.0201	0.0090	-0.0060	0.0223
Region						
Flanders	0.0243	0.0135	0.0352	0.0256	0.0144	0.0342
Brussels	-0.0647	-0.0807	-0.0492	-0.0682	-0.0839	-0.0529
Wallonia	0.0006	-0.0133	0.0155	0.0007	-0.0136	0.0149

Figure 4.11a: ESS-BE: $P_2(Z,k,\rho_{X,Z})$ for Gender, for small (sm) and extended (ex) auxiliary sets

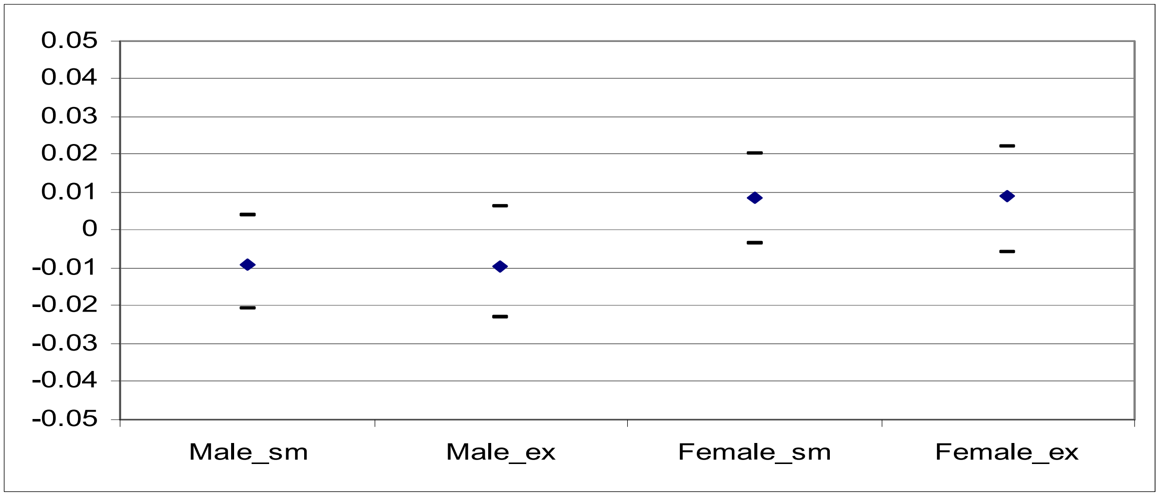


Figure 4.11b: ESS-BE: $P_2(Z,k,\rho_{X,Z})$ for Age Group for small (sm) and extended (ex) auxiliary sets

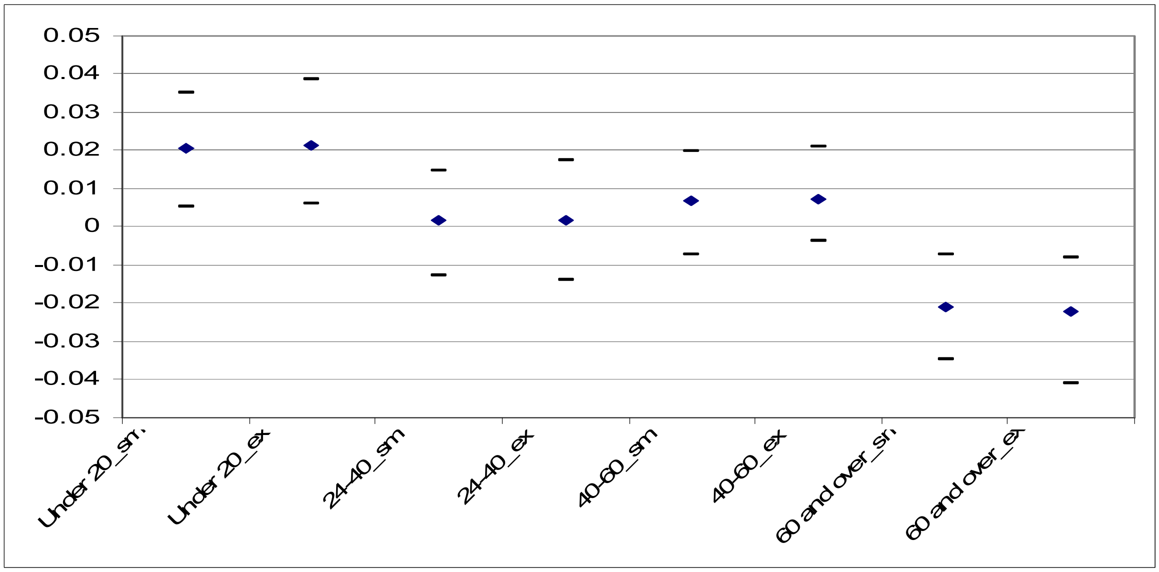
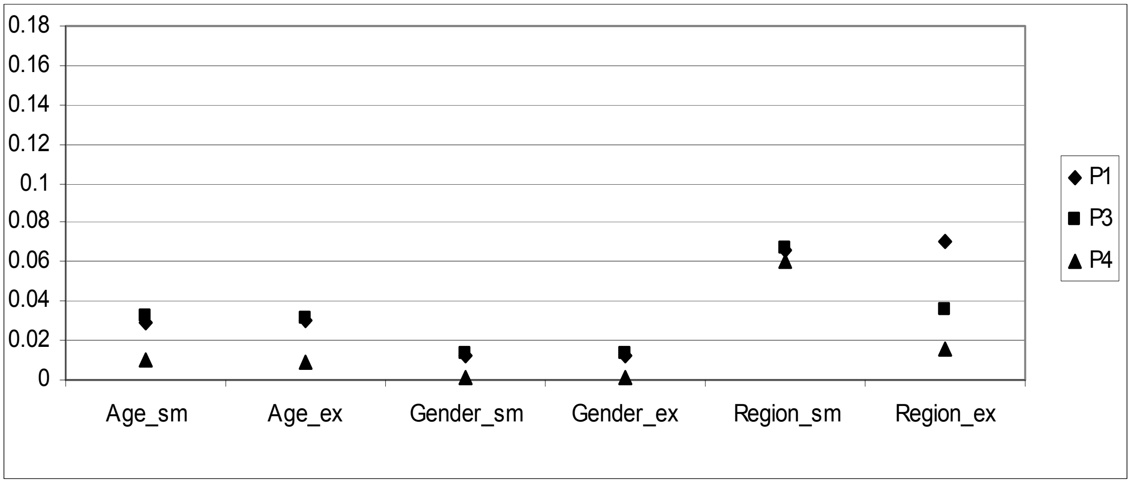


Table 4.12: ESS-BE: partial indicators $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and Region

Partial Indicator	Small Variable Set			Extended Variable Set		
	Age Group	Gender	Region	Age Group	Gender	Region
$P_1(Z,\rho_{X,Z})$	0.0290 (0.0167-0.0451)	0.0119 (0.0009-0.0280)	0.0662 (0.0508-0.0841)	0.0306 (0.0186-0.0514)	0.0125 (0.0009-0.0311)	0.0700 (0.0547-0.0869)
$P_3(Z,\rho_{X,Z})$	0.0323 (0.0201-0.0481)	0.0132 (0.0010-0.0288)	0.0676 (0.0517-0.0856)	0.0312 (0.0186-0.0463)	0.0138 (0.0008-0.0307)	0.0360 (0.0231-0.0531)
$P_4(Z,\rho_{X,Z})$	0.0101 (0.0025-0.0235)	0.0012 (-0.0007-0.0085)	0.0608 (0.0371-0.0868)	0.0088 (0.0012-0.0270)	0.0014 (-0.0009-0.0098)	0.0152 (0.0045-0.0320)

The above show the results based on the country dataset of Belgium, for both a small and extended set of auxiliary variables. The results for the partial indicator $P_2(Z,k,\rho_{X,Z})$ are similar for both sets of variables, which is obvious from Table 4.11 and Figures 4.11a and 4.11b. The age group category 'under 20' seems to be overrepresented while for the category '60 and over' there is sign of underrepresentation. From Figure 4.12 it is clear that the variables Region explains most of the lack of representativity, while Gender does not seem to have any significant contribution. The results are consistent for all partial indicators.

Figure 4.12: ESS-BE: $P_1(Z, \rho_{X,Z})$, $P_3(Z, \rho_{X,Z})$ and $P_4(Z, \rho_{X,Z})$ for **Age Group, Gender and Region for small and extended auxiliary variable sets**



G. LFS-SLO dataset: Sample size= 2,219 overall response rate: 69%
Auxiliary variable set: Gender (2), Age Group (7), Region (3),
R-indicator: 0.8155 (STD: 0.0199)

Table 4.13: LFS-SLO $P_2(Z, k, \rho_{X,Z})$ **for Categories of Age Group and Gender**

Category	$P_2(Z, k, \rho_{X,Z})$	Lower CI	Upper CI
Age Group			
18-29	-0.0052	-0.0216	0.0117
30-39	-0.0062	-0.0237	0.0114
40-49	-0.0107	-0.0286	0.0057
50-59	0.0108	-0.0066	0.0279
60-69	0.0099	-0.0071	0.0274
70-79	0.0123	-0.0069	0.0296
80+	-0.0105	-0.0306	0.0095
Gender			
Males	0.0075	-0.0057	0.0212
Females	-0.0074	-0.0208	0.0057

Figure 4.13a: LFS-SLO: $P_2(Z,k,\rho_{X,Z})$ for Gender

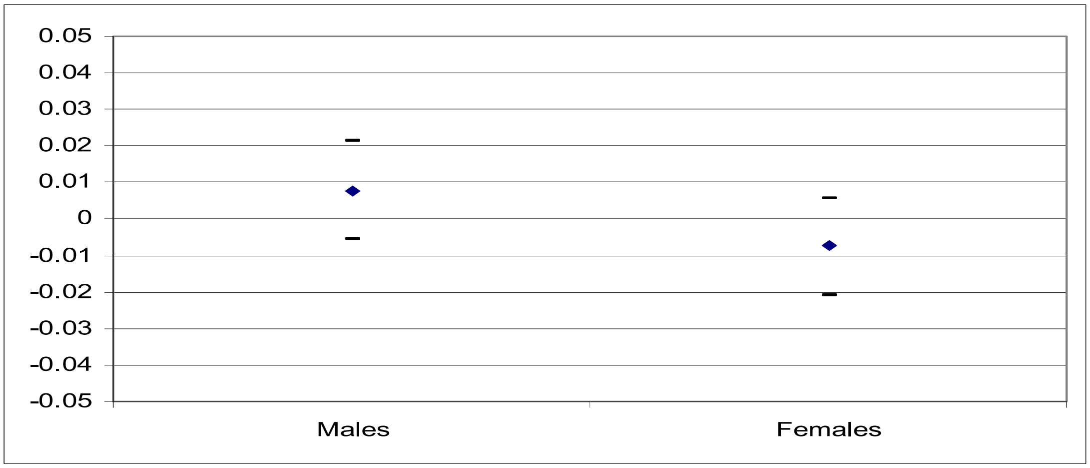


Figure 4.13b: LFS-SLO $P_2(Z,k,\rho_{X,Z})$ for Age Group

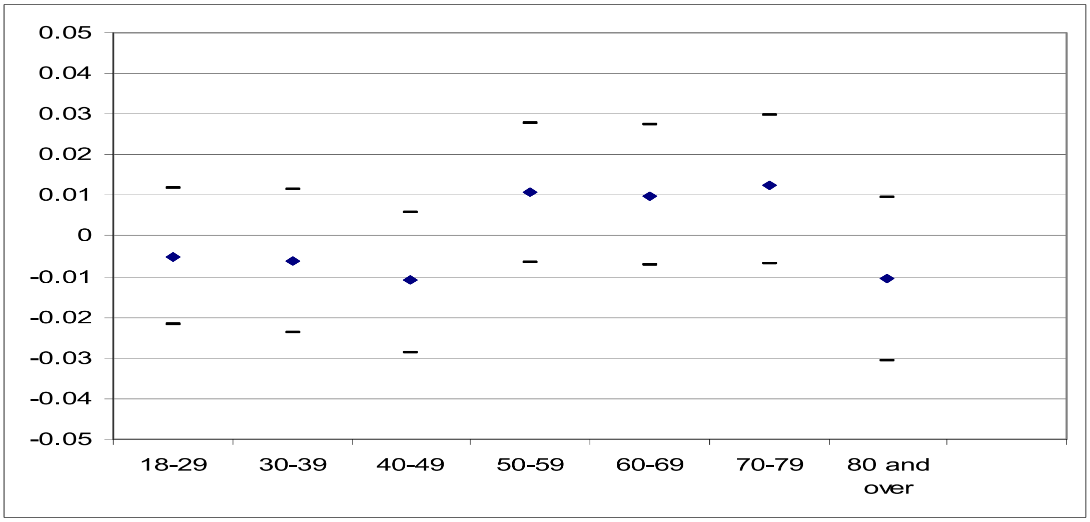
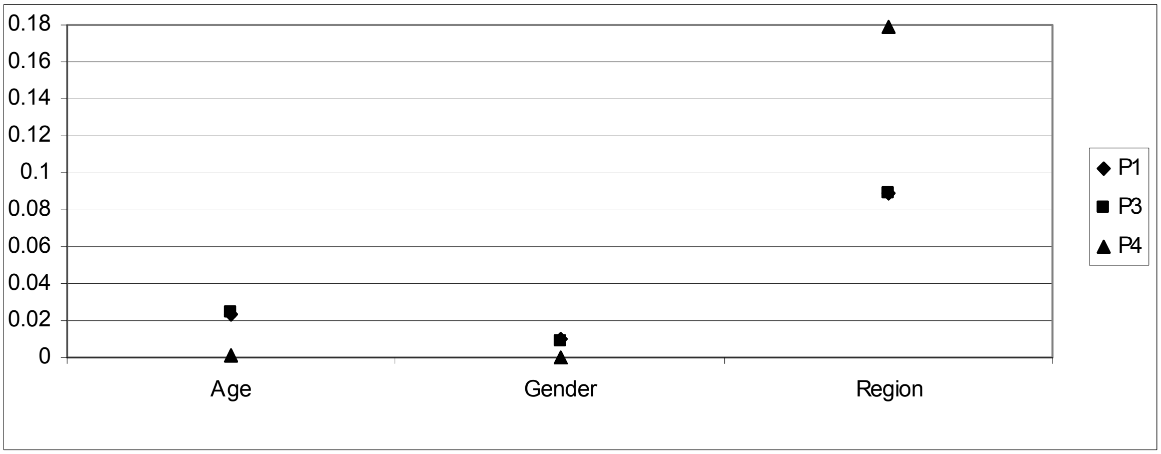


Table 4.14: LFS-SLO: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Age Group, Gender and Region

Partial Indicator	Age Group	Gender	Region
$P_1(Z,\rho_{X,Z})$	0.0234 (0.0170-0.0484)	0.0096 (0.0006-0.0274)	0.0885 (0.0745-0.1123)
$P_3(Z,\rho_{X,Z})$	0.0241 (0.0177-0.0492)	0.0094 (0.0006-0.0263)	0.0884 (0.0747-0.1124)
$P_4(Z,\rho_{X,Z})$	0.0015 (-0.0019-0.0217)	0.0002 (-0.0010-0.0074)	0.1790 (0.1029-0.2009)

Figure 4.14: LFS-SLO: $P_1(Z, \rho_{X,Z})$, $P_3(Z, \rho_{X,Z})$ and $P_4(Z, \rho_{X,Z})$ by Age Group, Gender and Region



4.2 Business Data

H. STS-IND-CBS dataset: Sample size= 64,413 overall response rate: 78.7% after 30 days

Small auxiliary variable set: Business type (23), Business size (5) R-Indicator: 0.933 (CI: 0.927-0.940)

Extended auxiliary variable set: Business type (23), Business size *VAT (28) R-Indicator: 0.918 (CI: 0.913-0.922)

Figure 4.15: STS-IND-CBS: $P_2(Z, k, \rho_{X,Z})$ for Business Type for small (sm) and extended (ex) auxiliary sets

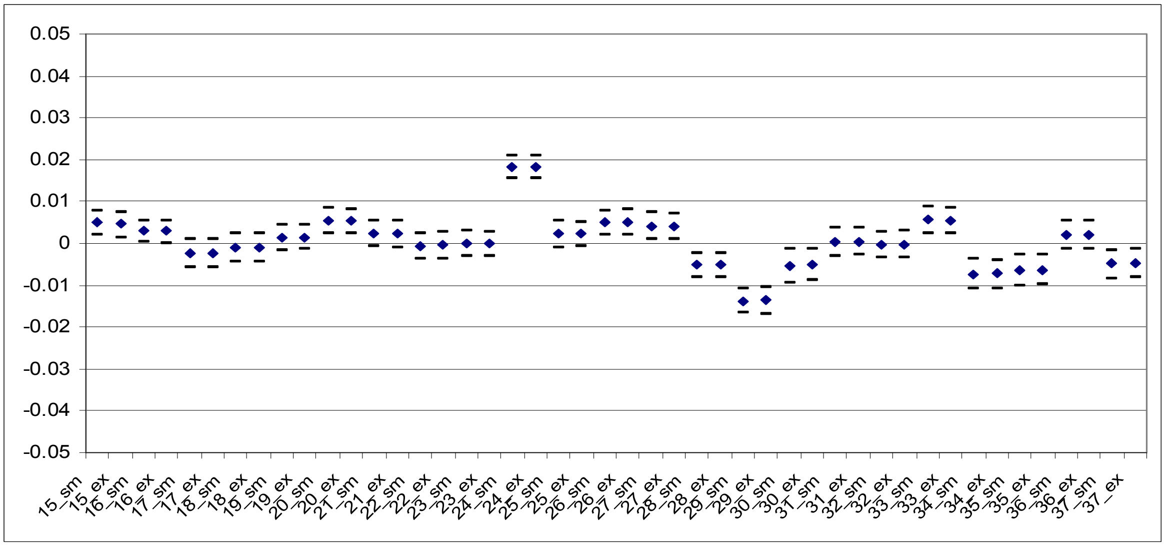


Table 4.15: STS-IND-CBS: $P_2(Z,k,\rho_{X,Z})$ for Categories of Business Type

Category	Small Variable Set			Extended Variable Set		
	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Business Type						
15	0.0050	0.0019	0.0078	0.0047	0.0015	0.0076
16	0.0029	0.0002	0.0055	0.0029	0.0000	0.0054
17	-0.0024	-0.0057	0.0009	-0.0023	-0.0057	0.0009
18	-0.0011	-0.0043	0.0024	-0.0011	-0.0045	0.0023
19	0.0014	-0.0018	0.0044	0.0014	-0.0015	0.0043
20	0.0054	0.0025	0.0083	0.0053	0.0025	0.0082
21	0.0025	-0.0007	0.0054	0.0023	-0.0009	0.0053
22	-0.0006	-0.0038	0.0024	-0.0004	-0.0036	0.0028
23	0.0000	-0.0029	0.0029	0.0000	-0.0031	0.0028
24	0.0183	0.0157	0.0209	0.0181	0.0155	0.0208
25	0.0022	-0.0011	0.0053	0.0022	-0.0006	0.0051
26	0.0051	0.0021	0.0078	0.0050	0.0021	0.0080
27	0.0041	0.0011	0.0073	0.0040	0.0009	0.0071
28	-0.0052	-0.0082	-0.0023	-0.0051	-0.0082	-0.0022
29	-0.0137	-0.0167	-0.0107	-0.0136	-0.0169	-0.0105
30	-0.0054	-0.0096	-0.0015	-0.0051	-0.0089	-0.0012
31	0.0004	-0.0029	0.0036	0.0004	-0.0027	0.0037
32	-0.0003	-0.0033	0.0028	-0.0003	-0.0034	0.0029
33	0.0056	0.0023	0.0087	0.0053	0.0022	0.0085
34	-0.0073	-0.0107	-0.0038	-0.0072	-0.0107	-0.0040
35	-0.0064	-0.0100	-0.0027	-0.0063	-0.0097	-0.0027
36	0.0021	-0.0013	0.0054	0.0021	-0.0013	0.0054
37	-0.0048	-0.0083	-0.0017	-0.0048	-0.0080	-0.0012

Table 4.16: STS-IND-CBS: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Business Type

Partial Indicator	Small Variable Set	Extended Variable Set
	Business Type	
$P_1(Z,\rho_{X,Z})$	0.0293 (0.0264-0.0323)	0.0289 (0.0259-0.0318)
$P_3(Z,\rho_{X,Z})$	0.0264 (0.0235-0.0297)	0.0255 (0.0224-0.0286)
$P_4(Z,\rho_{X,Z})$	0.0175 (0.0133-0.0217)	0.0145 (0.0106-0.0184)

Figure 4.16: STS-IND-CBS: small and extended auxiliary variable set $P_1(Z, \rho_{X,Z})$, $P_3(Z, \rho_{X,Z})$ and $P_4(Z, \rho_{X,Z})$ by Business Type

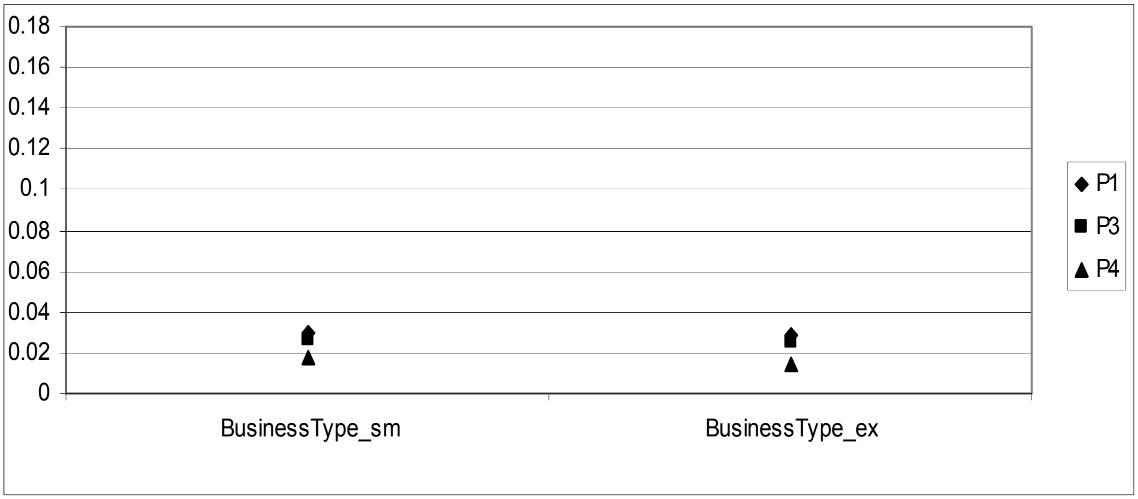


Figure 4.15 and Table 4.15 show the results for the partial indicator P_2 for the Dutch Short Term Statistics on Industry 2007 survey. The results are shown for both a small and extended auxiliary dataset. The small set contains the variables Business type (23 categories) and Business size (5 categories). The extended set includes the interaction term Business size*VAT (28 categories) as a substitute of Business Size to capture the effects of value added tax. The partial indicator P_2 shows consistent results for both the small and extended auxiliary variable set. Several categories of Business Type are under- or overrepresented in this business survey. In view of that, this variable seems to have a significant contribution to the lack of representativity. The same conclusion can be drawn by the results depicted in Figure 4.16. Even when correcting for the effect of other variables (conditioning on the other variables), Business Type explains a part of the bias or lack of representativity. Also for the conditional partial indicators, the results are consistent for both the small and extended variable sets.

The next figures show the results for the Dutch Short Term Statistics on Retail 2007 survey. Similar conclusions can be drawn for this survey compared to the STS-IND survey. Once more, the variable Business Type has a (small) contribution to the lack of representativity. Furthermore, it is possible to indicate the different categories which are under- or overrepresented in the survey.

STS-RET-CBS dataset: Sample size= 93,799 overall response rate: 78.0% after 30 days

Small auxiliary variable set: Business type (7), Business size (9)
R-Indicator: 0.946 (CI: 0.940-0.952)

Extended auxiliary variable set: Business type (7), Business size*VAT (42)
R-Indicator: 0.879 (CI: 0.873-0.886)

Table 4.17: STS-RET-CBS $P_2(Z,k,\rho_{X,Z})$ for Categories of Business Type

Category	Small Variable Set			Extended Variable Set		
	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Business Type						
521	-0.0001	-0.0027	0.0025	-0.0006	-0.0035	0.0021
522	0.0104	0.0074	0.0131	0.0094	0.0065	0.0122
523	-0.0016	-0.0043	0.0013	-0.0020	-0.0050	0.0008
524	-0.0021	-0.0040	0.0001	-0.0041	-0.0062	-0.0021
525	0.0074	0.0040	0.0107	0.0071	0.0039	0.0104
526	-0.0069	-0.0096	-0.0044	-0.0026	-0.0052	0.0002
527	-0.0008	-0.0063	0.0051	-0.0000	-0.0057	0.0055

Figure 4.17: STS-RET-CBS: $P_2(Z,k,\rho_{X,Z})$ for Business Type for small (sm) and extended (ex) auxiliary variable sets

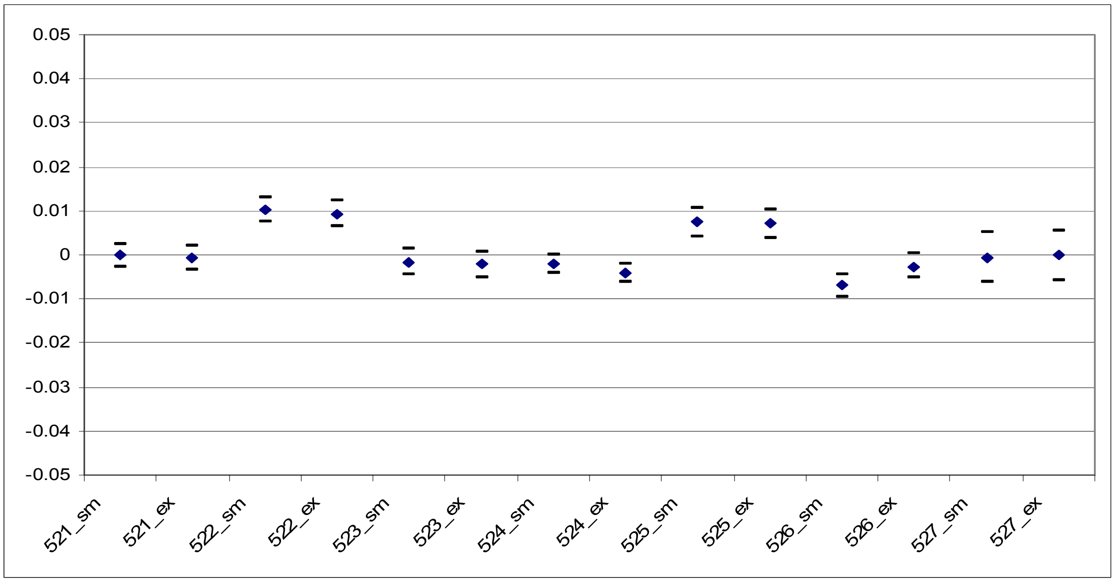
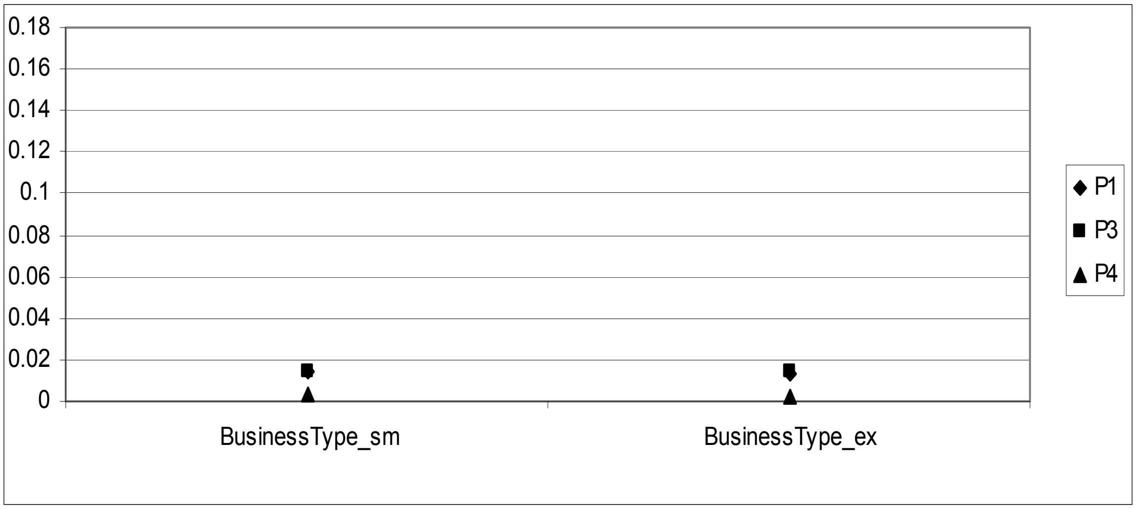


Table 4.18: STS-RET-CBS: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Business Type

Partial Indicator	Small Variable Set	Extended Variable Set
	Business Type	
$P_1(Z,\rho_{X,Z})$	0.0147 (0.0116-0.0178)	0.0129 (0.0097-0.0178)
$P_3(Z,\rho_{X,Z})$	0.0140 (0.0109-0.0173)	0.0142 (0.0111-0.0175)
$P_4(Z,\rho_{X,Z})$	0.0037 (0.0020-0.0054)	0.0023 (0.0012-0.0034)

Figure 4.18: STS-RET-CBS: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ for Business Type for small and extended auxiliary variable sets



In the subsequent, the results for the Slovenian business survey are presented for the auxiliary variable set containing the variables Business Type (14 categories) and Business Size (4 categories). The sample size equals 1,998 and an overall response rate of 88 percent was achieved. The estimated R-indicator is equal to 0.8548. The calculated values for partial indicator P_2 are illustrated in Figure 4.19a and 4.19b for the variables Business Size and Business type, consecutively. The partial indicators P_1, P_3 and P_4 are presented in Table 4.20 and Figure 4.20. The partial indicator P_2 show that the micro businesses are underrepresented in the survey, while medium and large business are overrepresented. Similar to the business surveys in the Netherlands, P_2 indicates several categories of Business Type that are under- or overrepresented.

ICT-SLO: Sample size= 1,998 overall response rate: 88%
Auxiliary variable set: Business type (14), Business size (4)
R-Indicator: 0.8548 (STD:0.0144)

Figure 4.19a: ICT-SLO: $P_2(Z, k, \rho_{X,Z})$ for Size of Business

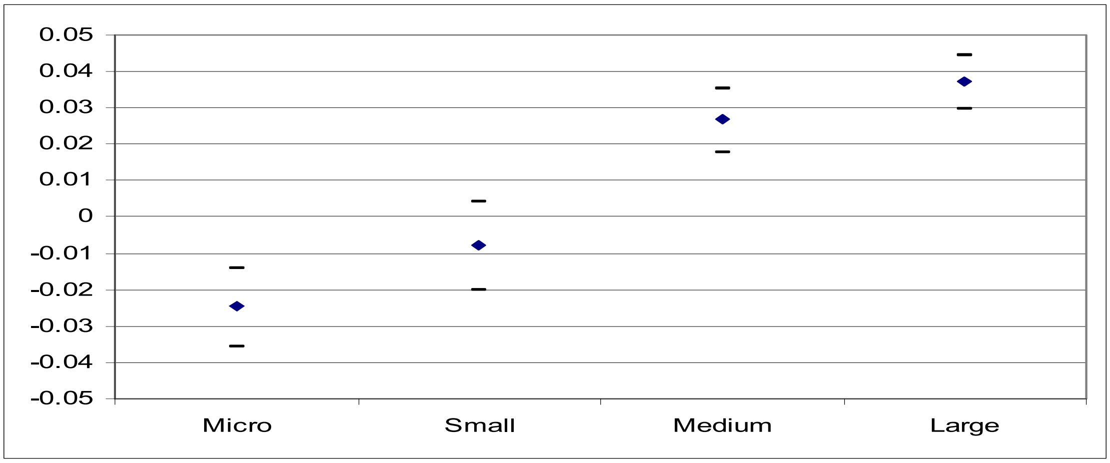


Figure 4.19b: ICT-SLO: $P_2(Z, k, \rho_{X,Z})$ for Type of Business

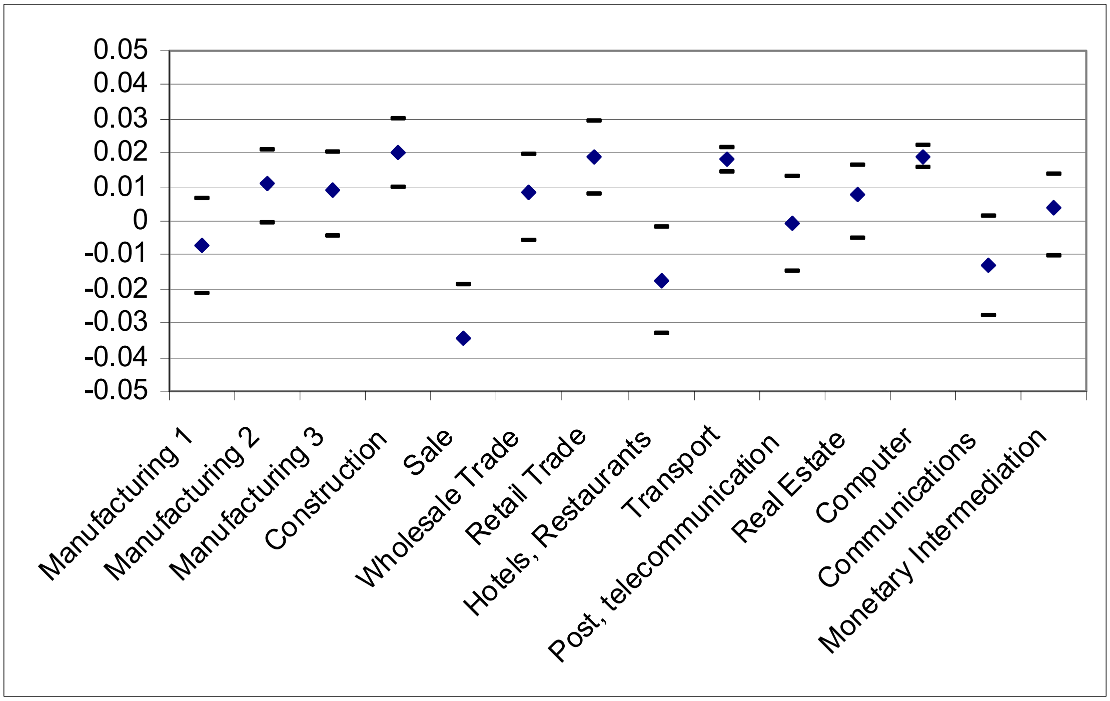


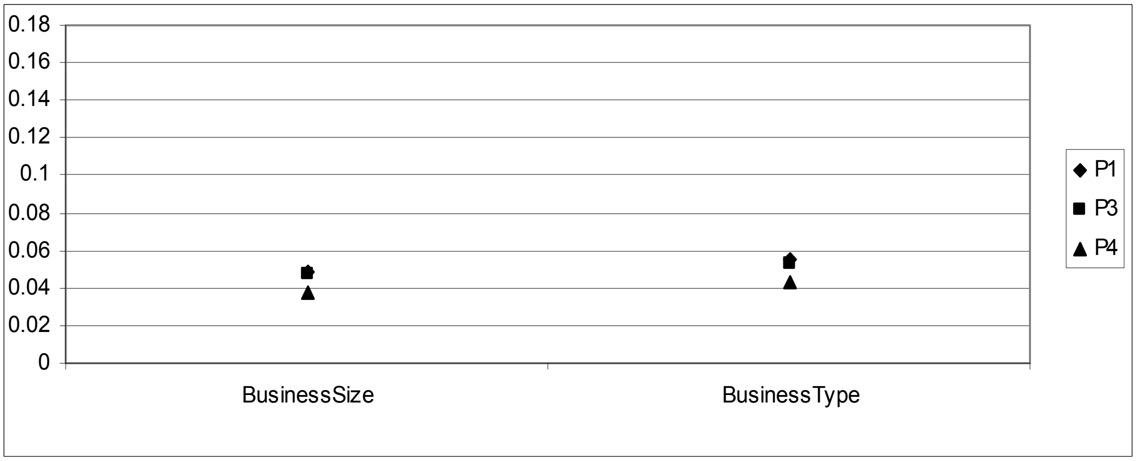
Table 4.19: ICT-SLO $P_2(Z,k,\rho_{X,Z})$ for Categories of Size and Type of Business

Category	$P_2(Z,k,\rho_{X,Z})$	Lower CI	Upper CI
Size of Business			
Micro	-0.0246	-0.0355	-0.0143
Small	-0.0078	-0.0203	0.0040
Medium	0.0269	0.0177	0.0353
Large	0.0371	0.0296	0.0444
Type of Business			
Manufacturing 1	-0.0074	-0.0214	0.0067
Manufacturing 2	0.0109	-0.0009	0.0206
Manufacturing 3	0.0088	-0.0048	0.0203
Construction	0.0202	0.0095	0.0299
Sale	-0.0343	-0.0509	-0.0190
Wholesale Trade	0.0084	-0.0058	0.0197
Retail Trade	0.0191	0.0080	0.0294
Hotels, Restaurants	-0.0173	-0.0331	-0.0018
Transport	0.0179	0.0146	0.0214
Post, telecommunication	-0.0006	-0.0151	0.0129
Real Estate	0.0077	-0.0055	0.0163
Computer	0.0190	0.0155	0.0222
Communications	-0.0133	-0.0277	0.0016
Monetary Intermediation	0.0039	-0.0107	0.0138

Table 4.20: ICT-SLO: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Size and Type of Business

Partial Indicator	Size of Business	Type of Business
$P_1(Z,\rho_{X,Z})$	0.0491 (0.0401-0.6056)	0.0553 (0.0461-0.0736)
$P_3(Z,\rho_{X,Z})$	0.0470 (0.0374-0.6023)	0.0534 (0.0442-0.0724)
$P_4(Z,\rho_{X,Z})$	0.0371 (0.0230-0.0536)	0.0429 (0.0291-0.0723)

Figure 4.20: ICT-SLO: $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ by Size and Type of Business



5. Conclusions from the Country Datasets

The evaluation of representatitivity of the country datasets in Section 4 had two goals: comparison of partial indicators across country datasets and investigating the impact of the response model (small versus extended variable sets).

With respect to the first goal, we compare some results of the household country datasets presented in Section 4 based on the small auxiliary variable set to estimate response propensities. Figure 5.1 compares the R-indicator with the partial indicator $P_1(Z,\rho_{X,Z})$ for the variable Region/Urbanicity on the household country datasets. As can be seen, there is no clear pattern with respect to high R-indicators resulting in high partial indicators for P_1 of Region/Urbanicity. A high R-indicator means less variability in the response probabilities, i.e. smaller variance, but the decomposition of this variance to calculate the between variance of P_1 has mixed results across the country datasets where some datasets have high P_1 and some have low P_1. Figure 5.2 compares the R-indicators with the partial indicator $P_2(Z,k,\rho_{X,Z})$ for the category of males on the country household datasets. Figure 5.2 shows that high R-indicators do not necessarily result in more representativity of males, i.e. a high P_2.

Figure 5.1: R-indicator and Partial Indicator $P_1(Z,\rho_{X,Z})$ on Region/Urbanicity for Country Household Datasets on Small Auxiliary Set

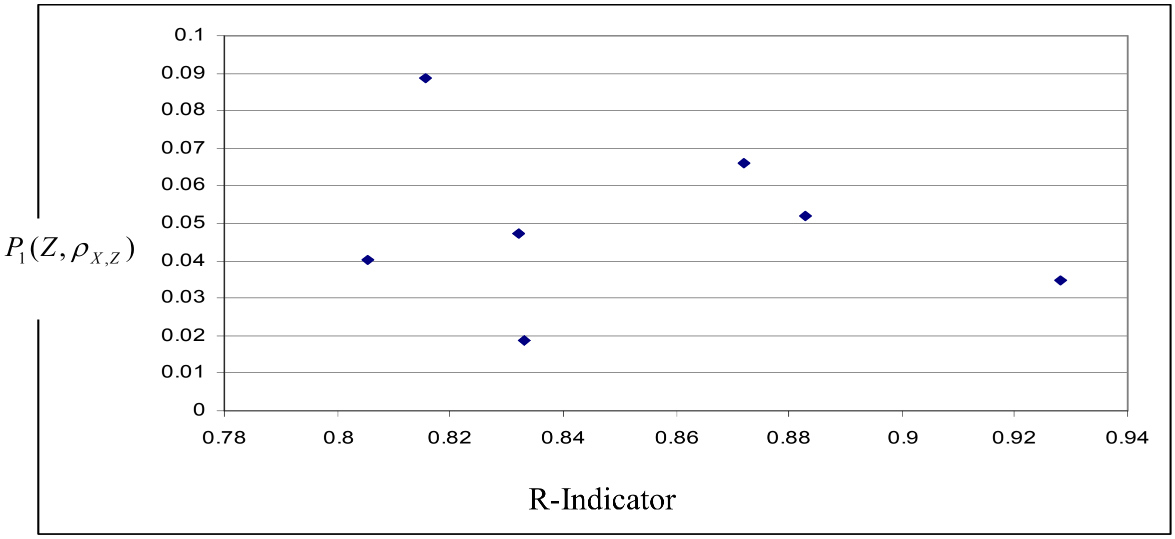


Figure 5.2: R-indicator and Partial Indicator $P_2(Z,k,\rho_{X,Z})$ for Males for Country Household Datasets on Small Auxiliary Set

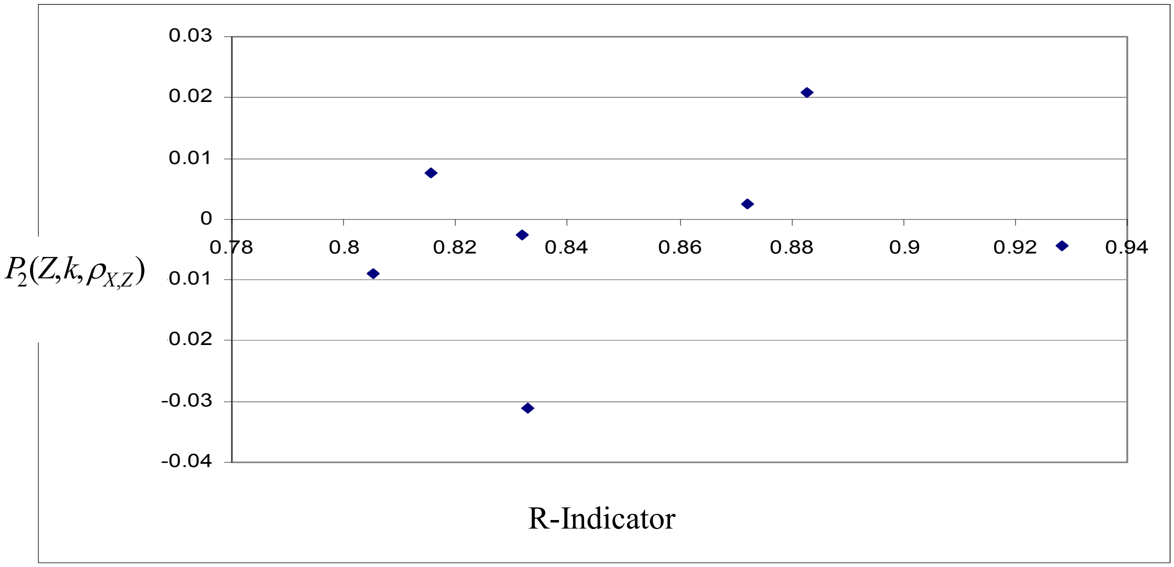


Figure 5.3 compares the R-indicator with the maximal absolute bias $\Delta B_m(Z \mid X)$ of the variable Region/Urbanicity and Figure 5.4 compares the R-indicator with the maximal absolute bias $\Delta B_m(Z \mid X)$ for the variable Age Group based on the household country datasets. With the exception of LFS-SLO which has a low R-indicator and a high maximal absolute bias for Region/Urbanicity and a low maximal absolute bias for Age

Group, other country datasets demonstrate that high R-indicators are associated with lower maximal absolute bias in Age Group and a higher maximal absolute bias in Region/Urbanicity.

Figure 5.3: R-indicator and Difference in Maximal Absolute Bias $\Delta B_m(Z \mid X)$ of Region/Urbanicity for Country Household Datasets on Small Auxiliary Set

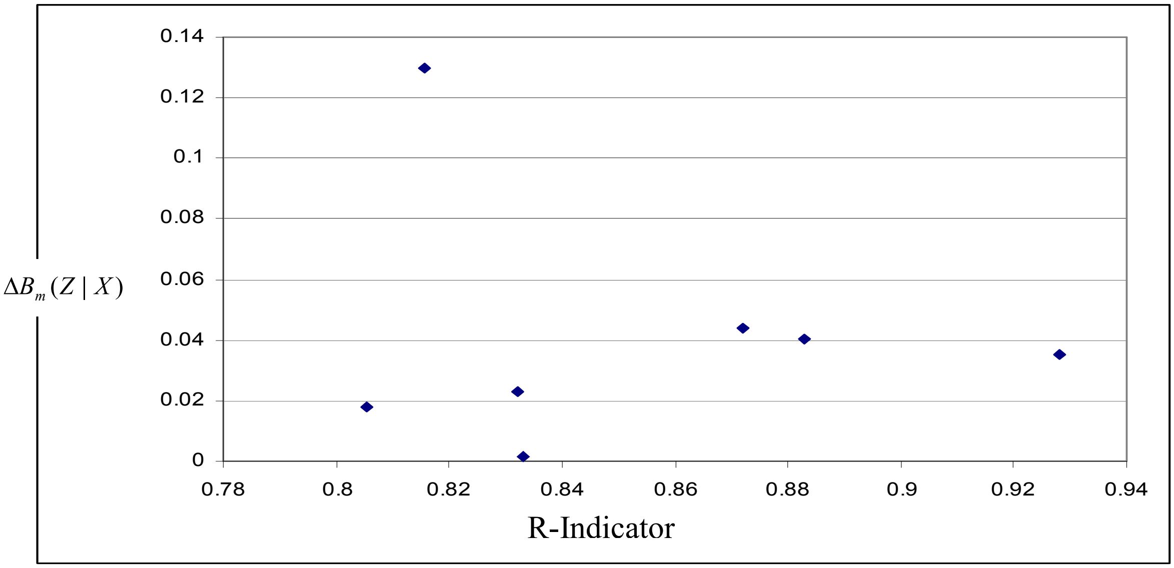
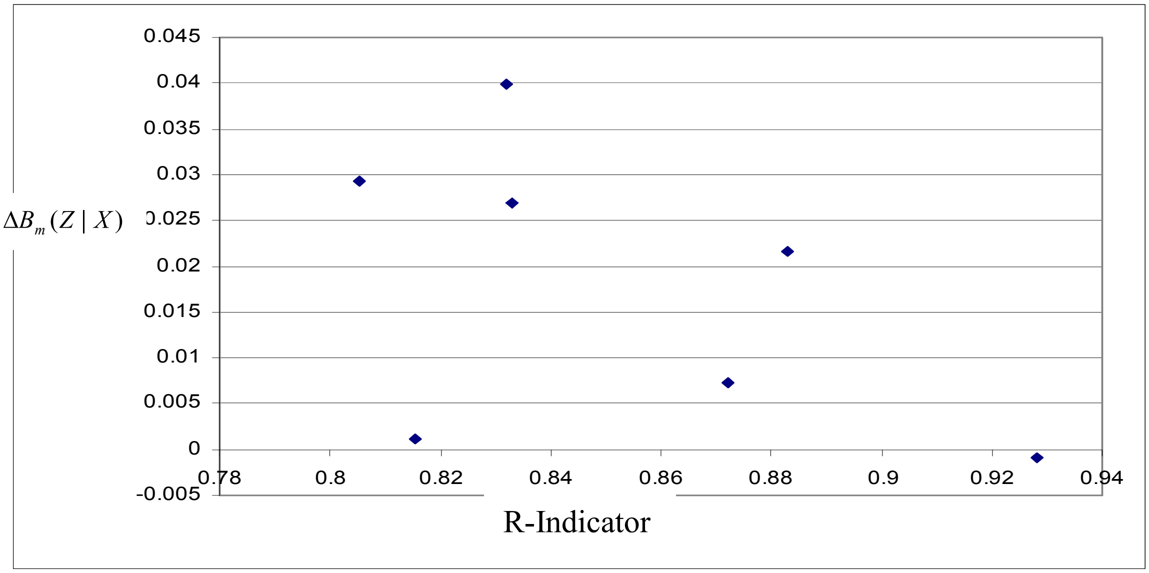


Figure 5.4: R-indicator and Difference in Maximal Absolute Bias $\Delta B_m(Z \mid X)$ of Age Group for Country Household Datasets on Small Auxiliary Set



In summary, Figures 5.1 to 5.4 conclude that there is a need for both R-indicators and partial indicators to fully understand where the lack of representativity is arising from in assessing survey quality and that the association between the R-indicator and partial indicators is mixed. In addition, it is clear that the lack of representativity for specific variables and their categories vary across country datasets which is likely due to different definitions and response rates.

With respect to the second goal, not all of the country datasets provided results of partial indicators for both small and extended auxiliary variable sets. For the CBS datasets (CSS-CBS, HS-CBS, STS-IND-CBS, STS-RET-CBS) there were little differences in the partial indicator $P_2(Z,k,\rho_{X,Z})$ between the small and extended auxiliary variable sets across categories of variables but for the variable level partial indicators $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$ there were slight reductions in the lack of representativity for some of the variables. This was not necessarily the case for the dataset ESS-BE where the extended variable set increased the $P_2(Z,k,\rho_{X,Z})$ across many of the categories of variables and had mixed results at the variable level partial indicators $P_1(Z,\rho_{X,Z})$, $P_3(Z,\rho_{X,Z})$ and $P_4(Z,\rho_{X,Z})$. The topic of model selection will be explored further in future work.

6. Discussion and Future Work

In this report we defined partial indicators for representative response, described how they may be used in different stages of survey processes, developed their theoretical properties and carried out both a simulation study and used real datasets to assess their impact on identifying variables and categories of variables that contribute to the lack of representativity. When used together with R-indicators and response rates, survey managers can target data collection resources to specific sub-groups contributing to the lack of representativity, identify variables that might be used in survey estimation procedures to reduce non-response bias, assess future strategies for data collection modes and methods for a particular survey and compare different surveys with respect to their representativity.

This paper can be viewed as a first exploration of partial indicators. From this exploration we conclude that the estimated indicators behave as expected with respect to their statistical properties. The analysis furthermore provides valuable insight into the size of confidence intervals for partial indicators and the strength of conclusions that can be drawn given realistic sample sizes. Much is still to be learned, however, about the interpretation of their values and the use in practical settings. Future research on R-indicators and partial indicators will be carried out in the following stages of RISQ (Work Packages 6 and 7), specifically for their use in data collection and assessing approximations to variance estimation for the partial indicators. Two pilots are planned for October-December 2009 where the R-indicators and partial indicators will be used for monitoring response representativeness during the field work. In addition, we will employ more advanced models that distinguish different causes for non-response and include more fieldwork paradata.

Acknowledgements: We thank the members of the RISQ project: Katja Rutar from Statistični Urad Republike Slovenije, Geert Loosveldt and Koen Beullens from Katholieke Universiteit, Leuven, Johan Fosen and Øyvin Kleven from Statistisk Sentralbyrå, Norway, Jelke Bethlehem and Fannie Cobben from Centraal Bureau voor de Statistiek, Netherlands, Ana Marujo from the University of Southampton, UK for their valuable input. We also thank Li-Chun Zhang from Statistisk Sentralbyrå, Norway on his contribution to the variance estimation of partial indicators.

References

Bethlehem, J., Schouten, B. (2008), Representativity Indicators for Survey Quality (RISQ), collaborative project, 7th Framework Programme FP7SSH20071, Socioeconomic sciences and the humanities Part 8, CBS, Voorburg, The Netherlands.

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap. New York: Chapman and Hall.

RISQ (2008), RISQ Data set documentations, Deliverable 1, available at www.r-indicator.eu .

Särndal, C-E and Lundström, S. (2008) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator, *Journal of Official Statistics*, 24, 167-191

Schouten, B., Cobben, F. and Bethlehem, J. (2009), Indicators for the Representativeness of Survey Response. Survey Methodology (to appear).

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N. and Skinner, C. (2009) How to use R-indicators?. Work Package 4, Deliverable 3, RISQ Project, 7th Framework Programme (FP7) of the European Union, available at www.r-indicator.eu

Shlomo, N., Skinner, C.J., Schouten, B., Bethlehem, J., Zhang, L. (2008) Statistical Properties of R-indicators, Work Package 3, Deliverable 2.1, RISQ Project, 7th

Framework Programme (FP7) of the European Union , available at www.r-indicator.eu

Appendix: Variance of Partial Indicators and Variance Estimation

In this appendix we provide analytic approximations to the standard errors of the partial R-indicators P_1 and P_3. The standard error for P_2 can be derived from that of P_1. We do not have an analytic approximation for P_4. We leave this to future research.

Let X be the auxiliary variables, taking values $j = 1,2,...,J$ and Z be a categorical variable for which the partial indicator is calculated with categories $k = 1,2,\ldots,K$.

Analytic approximation to standard error of P_1

$P_1(Z,k,\rho_X) = \sqrt{\Delta_b}$ with the index reflecting that it is a between variance, and

$$\Delta_b = \frac{N_{.k}}{N}(\bar{\rho}_{XZ=k} - \bar{\rho}_X)^2 \qquad \text{(A1)}$$

as shown in (10), and

$$\bar{\rho}_X = \sum_{j=1}^{J} \frac{N_{j.}}{N} \rho_{X=j} \quad \text{and} \quad \bar{\rho}_{XZ=k} = \sum_{j=1}^{J} \frac{N_{jk}}{N_{.k}} \rho_{X=j,Z=k} \text{ ,} \qquad \text{(A2)}$$

N_{jk} is the number of units with $(X,Z) = (j,k)$, $N_{.k}$ is the number of units with $Z = k$, $N_{j.}$ is the number of units with $X = j$. In addition, we assume that $N_{.k}$ is known.

The variable Z may or may not be part of the non-response model. In the case where Z is not part of the model, we have $\rho_{X=j,Z=k} = \rho_{X=j}$ and we can write:

$$\hat{\Delta}_b = \frac{N_{.k}}{N}(\sum_{i \in s_k} d_i \hat{\rho}_i / N_{.k} - \sum_{i \in s} d_i \hat{\rho}_i / N)^2 \qquad \text{(A3)}$$

where d_i is the design weight of unit i in the population U, s is the sample and s_k the sub-sample of s when $Z_i = k$. Let $\delta_i^k = 1$ if $Z_i = k$ and $\delta_i^k = 0$ otherwise. We can rewrite (A3) to:

$$\hat{\Delta}_b = \frac{N_{.k}}{N}(\sum_{i \in s} d_i \hat{\rho}_i \delta_i^k / N_{.k} - \sum_{i \in s} d_i \hat{\rho}_i / N)^2 . \qquad \text{(A4)}$$

As a linear approximation, we have:

$$\text{var}[\hat{P}_1(Z,k,\rho_X)] \approx \frac{1}{4} \frac{1}{E(\hat{\Delta}_b)} \text{var}(\hat{\Delta}_b) . \qquad \text{(A5)}$$

To approximate $\text{var}(\hat{\Delta}_b)$, we will use a second linearization by rewriting (A3) to

$$g_k(u_1,u_2) = \frac{N_{.k}}{N}\left(\frac{u_1}{N_{.k}} - \frac{u_2}{N}\right)^2 \quad \text{where} \qquad \text{(A6)}$$

$$u_1 = \sum_{i \in s} d_i u_{1i} = \sum_{i \in s} d_i \hat{\rho}_i \delta_i^k \text{ and } u_2 = \sum_{i \in s} d_i u_{2i} = \sum_{i \in s} d_i \hat{\rho}_i \,. \tag{A7}$$

The partial derivatives of g are easily derived as:

$$\frac{\partial g_k}{\partial u_1} = \frac{2}{N}\left(\frac{u_1}{N_{.k}} - \frac{u_2}{N}\right) \text{ and } \frac{\partial g_k}{\partial u_2} = -\frac{2N_{.k}}{N^2}\left(\frac{u_1}{N_{.k}} - \frac{u_2}{N}\right) \tag{A8}$$

In addition, we have for the expectations

$$\mu_1 = E(u_1) = N_{.k}\overline{\rho}_{X,Z=k} \text{ and } \mu_2 = E(u_2) = N\overline{\rho}_X \tag{A9}$$

Now we linearize $g_k(u_1,u_2)$ to obtain:

$$g_k(u_1,u_2) = g_k(\mu_1,\mu_2) + \left.\frac{\partial g_k}{\partial u_1}\right|_{u=\mu}(u_1-\mu_1) + \left.\frac{\partial g_k}{\partial u_2}\right|_{u=\mu}(u_2-\mu_2) \tag{A10}$$

Using the linear approximation $\hat{\rho}_i = \rho_i + z_i'(\hat{\beta}-\beta)$ and $z_i = \nabla h(x_i'\beta)x_i$, the derivative of the logistic link function h, it follows that $\text{var}(\hat{\Delta}_b)$ is approximately equal to $\text{var}(\sum_{i \in s} d_i t_i)$, where:

$$\begin{aligned} t_i &= \frac{2}{N}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)u_{1i} - \frac{2N_{.k}}{N^2}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)u_{2i} \\ &= \frac{2}{N}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)(\delta_i^k - \frac{N_{.k}}{N})(\rho_i + z_i'(\hat{\beta}-\beta)) \end{aligned} \tag{A11}$$

From (A11) we find:

$$\text{var}(\sum_{i \in s} d_i t_i) = \tag{A12}$$

$$\text{var}(\sum_{i \in s} d_i \frac{2}{N}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)(\delta_i^k - \frac{N_{.k}}{N})\rho_i + \sum_{i \in s} d_i \frac{2}{N}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)(\delta_i^k - \frac{N_{.k}}{N})z_i'(\hat{\beta}-\beta))$$

The first and second term of (A12) are of order, respectively, $O_p(n^{-1/2})$ and $O_p(n^{-1/2})O_p(n^{-1/2})$. Therefore, we can neglect the second term and approximate the variance by a standard design based variance estimator, where $\sum_{i \in s} d_i \phi_i$, with

$$\phi_i = \frac{2}{N}(\overline{\rho}_{X,Z=k} - \overline{\rho}_X)(\delta_i^k - \frac{N_{.k}}{N})\rho_i \tag{A13}$$

is treated as a linear estimator based on the sample and ϕ_i is a constant associated with unit i. We estimate the variance by replacing ϕ_i with $\hat{\phi}_i$. Finally we divide by $4\hat{\Delta}_b$ to obtain an approximation of $\text{var}[\hat{P}_1(Z,k,\rho_X)]$.

In general $N_{.k}$ may not be known and we may need to estimate it by the sample-based estimator $\hat{N}_{.k} = \sum_{s_k} d_i$. This will introduce a small additional loss of precision.

Analytical approximation to standard error of P_3

We calculate a variance for the partial indicator: $P_3(Z,k,\rho_{X,Z}) = \sqrt{\Delta_w}$, with the index w reflecting that it is a within variance, where

$$\Delta_w = \sum_{j=1}^{J} \frac{N_{jk}}{N} (\rho_{X=j,Z=k} - \bar{\rho}_{X=j})^2 \quad \text{and} \tag{A14}$$

$$\rho_{X=j} = \sum_{k=1}^{K} \frac{N_{jk}}{N} \rho_{X=j,Z=k} \tag{A15}$$

Consider the estimator:

$$\hat{\Delta}_w = \frac{1}{N} \sum_{j=1}^{J} \sum_{i \in s_j} d_i (\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})^2 = \frac{1}{N} \sum_{i \in s} d_i (\hat{\rho}_i - \sum_{j=1}^{J} \delta_i^j \hat{\bar{\rho}}_{X=j})^2 \tag{A16}$$

where s_j is the sub-sample of s for which $X_i = j$, and $\delta_i^j = 1$ if $Z_i = j$ and $\delta_i^j = 0$ otherwise.

Again, we employ a linear approximation of the variance of the estimator $\hat{P}_3$:

$$\text{var}[\hat{P}_3(Z,k,\rho_{X,Z})] \approx \frac{1}{4} \frac{1}{E(\hat{\Delta}_w)} \text{var}(\hat{\Delta}_w). \tag{A17}$$

We define $\phi_i = (\rho_i - \sum_{j=1}^{J} \delta_i^j \bar{\rho}_{X=j})^2$ and write $\hat{\phi}_i = \phi_i + \left(\frac{\partial \hat{\phi}_i}{\partial \hat{\beta}} \right) \Bigg|_{\hat{\beta}=\beta} (\hat{\beta} - \beta)$.

Using the same argument as for (A12), we approximate the variance by a standard design based variance estimator where $\sum_{i \in s} d_i \phi_i$ is treated as a linear estimator based on the sample and ϕ_i is a constant associated with unit i . We estimate the variance by replacing ϕ_i with $\hat{\phi}_i$. Finally we divide by $4\hat{\Delta}_w$ to obtain an approximation of $\text{var}[\hat{P}_3(Z,k,\rho_{X,Z})]$.