Representativeness indicators for measuring and enhancing the composition of survey response

Barry Schouten (Statistics Netherlands)
Natalie Shlomo and Chris Skinner (University of Southampton)

www.R-indicator.eu

Why indicators for representative response?

- Response rate is insufficient indicator of quality response
  - Response rate limits maximal impact of nonresponse
  - Literature gives examples where increased response rate corresponded to increased nonresponse bias (e.g. Peytcheva & Groves 2009 for recent examples)

- There is a need for indicators that enable
  1. comparison of response quality in different surveys
  2. comparison of response quality over time in one survey
  3. monitoring of response quality during data collection
  4. optimization of data collection resources

What is representative response?

- Limitations
  - *Dependence on external information*: No statement about the representativeness of response is possible without information that is auxiliary to the survey.
  - *Dependence on sample size:* The strength of any statement about the nature of response to a survey will depend on the sample size.
  - *Non-response adjustment*: Indicators are not designed for the selection of weighting variables but for the evaluation and enhancement of response.
  - *Less is better?*: One may always attain representative response with respect to some standard by simply erasing (or sub-sampling) overrepresented groups in the response.

What is representative response?

- Definition: *Response is representative with respect to X if the response propensities are constant for X.*
- Ideal situation: Availability of a "super" vector ***x*** that fully explains response behaviour
- R-indicator: *the variation in response propensities*

$$R(\rho_X) = 1 - 2S(\rho_X)$$

- Unconditional partial R-indicator for a single variable Z: *the between variance of response propensities*

$$P_u(Z) = S(\rho_Z) \qquad P_u(Z,k) = \sqrt{\frac{N_k}{N}}(\rho_{Z_k} - \rho)$$

What is representative response?

- Definition: *Response is conditionally representative with respect to Z given X when the conditional response propensities are constant for Z.*

- Conditional partial R-indicator for a single variable Z: *the within variation in response propensities given stratification on X*

$$P_c(Z \mid X) = \sqrt{\frac{1}{N-1}\sum_U (\rho_{X,Z}(x_i, z_i) - \rho_X(x_i))^2}$$

$$P_c(Z, k \mid X) = \sqrt{\frac{1}{N-1}\sum_U Z_k (\rho_{X,Z}(x_i, z_i) - \rho_X(x_i))^2}$$

What is representative response?

- Maximal absolute nonresponse bias

$$\frac{|B(\hat{\bar{y}}_r)|}{S(y)} = \frac{|Cov(y,\rho_Y)|}{\rho S(y)} \leq \frac{S(\rho_Y)}{\rho} \leq \frac{S(\rho_{\mathbb{N}})}{\rho} = \frac{1-R(\rho_{\mathbb{N}})}{2\rho}$$

- We have to resort to available *X*

$$B_m(X) = \frac{1-R(\rho_X)}{2\rho}$$

- Difference in nonresponse bias when adding *Z*

$$\Delta B_m(Z,k \mid X) = B_m(X,Z_k) - B_m(X) = \frac{R(\rho_X) - R(\rho_{X,Z_k})}{2\rho}$$

How to estimate indicators?

- Indicators based on estimated variation of estimated response propensities

- Situation 1: Sample is linked to *X*
  - Replace population means by design-weighted sample means
  - Estimate response propensities using regression on sample

- Situation 2: Population contingency tables for *X*
  - Replace population means by propensity-weighted response means
  - Estimate sample covariance matrix by population covariance matrix of *X*

- Situation 2: Population marginal counts for *X*
  - Replace population means by propensity-weighted response means
  - Estimate sample covariance matrix by combination of response covariance matrix and marginal counts of *X*

How to estimate indicators?

- Consequence: Indicators have precision and may be biased
- Situation 1: Bias results from plugging in estimated response propensities, but the bias can be adjusted for.
- Situations 2 and 3: Additional bias results from picking up variation in (unknown) sample composition of *X*.

How to use indicators?

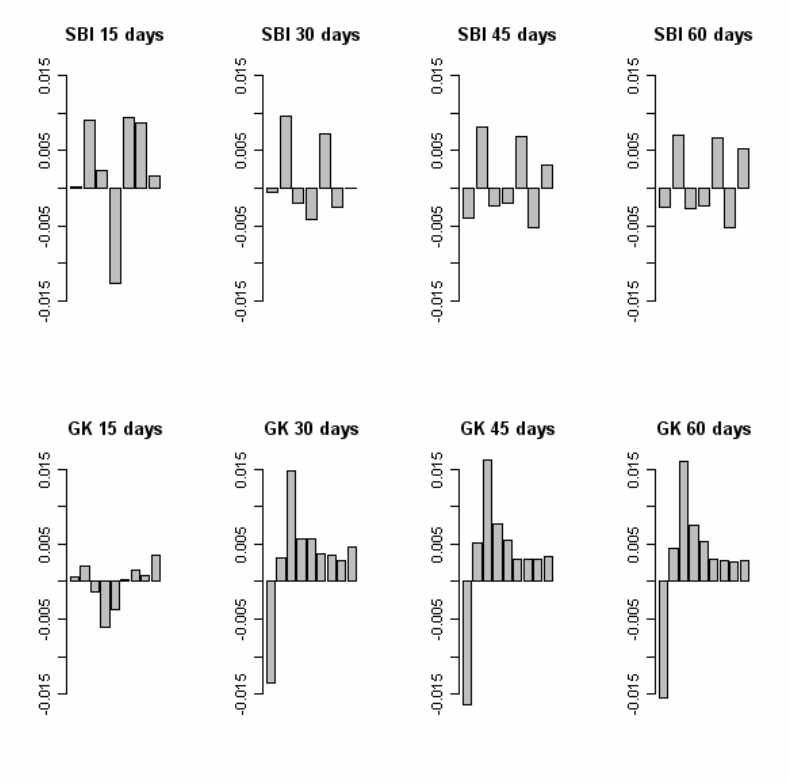
- Always present indicators and maximal bias together with *X*
- Provide confidence intervals
- Use the same *X* when comparing two surveys or monitoring a single survey in time
- Choose *X* based on intended use
    - Comparison of surveys: General variables, simple models
    - Comparison of one survey: Add variables related to survey items, models may include paradata
    - Data collection: Distinguish different causes of nonresponse
- Use partial indicators to identify groups that need additional effort (responsive or adaptive designs)

Example 1: Dutch Short Term Statistics 2006

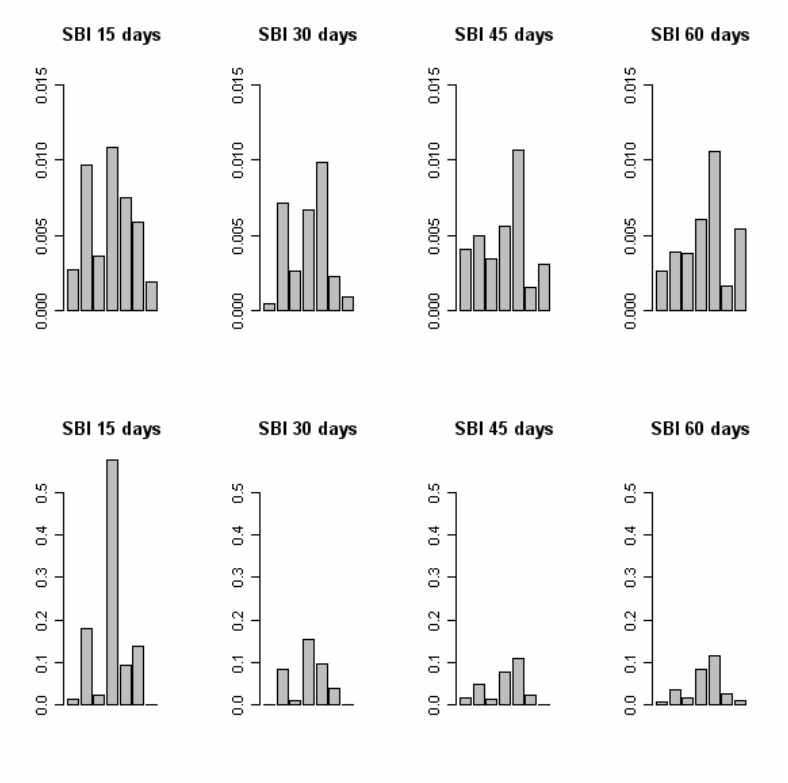
Survey		*X = Business size + type*				*X = Business size x VAT + type*			
		15days	*30days*	*45days*	*60days*	*15days*	*30days*	*45days*	*60days*
Industry	R	92,1%	93,3%	94,0%	94,2%	90,5%	91,8%	93,1%	93,3%
	CI	91,3-92,8	92,7-94,0	93,5-94,4	93,8-94,6	89,7-91,3	91,3-92,2	92,6-93,5	92,8-93,8
	B	8,1%	4,2%	3,5%	3,3%	9,7%	5,2%	4,1%	3,8%
Retail	R	96,1%	94,6%	94,0%	94,1%	88,1%	87,9%	88,3%	89,0%
	CI	95,4-96,7	94,0-95,2	93,5-94,5	93,6-94,6	87,3-88,8	87,3-88,6	87,6-88,9	88,3-89,6
	B	3,9%	3,5%	3,5%	3,3%	12,0%	7,7%	6,8%	6,2%

Example 2: Dutch Short Term Statistics retail 2006

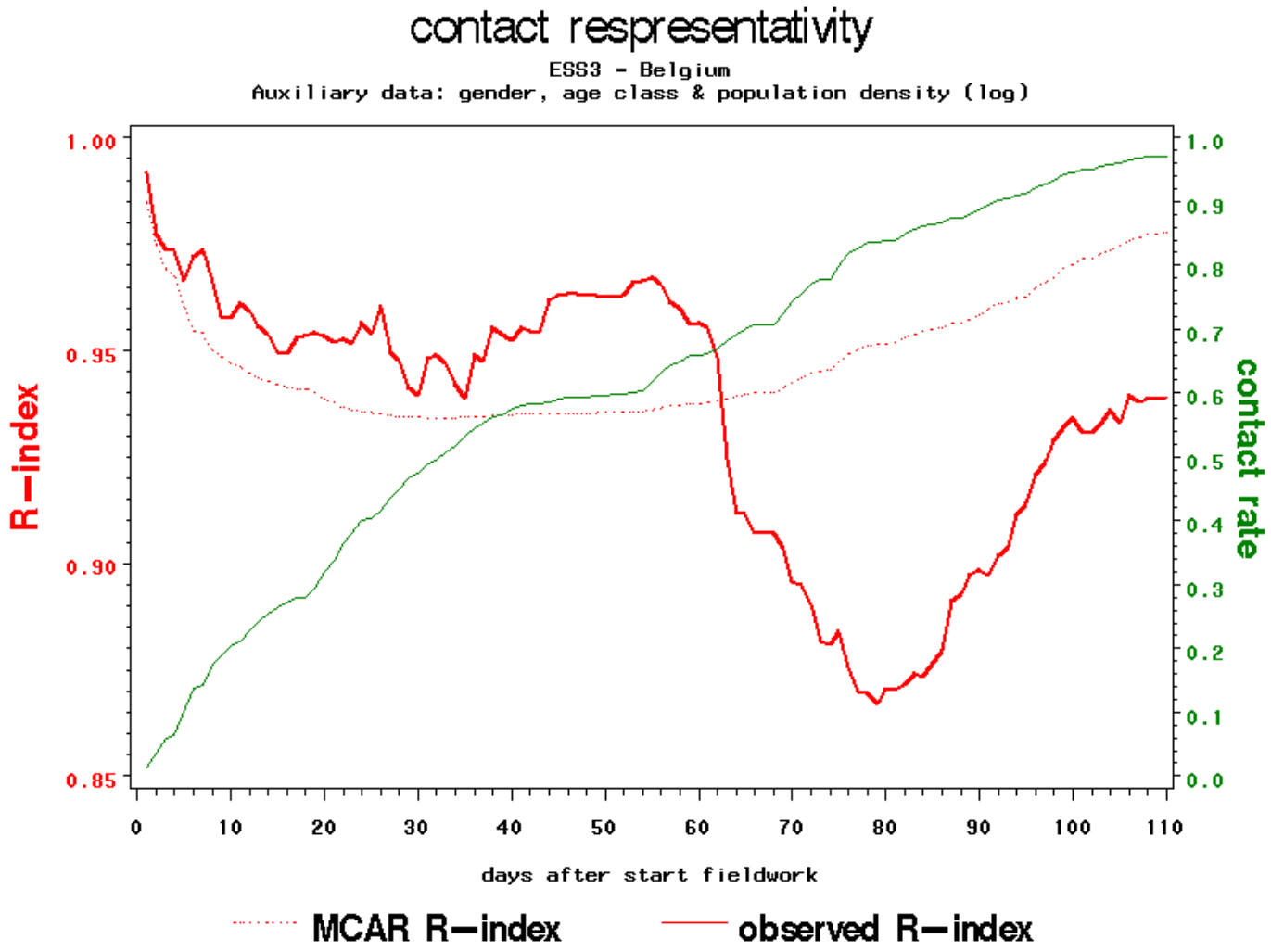
Unconditional partial indicators for business type (SBI) and business size (GK)



Conditional partial indicators and difference in maximal bias for business type (SBI)

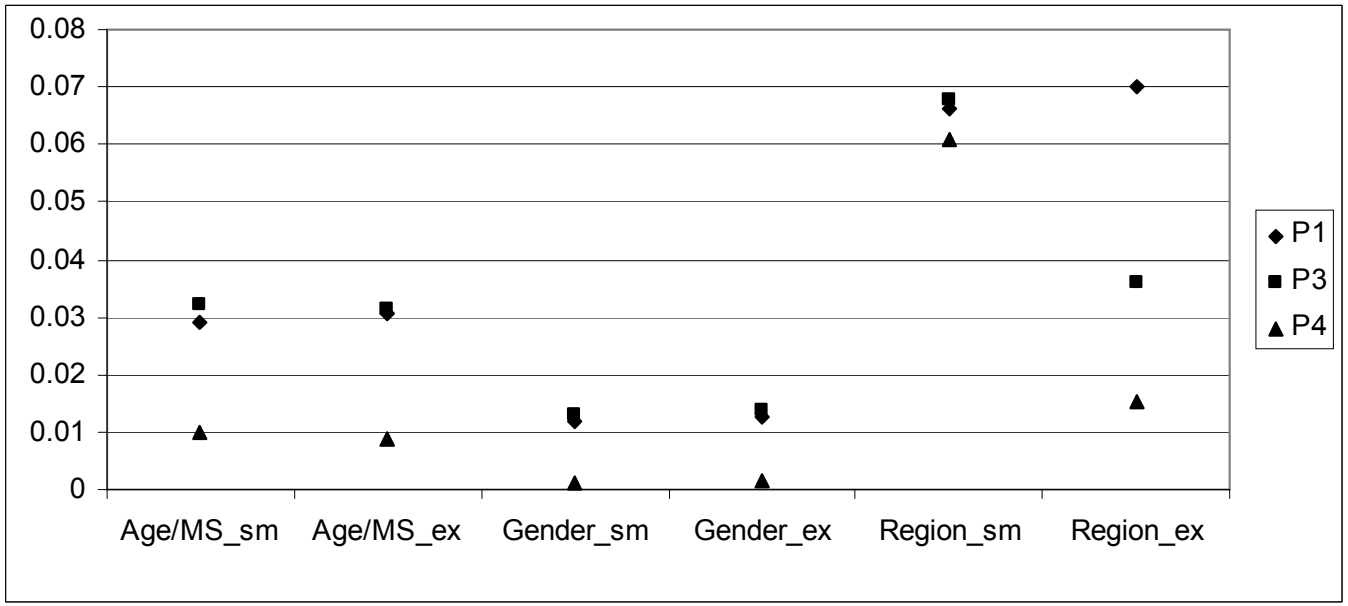


Example 3: Making contact in ESS Belgium



Example 4: Response to ESS Belgium

Unconditional partials (P1), conditional partial indicators (P3) and difference in maximal bias (P4) for age x marital status, age and region given small and full sets of auxiliary variables.



Example 5: Health Survey + Consumer Satisfaction Survey 2005

Forward variable selection
A = gender, B = age x marital status, C = urbanization, D = house value, E = paid job,
F =household type and G = ethnic background

Health Survey 2005	*Consumer Satisfaction Survey 2005*
Full model = A+B+C+D+E+F+G R=80,8% CI=(79,4 – 82,3)	Full model = A+B+C+D+E+F+G 82,1% (80,7 – 83,4)
B R=85,5% CI=(84,0 – 87,0)	B R=84,6% CI=(83,2 – 86,0)
B+G R=82,9% CI=(81,4 – 84,2)	B+F R=83,2% CI=(81,8 – 84,6)
B+G+C R=81,7% CI=(80,3 – 83,2)	B+F+G R=82,8% CI=(81,4 – 84,2)
B+G+C+F R=81,2% CI=(79,7 – 82,8)	B+F+G+E R=82,5% CI=(81,2 – 84,0)
Final selection = B+G+C+F+E R=81,0% CI=(79,6 – 82,4)	Final selection = B+F+G+E+A R=82,4% CI=(81,0 – 83,8)

Future work

- Future research
  - Bias adjustment of population-based indicators
  - Explore use of indicators in data collection monitoring
  - Two pilots with optimization of indicators given fixed costs
  - Optimal survey designs

- Potential extensions
  - Other causes for missing data: item-nonresponse, panel attrition, linkage, under coverage
  - Registers: completion representativeness as function of *t*