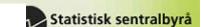# Representativity Indicators for Survey Quality RISQ

- Research objectives
- Work packages
- Planning of work packages
- Further elaboration of work packages

RISQ meeting, 7 – 8 April 2008, Voorburg

Research objectives

Development, testing and application of statistical indicators that measure the extent to which a survey or register is representative of the population under investigation.

Indicators to be used in all steps of data collection process:
• Analysis (WP 3 and 4)
• Processing (WP 5)
• Monitoring (WP 6)
• Control (WP 7)

Research objectives

Two types of indicators:
1. Indicators that enable comparison of surveys (R-indicators)
   - in time
   - in topic
2. Indicators that enable monitoring and controlling of data collection (partial R-indicators)

Work package 2

- Identification of data sets that will be used throughout the research.
- Linkage of data sets to fieldwork paradata
- Linkage of data sets to auxiliary variables from register data
- Documentation of data sets

Documented data sets
- Testing and validation (WP 3 – 6)
- Pilots (WP 7)

Work package 3

- Development of new Representativity Indicators
- Elaboration of existing Representativity Indicators
- Derivation of theoretical properties of Representativity Indicators
- Empirical validation of Representativity Indicators
- Selection of Representativity Indicators for use in RISQ

Deliverables
- Report Representativity indicators including the state-of-the-art
- Macro's and syntax in standard statistical software

Work package 4

- Development of theory and best practices that allow for the comparison of Representativity Indicators for different surveys and registers

Interpretation complicated by
- Dependency on sample size
- Dependency on available auxiliary information

Strategies
1. Employ a fixed pre-defined set of auxiliary variables
2. Use all "significant" available auxiliary variables
3. Incorporate population parameters

Deliverables
Report about how to use indicators in comparisons of surveys and registers

Work package 5

- Development of new partial Representativity Indicators
- Elaboration of existing partial Representativity Indicators
- Derivation of theoretical properties of partial Representativity Indicators
- Empirical validation of partial Representativity Indicators
- Selection of partial Representativity Indicators for use in RISQ

Deliverables
- Report partial R-indicators
- Macro's and syntax in standard statistical software

Work package 6

- Selection of process data
- Incorporation of process data in measurement of indicators

Deliverables
- Report about indicators that incorporate process data
- Report about use of indicators in monitoring data collection

Work package 7

- Investigation of the use of Representativity Indicators in differentiated data collection strategies
- Investigation of the use of Representativity Indicators in responsive data collection designs

Deliverables
- Work plan pilot studies
- A face-to-face and telephone pilot study
- Report Representativity indicators for the controlling of data collection

Work package 7

Details	CBS	SSB
Data collection mode	Telephone	Mixed-mode; telephone + face-to-face
Survey	Existing telephone survey; to be decided during June – July 2009	Level of living
Sampling design	Identical to selected telephone survey	Self-weighting, two-stage design
Sample size	2.000	1.500
Test period	September - October 2009	December 2009 – January 2010
Fieldwork period	November - December 2009	February – March 2010

Work package 7

Responsive designs - Groves & Heeringa (2006):
The ability to monitor continually the streams of process data and survey data creates the opportunity to alter the design during the course of data collection to improve survey cost efficiency and to achieve more precise, less biased estimates.

Differentiated data collection strategies:
• Different strategies for households/businesses with different background
• Static: Pre-defined
• Dynamic: Dependent on process data
• Goal: optimal balance between quality and costs

Work package 7

Ingredients:
1. Auxiliary variables
2. Strategies
3. Criteria to optimise quality versus costs

Work package 7

Strategies: $S = \{s_0, s_1, \ldots, s_K\}$

Auxiliary: $x \in X$

Distribution: $q(x)$

Costs: $c(x, s)$

Response: $\rho(x, s)$

Allocation: $p(s \mid x)$

Work package 7

Strategy probabilities $\qquad \displaystyle\sum_{s \in S} p(s \mid x) = 1$

Expected sample size $\qquad \displaystyle n = N \sum_{x} q(x)(1 - p(s_0 \mid x))$

Expected costs $\qquad \displaystyle C = N \sum_{x,s} q(x) p(s \mid x) c(x,s).$

Expected response rate $\qquad \displaystyle \overline{\rho} = \frac{N}{n} \sum_{x,s} q(x) p(s \mid x) \rho(s,x)$

Work package 7

Optimization:

$$\min_{p(s|x)} K(p(s\,|\,x); \rho(x,s), q(x))$$

given $\qquad C \le C_{\max}$

$$\min_{p(s|x)} C(p(s\,|\,x); \rho(x,s), q(x))$$

given $\qquad K \ge K_{\min}$

Candidate: MSE (mean square error).

Work package 7

Example: Minimise impact of worst-case-scenario on MSE

One can show that:

$$| B(\hat{\bar{y}}_{HT}) | = \left| \frac{\text{cov}(\tilde{\rho}, y)}{\bar{\rho}} \right| \leq \frac{S(\tilde{\rho})S(y)}{\bar{\rho}} \leq \frac{S(y)(1 - R(\tilde{\rho}))}{2\bar{\rho}}$$

$$\hat{R}(\hat{\tilde{\rho}}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\bar{\rho}}_{HT})^2} \qquad \text{(R-indicator)}$$

$$K = \frac{\hat{S}^2(\hat{\tilde{\rho}})\hat{S}^2(y)}{\hat{\bar{\rho}}^2} + \text{var}(\hat{\bar{y}}_{HT})$$

Planning of work packages

WP	Work package title	Lead participant number	Person months	Start month	End month	Indicator
WP 1	Project management	CBS	6,0	March 2008	June 2010	
WP 2	Data preparation and documentation	CBS	6,3	March 2008	June 2008	1 + 2
WP 3	Theory and validation	SOTON	11,2	March 2008	November 2008	1
WP 4	Indicators and comparison of surveys	CBS	5,0	August 2008	January 2009	1
WP 5	Indicators and data processing	SOTON	10,7	August 2008	April 2009	2
WP 6	Indicators and data collection monitoring	KUL	14,0	December 2008	October 2009	2
WP 7	Indicators and data collection control	SSB	17,9	June 2009	June 2010	2
WP 8	Exploitation	CBS	6,0	March 2008	June 2010	
	TOTAL		77,1			

Elaboration of work packages 3-7

Work package 3:
- What is a useful definition of representativity?
- What is an R-indicator?
- How to develop new indicators?
- Literature?
- Can we develop indicators that deploy population totals only?
- How to select and choose indicators?

Work package 4:
- How can we validate indicators?
- How do we deal with dependency on sample size?
- How do we deal with dependency on auxiliary information?

Elaboration of work packages 3-7

Work package 5:
- How to isolate the influence of single variables in indicators?
- How to interpret and use partial R-indicators?
- How to validate partial R-indicators?
- Can we base partial R-indicators on population totals?

Work package 6:
- How to make a selection of useful process data?
- How to incorporate the process data in (partial) R-indicators?
- How to cope with different interview modes?

Elaboration of work packages 3-7

Work package 7:
- Do we want to differentiate data collection beforehand?
- Do we want to use responsive designs?
- Do we need a theoretical framework for strategy differentiation?
- What strategies may be useful for the pilots?
- Can we use indicators as criteria for strategy allocation?
- Can we rely on models that predict (types of) nonresponse?
- How to design and use the pilots?
- How to deal with different interview modes?

Elaboration of work packages 3-7

Proposal:
• WP coordinators prepare detailed research plan + planning
• Presentation and tuning of plans at meeting in June

Work packages 1 and 8

- Extended progress reports every 6 months
- One page progress reports in between
- Cost statements every 6 months
- Video meetings in between plenary meetings
- Technical reviews (indicators, partial indicators, pilots)
- Coordination of overview papers/articles

- Conferences
    - Q2008, July 2008 (FP 7 session + special topic session)
    - Symposium Stat Canada, October 2008
    - ISI, September 2009
    - Other?