Workpackage 3

Accuracy Measurement of Composite Indicators

Deliverable 3.4

2008

List of contributors:

Ralf Münnich, University of Trier; Kersten Magg, Dominik Ohly, and Rolf Wiegert, University of Tübingen

Main responsibility:

Ralf Münnich, University of Trier; Kersten Magg, Dominik Ohly, and Rolf Wiegert, University of Tübingen

CIS8–CT–2004–502529 KEI

The project is supported by European Commission by funding from the Sixth Framework Programme for Research.

`http://europa.eu.int/comm/research/index_en.cfm`

`http://europa.eu.int/comm/research/fp6/ssp/kei_en.htm`

`http://www.cordis.lu/citizens/kick_off3.htm`

`http://kei.publicstatistics.net/`

Preface

Within workpackage 3 of the KEI project, quality aspects and statistical properties of indicators were presented and discussed. After introducing general concepts of data quality (deliverable 3.1) special emphasis was laid on quality of indicators (deliverable 3.3) as well as on the impact and handling of missing data (deliverable 3.2).

Workpackages 5 and 7 were dedicated to investigate the methodology of composite indicators including their sensitivity analysis. The main study on the KEI dataset (cf. workpackage 2) was performed on macro data, i.e. country specific aggregated data. Since many of these data come from sample surveys, it seems adequate to consider the surveying process as one important aspect of data and thus indicator quality. This was considered in workpackages 3.1 and 3.3.

The aim of this deliverable is to present techniques which help to improve the quality of indicator estimates and allow to measure the accuracy of the estimates. As one major problem with regards to accuracy measurement, the role of National Statistical Institutes for delivering the appropriate information will be discussed in connection with transferring the concepts from single to composite indicators. Herewith, the question arises whether first macro information suffices to measure the single indicators' accuracy, and second, if quality information on single indicators provide the necessary information for measuring the accuracy of composite indicators.

The authors would like to thank the KEI partners for valuable discussions which helped to improve the study.

Contents

List of Figures

List of Tables

Table of Abbreviations and Symbols

The following table lists the most important abbreviations and symbols appeared in this report. In general, statistical parameter are kept in Greek letters. Estimates are characterised with a $\widehat{\ }$ above the parameter of interest. Vectors and matrices are printed in bold text.

ALLBUS	German general social survey
AV	asymptotic variance
B	bias
BStatG	Gesetz über die Statistik für Bundeszwecke
CAL	calibration estimator
CI	confidence interval
CIoILC	Composite Index of Individual Living Conditions
CV	coefficient of variance
DSP	Dauerstichprobe befragungsbereiter Haushalte
E	expectation operator
EBLUP	empirical best linear unbiased predictor
ECHP	European Household Panel
ESS	The European Social Survey
EU	European Union
EUROSTAT	Statistical Office of the European Communities
EUSI	European System of Social Indicators
EU-SILC	The European Community Statistics on Income and Living Conditions
EVS	German sample survey of income and expenditure
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.

GMC	German Microcensus
GREG	generalised regression estimator
HH	household
HHTYP	type of household
HT	Horvitz-Thompson estimator
ICT	The European Community Statistics on information and communications technologies
ID	identification number
INC	income
ISCED	International Standard Classification of Education
JRC	Joint Research Center of the European Commission in Ispra (KEI partner)
KBE	Knowledge Based Economy
KUL	Katholieke Universiteit Leuven (KEI partner)
LWR	Laufende Wirtschaftsrechnungen
min!	minimization, e.g. in context of mathematical programming
MSE	mean squared error
NA	missing value
NSI	National Statistical Institute
NSM	national sample mean
NUTS	Nomenclature des Unités Territoriales Statistiques
OLS	ordinary least squares
Pers	person
RMSE	root mean squared error
RRMSE	relative root mean squared error
SAE	small area estimation
SI	single indicator
SISoL	sub-indicator `standard of living`
SRS	simple random sampling

STSRS	stratified simple random sampling
SUF	scientific use file
SYNTH	synthetic estimator
SZH	social affiliation of head of household
V	variance operator
WLS	weighted least squares
ZUMA	The Centre for Survey Research and Methodology

Symbols

$\forall$	for all
$d_k = 1/\pi_k$	inverse inclusion probability
$\mathbf{d}$	vector of design weights
$\mathbf{D}$	diagonal matrix of design weights
e_k	residual corresponding to sample element k
$\mathbf{e}$	vector of residuals
$F(\cdot)$	calibration function
$G(\cdot)$	distance function
g_{di}	g-weight in area d for element i
g_k	g-weight for element k
$\mathbf{g}$	vector of g-weights
$\mathbf{G}$	matrix of g-weights
L	lower boundary for g-weights
M	number of households in population
N	number of elements in population
N_d	number of elements in area d
$\widehat{N}$	estimated number of elements in population
m	number of selected households in sample
n	number of selected elements in sample

n_d	number of selected elements in sample in area d
n_i	number of elements selected from household unit $\mathcal{U}_\iota$
p	number of auxiliary variables
R	number of runs in Monte-Carlo simulation study
s_d	sample in area d
$\mathcal{S}$	sample of elements
$\mathcal{S}_M$	sample of households
$\mathcal{S}_I$	sample of individuals
$\mathcal{S}_K$	sub-sample of Kish individuals
$\mathcal{S}_\iota$	n_ι elements selected from $\mathcal{U}_\iota$
u_d	area-specific error
U	upper boundary for g-weights
$\mathcal{U}$	population of elements
$\mathcal{U}_M$	population of households
$\mathcal{U}_I$	population of individuals
$\mathcal{U}_K$	sub-population of Kish individuals
$\mathcal{U}_\iota$	household unit sized N_ι persons
$\mathbf{v}_{i,k'}$	vector of auxiliary information known for Kish individual k'
w_k	calibrated weights for element k
$\mathbf{w}_i$	vector of calibrated weights
$\mathbf{W}$	diagonal matrix of calibrated weights
$\boldsymbol{x}_i$	vector of auxiliary information known for household i
$\mathbf{X}$	matrix of auxiliary information
y_k	observation for element k
$\mathbf{y}$	vector of observations
z_{di}	domain indicator for area d and element i
$z_{1-\alpha}$	$(1-\alpha)$ quantile of normal distribution
$\mathbf{z}_i$	vector of auxiliary information known for individual k
$\boldsymbol{\beta}$	vector of regression coefficients

$\widehat{\boldsymbol{\beta}}$	estimated vector of regression coefficients
ε_i	independent errors for element i
ε_{di}	independent errors in area d for element i
$\boldsymbol{\lambda}$	vector of Lagrange multipliers
μ	mean value
$\widehat{\mu}_Y$	estimated mean value of variable Y
$\boldsymbol{\mu}$	vector of Lagrange multipliers
γ_d	optimal weight for area d
$\boldsymbol{\gamma}$	vector of Lagrange multipliers
π_i	inclusion probability of element i
π_{ij}	second order inclusion probability
τ	total value
$\widehat{\tau}$	estimated total value
$\widehat{\tau}_{\text{HT}}$	Horvitz-Thompson estimate
$\widehat{\tau}_{\text{CAL}}$	Calibration estimate
$\widehat{\tau}_{Y_d}$	estimated total value of variable Y in area d
$\boldsymbol{\tau}_x$	vector of known totals for $\mathcal{U}_M$
$\boldsymbol{\tau}_{X_d}$	vector of known totals in area d
$\boldsymbol{\tau}_z$	vector of known totals for $\mathcal{U}_I$
$\boldsymbol{\tau}_v$	vector of known totals for $\mathcal{U}_K$
θ	parameter to be estimated
ω_g	set of g-weights
ω_i	inverse of the inclusion probability of individual i

Chapter 1

Introduction

Measuring the development of the impact of policy making on the economy is mainly dedicated to measuring changes of indicator values. The European Commission urges the needs of using adequate concepts to measuring the fulfilment of the goals of the Lisbon agenda. In many areas, special concepts for sets of indicators are used, e.g. the Laeken indicators for measuring poverty and social exclusion. Additionally, in several areas composite indicators are used where the *key figures in science, technology and innovation* are a good example (cf. `http://ec.europa.eu/invest-in-research/monitoring/statistical01_en.htm`).

The given single indicators, also as parts of composite indicators, in general use aggregated or so-called macro data which are published, e.g. on the Eurostat New Cronos database. In practice, mostly the data are used as they are as given *precise* numbers. Rarely the origin of the data as survey data is considered with a measurable inference. Quality reports certainly should point out the accuracy of the given values of interest (cf. deliverable 3.3). In case, these *variance estimates* are not available, one would have to apply variance estimation methods (for a thorough overview, the interested reader is referred to the reports from the DACSEIS project: `http://www.dacseis.de`) on the micro dataset. These computations, in general, still have to be performed by the data producer which in many cases is Official Statistics.

The aim of this deliverable is to present the methodology of improving estimates of indicator values and to measure its accuracy, i.e. the application of standard variance estimation methods. The general methodology refers to Horvitz-Thompson estimation and its improvements via calibration. Additionally, one example will be elaborated that allows to disaggregate the estimates into smaller subgroups, the so-called small area estimation methodology.

The calibration approach and the small-area estimation methods aim to improve the Horvitz-Thompson estimates and hence are of greater precision in estimating the characteristics of interest. Calibration methodology, as developed by Deville and Särndal (1992) and Deville et al. (1993), has the desirable additional feature of correcting for non-sampling errors, especially non-response and coverage errors. Small area estimation is particularly good (see on this point inter al. Rao, 2003) at tabling official reference numbers for both small areas (regional subgroups) and small domains (disaggregation by context variables), where classical direct estimators are likely to perform poorly. Both

groups of methods use auxiliary information known for the population of interest (persons, households, companies). This auxiliary information may consist of e.g. socio-demographic characteristics.

Against the background of the KEI project, in respect of the database and construction of composite indicators two prerequisites need to be met:

- Indicators are calculated on the basis of microdata. In statistical terms microdata chiefly consist of random samples taken from a suitably demarcated basic population. The statistical evaluation of indicators requires that this sampling process be adequately represented. What is needed, is a database comprising a sufficient number of observations and variables, so that i) samples can be taken in step with actual practice, and ii) the potential (in terms of estimation theory) of the afore-mentioned estimation methods can be shown.

- What is also needed, is a composite indicator, recognised by social scientist and economists alike, to supply information for e.g. the KBE. For each single indicator that forms part of a complexly composed composite indicator, associated microdata (individual data) will need to be accessed for cross-sectional and longitudinal analysis.

For a multiplicity of relevant composite indicators, either we find that we have too few individual data, classified into longitudinal and cross-sectional or we find that we have none at all (possibly due to data-protection regulations). This holds especially for indicators used in the KBE. Also, the use of indicators can involve considerable overhead costs, if and when data have to be obtained from companies or from non-governmental institutes. Thus, there is a permanent trade-off between data availability and composite indicators - the latter often only scientifically defined or else conforming to political agendas.

In light of the restrictions outlined above, the studies to be presented below will target the composite index of individual living conditions of the `Centre for Survey Research and Methodology` (ZUMA)[1]. This composite index is listed inter al. on the information server on composite indicators[2], operated by the `Joint Research Centre` (JRC) of the European Commission. It has both direct and indirect relevance for KBE.

Meeting in Lisbon in 2002, the European Council set itself the ambitious goal of targetting poverty and social exclusion. Successes were subject to annual monitoring, based on suitable index figures. The ZUMA index sheds light on the living standard of households and the quality of life of persons. Into the index entered numerous characteristics collected by the `Community Statistics on Income and Living Conditions` (EU-SILC), meaning that these statistical variables are available for comparison in the EU member states.

[1]The Centre for Survey Research and Methodology became part of the GESIS organization (`Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.`). GESIS is an institution devoted to research and service. It provides information, consultation, and data, supports and facilitates scientific work. For more information on GESIS cf. `http://www.gesis.org/`. In the meanwhile both are merged as GESIS – Leibniz-Institut für Sozialwissenschaften.

[2]Cf. `http://composite-indicators.jrc.ec.europa.eu/`

In Germany, EU-SILC samples are taken from the permanent sample, i.e. a fixed pool of households willing to be surveyed (`Dauerstichprobe befragungsbereiter Haushalte [DSP]`), which, at the time of writing, is still further developed. All households exiting the `German Microcensus` (GMC) are asked if they wish to participate in any further (voluntary) collections of data for official statistics. All households participating in the EU-SILC will, as from 2009, have previously participated in the GMC.

This database permits a semi-synthetic basic population to be built within the scope of the GMC. The EU-SILC characteristics of interest for the composite indicator (i.e. the frequencies of characteristics) are estimated with the help of Logit models. In the event that EU-SILC is a representative sub-sample of the GMC, the results of Logit modelling can be applied to the GMC to generate artificial universes. The structure given by EU-SILC variables of interest can be represented in GMC. The compiled population, containing original GMC data as well as generated EU-SILC characteristics (semi-synthetic), enables complete EU-SILC samples to be taken (14,100 households).

Using Monte-Carlo simulation, the alternative estimation methods are evaluated for some 10,000 independently taken EU-SILC samples. This allows the influence of both the sample design and the auxiliary variable set (used for calibration purposes) to be studied. The composite indicator is estimated for three different sample designs, two different sample sizes, and six different sets of variables. The results are graphically displayed. In particular, the study addresses difficulties arising in connection with estimating the variance of composite indices.

The deliverable is structured as follows. The composite index of individual living conditions is presented in Section 2.1. Section 2.2 describes the data sources collected, on which the subsequent analyses are based. The manner in which the datasets were processed is set out in Section 2.3. Owing to the non-availability of several variables, which - according to ZUMA's original definition - need to be entered into the composite indicator, the latter is redefined in Subsection 2.3.4. The frequency distributions of the single indicators and the composite indicator are shown graphically.

The calibration technique for improving the projection of the composite indicator is explained in Section 3.2. The course and goal of the simulation study are treated in Section 3.3. In it details are given of the auxiliary variable set used, the sample design followed, the units used for accuracy measurement and interpretation of the simulation results, as well as some explanations on how to read the evaluation graphics.

The results of the composite indicator are given in Section 3.4. In an annex (3.5) some of the results, namely those for the single indicators, are evaluated. Estimating the variance of the composite indicator is the task of Section 3.6. In Section 3.7, the authors draw a balance of their results.

Chapter 4 is devoted to the small area estimation (SAE) on single and composite indicators. After an introduction and motivation of the methodology in Section 4.1, the following Section 4.2 will describe the set of used and evaluated direct and indirect estimators. Section 4.3 explains the small area specific subjects of the simulation study, and Section 4.4 shows a selection of representative results from the broad simulation study.

A summary of the results is given in the final Chapter 5.

Appendix A contains a complete list of variables, a tabular overview of the auxiliary variable set, and a frequency distribution of the regressors. Appendix B gives detailed treatment of the correlative relationships between the variables and their frequency distribution in the individual datasets. This includes a comparison of the latter. Appendix C sets out the sampling plan for the 581 strata.

Chapter 2

The Composite Index of Individual Living Conditions

Since no data were forthcoming by other means, using `the composite index of individual living conditions`[1] (CIoILC) took shape as a feasible way to conduct the desired analyses and generate results.

2.1 Content-related basics

The composite indicator addresses one of the six major dimensions of social development defined in the `European System of Social Indicators` (EUSI)[2], especially for the KBE. In the initial phase, the development of the EUSI was funded by the European Commission as a sub-project of the EU-Reporting-Project within the 4^{th} Framework TSER-Programme. It aims to develop a theoretically and methodologically well-grounded set of measurement dimensions and indicators for monitoring the quality of live and societies across Europe. The measurement dimensions considered reflect the following concepts and basic dimensions:

- the concept of quality of live
  - objective living conditions
  - subjective well-being
- the concept of social cohesion
  - disparities, inequalities and social exclusion
  - social relations, ties and inclusion
- the concept of sustainability
  - preservation of human capital
  - preservation of natural capital.

[1] Cf. `http://www.gesis.org/en/services/data/social-indicators/eusi/total-life/`
[2] Cf. `http://www.gesis.org/en/services/data/social-indicators/eusi/`

The CIoILC presents as a summary measure of the quality of objective living conditions and consists of the following seven indices:

No.	index	score points	transformation of score points
1	`income/ standard of living`	[4 ; 20]	[1 ; 5]
2	`housing`	[0 ; 5]	[1 ; 5]
3	`housing area`	[0 ; 3]	[1 ; 5]
4	`education`	[1 ; 5]	[1 ; 5]
5	`health`	[0 ; 6]	[1 ; 5]
6	`social relations`	[0 ; 4]	[1 ; 5]
7	`work`	[2 ; 4]	[2 ; 4]
		total score:	[8 ; 34]

Indices `education` and `work` are single indicators consisting of one statistical variable. All others indices consists of several statistical variables. Indices `housing`, `housing area`, `health` and `social relations` are each composed of three binary or categorical variables.

The score which results per index subject to the number of variables or variable categories is limited by linear transformation to the interval of 1 to 5 points. The maximal score to attain is 34, the minimum score 8. The un-weighted average across the seven indices ranges from 1.14 to 4.86 points.

Index 1 is composed of four sub-indices: `equivalised household net income`, `affordability of goods`, `possession of durables`, and `ability to make ends meet`. Entered into the sub-indices of `affordability of goods` and `possession of durables` are (in each case) four characteristics, assumed to be equivalent and so able to be uniformly evaluated.

The composite indicator is a linear statistic. Evaluation of individual characteristics can only, to a small extent, reflect the social evaluation of goods or their intrinsic economic value. Equidistant point-scoring could be modified in favour of an evaluation which is, say, oriented to the distribution of goods or the frequency distribution of characteristics within the population.

The extent to which weighting the indices equally does justice to their substantive significance also merits discussion, and readers will have their own views on this. Here readers are referred to the studies of KEI partners, KUL and JRC, who discuss the methods used in weighting the individual indices.

The primary goal of this simulation study is to statistically evaluate estimation methods within the context of a KBE. Changing the point-scoring system or adapting the weighting of individual components of the composite indicator can lead to a changed correlation between the targetted indices and the available regressors. Since in regression models the explanatory (exogenous) variables are usually selected in dependence on the targetted characteristic, the issue of point-scoring and weighting is not further pursued here. Rather, what try to do is implement the composite index as defined by ZUMA and shed light on the difficulties users face in doing so.

ZUMA has calculated the composite index for the European Community Household Panel[3] (ECHP), which was taken annually from 1994 to 2001 and with some 65,000 households participating throughout Europe. The EU successor study, EU-SILC, which was set up in 2004 and whose data are thoroughly up-to-date, does not contain - and here it differs from the ECHP - the following variables:

- new clothes (index income/ standard of living),
- membership in club or organization (index social relations)
- frequency of meeting friends or relatives (index social relations).

Thus, for the index of social relations two of the three co-opted variables are missing. So as to avoid missing out on this index, a further dataset was included in the analysis: the European Social Survey (ESS).

The datasets are presented in Section 2.2. The task of Section 2.3 is to process the assembled data (from GMC, EU-SILC and ESS) and to compare the datasets. Conclusions are drawn as to how far the composite index can be operationalised. Tables 2.5 and 2.6 illustrate the composite index that - in deviation from the original definition - is implemented within the frame of the simulation study.

A closer analysis of the datasets can be found in Annex B. Here the datasets used are contrasted in terms of their respective merits. The empirical frequency distributions and variable correlations are evaluated. All the frequency and correlation tables for the variables - as an addendum to this report - can be found in the Excel file, Deliverable33b.xls. And finally, the EU-SILC and ESS characteristics generated using Logit models are assessed for substantive plausibility.

2.2 Description of surveys

2.2.1 The German Microcensus

The German Microcensus (GMC) is a representative annual sample survey of the official German statistics. It is a household survey providing information about economic and social situations of household members as well as the labor market, education and employment of the active economic population. The Federal Statistical Office in Germany (DESTATIS) and the offices of the Federal States share the task of technical preparation and organizing, surveying the household data and editing the data.

The GMC is conducted as a one step stratified area or cluster survey, which has a selection quota of one percent of households in Germany (all citizens of Germany). Stratification is carried out for the Federal States, their administration divisions, regional stratums and building stratums. Sample units are the clusters of households or dwellings (sized approximately nine).

[3]For further information cf. http://www.eds-destatis.de/en/microdata/echp.php

domain	variable	categories of variable	score	calculation of subindex
`income / standard of living`	equivalised household net income as a percentage of the national median	less than 60% of the national median	1	Subindex `income/ standard of living` is an average scale across its four components ranging from 1 to 5.
		60% - < 90% of the national median	2	
		90% - < 110% of the national median	3	
		110% - < 140% of the national median	4	
		140% of the national median and above	5	
	affordability of:		number of affordable items: 0 = 1; 1 = 2; 2 = 3; 3 = 4; 4 = 5;	
	keeping home adequately warm	yes/no		
	annual holiday trip	yes/no		
	new clothes	yes/no		
	meat every second day	yes/no		
	possession of durables:		number of affordable items: 0 = 1; 1 = 2; 2 = 3; 3 = 4; 4 = 5;	
	car	yes/no		
	colour TV	yes/no		
	dishwasher	yes/no		
	telephone	yes/no		
	ability to make ends meet	with great difficulty	1	
		with difficulty	2	
		with some difficulty	3	
		fairly easily	4	
		easily, very easily	5	
`housing`	rooms per person	less than 1 room (excl. kitchen)	0	Subindex `housing` is defined as score sum of the three variables. Transformation of score: 0 = 1 1 = 2 2 = 3 3 = 4 4 = 5
		one room	1	
		more than 1 room	2	
	bath/WC available	both available	1	
		else	0	
	state of repair: leaky roof; damp walls, rot in window frames or floors	neither leaky roof nor dampness/ rot	1	
		else	0	
`housing area`	noise from neighbours or outside	yes/no	no = 1	Subindex `housing area` is defined as score sum of the three variables. Transformation of score: 0 = 1.00 1 = 2.33; 2 = 3.66 3 = 5.00;
	any pollution, crime or other environmental problem caused by traffic/industry	yes/no	no = 1	
	crime or vandalism in the area	yes/no	no = 1	

Table 2.1: Composite index of individual living conditions acc. to ZUMA definition (1)

domain	variable	categories of variable	score	calculation of subindex
`education`	education level	less than second stage of secondary level (ISCED 0-2)	1	Subindex `education` is equal to variable education level.
		second stage of secondary level (ISCED 3)	3	
		third level or still at school for a third level education (ISCED 5-7)	5	
`health`	self-rated health status	bad, very bad	0	Subindex `health` is defined as score sum of the three variables. Transformation of score: 0 = 1.00 4 = 3.66 1 = 1.66 5 = 4.33 2 = 2.33 6 = 5.00 3 = 2.99
		fair	1	
		good	2	
		very good	3	
	chronic health problem	yes	0	
		no	1	
	hampered in daily activities by any health problem	yes, severely	0	
		yes, to some extend	1	
		no	2	
`social relations`	household size	one person	0	Subindex `social relations` is defined as score sum of the three variables. Transformation of score: 0 = 1 1 = 2 2 = 3 3 = 4 4 = 5
		more than one person	1	
	membership in club or organisation	yes	1	
		no	0	
	fequency of meeting friends or relatives	never, less often than once a month	0	
		once/twice a month	1	
		on most days, once/twice a week	2	
`work`	activity status	working or economically inactive and not looking for work	4	Subindex `work` is defined as cross-combination of the two variables characterising the activity status.
	looking for work reasons for not looking for work	not working but looking for work or not working and not looking for work due to discouragement	2	

Table 2.2: Composite index of individual living conditions acc. to ZUMA definition (2)

The GMC is designed as a rotation panel (only a quarter of the units will be substituted by rotation of units) with a period of four years. Approximately 370,000 households with 820,000 respondents are selected by random. Data is gathered by interview questioning and filling in a questionnaire. The data of the 2004 GMC referred to the period from March 22^{nd} to the 28^{th} (reference week concept). Since 2005 the survey process takes place within a year (continuous microcensus).

The catalogue of relevant items of the GMC is as follows:

- Obligatory for all households is a fixed program that provides basic-relevant socio-economic data; an analysis of the labor market should be possible with this data.
- A complementary program which contains about 0.45 percent of the sample units to gain more detailed information for employment and common education and further training programs (European Labor Force Survey).
- An additional program with a four-year-period to gain further information about commuting to and from work, health, migration and dwelling conditions
- An ad-hoc module to get responses of persons whom are in a ten percent of the selected sample units; their responses are the basis of detailed labor market analysis.

The participation within the GMC is mandatory. The response rate was about 97 percent in the last few years. The unit non-response of approximately three percent is to trace back to households that could not be met by interviewer. Non-deficiencies are compensated by a special ad-hoc-weighting. The whole program is mandatory due the GMC law of 1996 and the `Bundesstatistikgesetz` (BStatG) from 1987. Additionally the instruction numbers 577/98 of the European council is a legal basis for the EU-Labor Force Survey.

The buildings are at first regionally stratified. Within these strata they are stratified once again by the size of the building and then arranged by region, street and house number. At last they are grouped to clusters that form the sample units. These units are sorted in terms of region. `Zones` are formed by 100 consecutive sample units. One sample unit is selected at random from each zone (semi-systematic sampling). Each dataset is weighted by a compensation factor for the known non-responses and adjusted by means of key data from the continuous population updating procedure. Including proxy-interviews the unit non-response is very low: ca. 3%. The sampling error of the interesting variables is calculated by the product of the variance under unrestricted random sampling and a design effect.

According to QUATEMBER et al. (2004) the structure of the sample design of the GMC can be described as follows:

Population of buildings
|
Stratification
(stratification variables: region and building-size)
|
Definition of sample units
(= clusters of households)
|
Semi-systematic selection of sample units

The Federal statistical law regulates the circulation of the Micro data to universities and other research institutions. Only 70% of sample data is available for extern users in the above mentioned institutions. For this purpose data must be made anonymous in a certain way, the `Faktische Anonymisierung`. This technique is based on random changes of the original data that can only be identified within the original units by a lot of painstaking effort, time and cost. Masking the data in this procedure ensures that the totals remain unchanged.

The essentials of these anonymising methods are:

- Drawing sub-samples from the original data (a 70% sample is drawn by systematic random sampling on household level);
- Coarsening the geographical information (NUTS 2 and lower)
- Coarsening and cancelling additional items (national groups with less than 50.000 members; totals referring to groups or variables less than 5.000 persons in the population).

Detailed and current information on the GMC is available at DESTATIS [4] and the internet platform of GESIS[5]. Information to data access for research activities is available at the German Research Data-center of the Federal Statistical Office and the state offices of the Länder[6].

The following sub-section introducing to EU-SILC and to the Access Panel is mainly based on EUROSTAT (2004), Körner et al. (2005), and Körner et al. (2006). Detailed information to the implementation of the Access Panel in the official statistics can be found in Körner et al. (2003).

2.2.2 The European Community Statistics on Income and Living Conditions

In Amsterdam, 1999, the `European Union` (EU) decided upon a task for all members to effectively combat poverty and social exclusion in the EU. Following this plan, the state leader of the member states resolved in the European Council in Lisbon (March 2000) the common target was to decrease poverty and social exclusion in the EU by 2010.

In this context, the member states agreed on observing the success of this task through a system of reports and indicators as a monitoring system. The setup of a reliable and compatible data base in order to construct and to calculate such indicators is one of the tasks of the European Community Statistics on Income and Living Conditions (EU-SILC). The agreement to generate this new statistic took place in 2003.

The EU-SILC will provide the database for calculating reliable and sustainable indicators for poverty and social exclusion. Poverty is defined by the European Commission as a

[4] Cf. `http://www.destatis.de`
[5] Cf. `http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/index.htm`
[6] Cf. `http://www.forschungsdatenzentrum.de/bestand/mikrozensus/index.asp`

lack of personal income as well as a lack of other sufficient sources related to the range of living standard defined and accepted in the relevant society. This is a relative concept of a standard of living in different countries.

Connected with this concept is social exclusion. Persons whom cannot fully share in the pursuit of social standard can be regarded as socially excluded in society. The main duty of the EU-SILC is the presentation of income data of private households. Next to this definition of household and income, the concept of social exclusion is of specific interest with the EU-SILC; but an operable definition is still missing. Monetary criteria alone are not sufficient enough to measure poverty and social exclusion. Improving the measurement of poverty and social exclusion beyond monetary aspects is a highly sophisticated multi-dimensional problem.

For certain population sections the EU-SILC provides a catalogue of factors determining social exclusion. A few very important factors are considered: economic insolvency, the physical and social environment, health, and the entrance to medical welfare services. To carry out the EU-SILC, a catalogue of criteria was fixed containing about 200 target variables. They are gathered as follows: socio-demographic information, e.g.

- structure of households
- education and training
- persons with employment
- income, housing and expenditures
- care of children younger 12 years
- social exclusion
- health.

To measure economic welfare of households is not adequate with only the observation of the monetary components. The concept of income that the EU-SILC takes into account non-monetary components, e.g. production of goods of small agricultural holdings being of great importance for rural regions in Middle and Eastern Europe. By counting of fictive rents regarding properties of housing and flats, it is possible to generate a quasi-income. Thus, income of property holders and of those paying a rent can be compared. Generally speaking, the EU-SILC is expected to become the EU reference source for producing structural indicators on social exclusion for the annual spring report to the European Council. It will provide two types of annual data:

- Cross-sectional data pertaining to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions, and
- Longitudinal data pertaining to individual-level changes over time, observed periodically over a four year period.

The first priority is to deliver comparable, timely and high quality cross-sectional data. Longitudinal data will be limited to income information and a limited set of critical qualitative, non-monetary variables of deprivation aimed at identifying the incidence and dynamic processes of persistence of poverty and social exclusion among subgroups in the population. The longitudinal component will also be more limited in sample size compared to the primary, cross-sectional component. Furthermore, for any given set of individuals, micro-level changes will be followed up only for a limited duration, such as a period of four years. For both the cross-sectional and longitudinal components, all household and personal data will be linkable. Furthermore, modules providing updated information in the field of social exclusion will be included starting in 2005.

The reference population of the EU-SILC is all private households and their current members residing in the territory of the member states at the time of data collection. Persons living in collective households and in institutions are generally excluded from the target population.

The `Statistical Office of the European Communities` (EUROSTAT) and the member states developed the technical aspects of the instrument. Five Commission Regulations (`Sampling and tracing rules`, `Definitions`, `list of primary target variables`, `Fieldwork aspects and imputation procedures`, `Quality reports`) implementing the Framework Regulation were elaborated.

The starting date for the EU-SILC is 2004 for the 12 member countries (except for Germany, the Netherlands, and the United Kingdom) Estonia, Norway and Iceland. All other countries (including the new members) started in 2005.

Since improving timeliness for statistics of the members of the EU has been one of the core objectives of the EU-SILC, priority has been given to the delivery of timely and comparable cross-sectional data. Another characteristic is flexibility in terms of data sources and sampling design. EUROSTAT strongly encourages the use of existing data sources, whether they are surveys or registers and the use of national sampling design. EUROSTAT recommends an integrated design (rotational design) for those countries planning to launch a new survey. This design aims to be the most cost effective as well as efficient for both cross-sectional and longitudinal requirements.

For all components of the EU-SILC, the cross-sectional and longitudinal data shall be based on a nationally representative probability sample of the population residing in private households, irrespective of language, nationality or legal residence status. All private households and all persons aged 16 and over within the household are eligible for the operation. By way of exception, Germany shall supply in the 2005 EU-SILC data for one fourth based on probability sampling and for three fourths based on quota samples, the latter to be progressively replaced by random selection so as to achieve fully representative probability sampling by 2008. Quota sampling will be 75% in 2005, 50% in 2006 and 25% in 2007.

The ultimate units used in the sample selection may be addresses, households or persons, each unit selected with a known probability. The analysis units can be households, all members, adult members, or possibly a subsample of adult members. These are the units to which the information collected pertains. Their probabilities of selection are determined through their association with the sample household.

For the period 1995 to 2001 the ECHP was used for the numerical calculation of indicators. The ECHP was a survey harmonized within the European members and the National Statistical Offices. Regarding the survey itself and the field work applied, e.g. a unified questionnaire and comparable survey techniques, this is a concept of input harmonization.

Parts of the statistical processing, for instance weighting and imputation, were directly carried out by EUROSTAT. In this context the high level of unification lead to a smoothing of national items, especially for the taxes and social insurances systems. Beyond this, the data seem to not be adequate for political purposes and politics itself. These arguments caused the decision to cancel the ECHP in 2001.

Principally the methodology of the EU-SILC follows the methodology of the ECHP, but introducing innovative principles eliminated some problems with the ECHP, mentioned above. The new statistic that the EU-SILC is based on is a frame of descriptions decreed by the European parliament and the European Council (see above). This frame regulates definitions, contents, sample sizes, methodological prescription as well as a schedule for all members. Flexible elements of the survey itself and details of the technical carrying-out are guided by the seven instructions of the European Commission.

The instructions within the instruction frame together with the executive order allow for an ample scope on the design of the EU-SILC. These instructions and orders deal with details of sampling and projections as well as with the practice of surveying and editing. However, some test results of the survey showed certain problems that are caused by the complex structure of the survey itself. As a survey provides data for national and European needs it was not possible to get reliable results of an integral EU-wide level. The burden load which varies from country to country was too high for the respondents in different member states.

In this strange situation the German Official Statistic decided to carry out an independent survey for the EU-SILC. Corresponding with the federal structure of the statistic in Germany, presentation, analysis and editing of the results for Germany is the duty of DESTATIS while the Federal States Offices survey and collect the data. The EU-SILC is described as `Leben in Europa` in Germany for better public recognition. The basic idea of the new survey is to permanently recruit households who have participated in the largest random sample of the general population in Germany, the microcensus. All households who have participated in their last microcensus interview (normally after four years) will be asked whether they are ready to participate from time to time in voluntary household surveys of official statistics.

Compared to other sampling frames for household surveys, the permanent sample is characterized by a number of special features that are essential for its effective use. As briefly mentioned, extensive socio-economic information is available from the last microcensus interview. This information can be used for an effective correction of the non-response bias occurring during the recruitment stage. In the context of the development of effective weighting techniques it is important to note that the socio-economic information from the microcensus is available for both participants and non-participants of the permanent sample.

The DSP could be used as a frame for most of the voluntary household and person surveys in Germany. Respecting design effects and panel mortality, the EU-SILC sample drawn from the access panel contains 14,100 households.

The anonymisation of micro-data is restricted by some special properties of the EU-SILC. This is caused by the hierarchical structure of data consisting information on household and personnel level. In addition, some longitudinal data of the panel are available. The Allocation of datasets especially for scientific use is fixed for the end of the year 2007 related to the survey of 2005. Longitudinal should be available for EUROSTAT in 2008, at the latest. As the 2006 EU-SILC data was not available before spring 2008, this data could, not be investigated in the KEI project.

2.2.3 The European Social Survey

The European Social Survey (ESS) was installed to measure the changing of social attitudes and values in the European nations discovering cross-cultural and cross-national commons and differences in individuality. The purpose is to chart and interpret the speed and direction of change scoped public attitudes of European members.

Some characteristics owing to the ESS will be concisely outlined in the following chapter which is mainly based on information provided on the ESS website[7]. Explicitly, we refer to Stoop et al. (2002) and point out the ESS Blueprint[8].

The ESS was funded by the 5^{th} Framework Program of the European Commission and the European Science Foundation. The first ESS round started in 2001. The third wave carried out in 2006, was conducted in 26 European countries. The ESS was motivated due to the lack of a comparative long-term survey providing data. This data should be available at a low cost, freely accessible for survey research and cover all facets of individual attitudes, beliefs and behaviors which are important in understanding modern societies. The ESS was developed as a conceptually new survey to be conducted according to a rigorous methodological standard.

Each wave of the ESS should basically consist of a core set of questions for the observation of change and persistence in attitudes as well as a core for social and demographic attributes, and additionally modules for specific topics including space for methodological testing.

There are two questionnaires, both administered by interviewers: A face-to-face interview questionnaire of around one hour average duration consisting of roughly 240 items and a short supplementary questionnaire for self-completion or face-to-face.

Around one half of the interview questionnaire comprises core items, both socio-demographic and substantive in nature. The other half of the face-to-face interview comprises rotating items varying from round to round in two separate modules. The rotating element consists of two topic-specific modules per round. In ESS Round 3 there are two rotating modules each of fifty items to measure particular academic and policy concerns and debates that require examination in depth.

The core questionnaire items cover both socio-demographic and substantive themes. The content of the core will remain largely constant for each round. It includes independent

[7] Cf. `http://www.europeansocialsurvey.org/`

[8] Report prepared for the Standing Committee for the Social Sciences (SCSS) of the European Science Foundation (ESF); cf. `http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_details&gid=4&Itemid=80`

and dependent variables, the latter designed to measure shifts over time in what are considered to be key components of Europe's social fabric. These core questions have been designed in collaboration with a group of experts in different fields and cover the following subjects:

- public trust in government, politicians and other major institutions
- political interest and participation
- socio-political orientations
- issues of governance and efficacy at the national and international level
- underlying moral, political and social values
- social inclusion and exclusion
- national, ethnic and religious allegiances
- well-being, health and security
- demographic composition - age, gender, marital status, etc.
- education and occupational background
- financial circumstances
- household circumstances

The purpose of the supplementary questionnaire is twofold. The first part contains 21 questions on human qualities which are asked for all respondents. The second part is devoted to test new or additional questions. For this purpose the sample of respondents will be split into three sub-groups, each sub-group getting a different set of 12 questions. The supplementary questionnaire may be administered either as an extension of the main interview questionnaire or as a self-completion questionnaire.

Any nation may add items to the questionnaire for national rather than multinational use. But any additional country-specific questions may be inserted only after the ESS questions in sequence, whether in the interview or in the supplementary questionnaire.

The survey will be representative of all persons aged 15 and over resident within private households in each country, regardless of their nationality, citizenship or language. The sample is to be selected by strict random probability methods at every stage. Quota sampling is not permitted at any stage, nor is substitution of non-responding households or individuals.

The minimum number of interviews to be achieved in each country is 2,000, and the recommended number is 2,500. In any event, the minimum number of effective interviews (after discounting for design effects, e.g. geographical clustering) will be 1,500. In countries whose total population is less than two million, the minimum number is 1,000 interviews and an effective sample size of 800.

To ensure the quality and reputation of the ESS, it is vital that the survey achieves the highest possible response rate in each country. The proportion of non-contacts should not exceed three percent of all sampled units, and the minimum target response rate - after discounting ineligibles - should be 70 percent.

It is possible to over-sample particular sub-groups, e.g. with respect to low-response strata and certain minority groups. But it is important to assess the effect of such over-sampling on the design effect owing to differing selection probabilities.

Enhancing response rates an interviewer has to visit a sample unit at least four times before it is abandoned (non-respondent), including at least one visit in the evening and at least one at the weekend. These visits have to be spread over at least two different weeks.

The sampling procedure used, and the response obtained, will be fully documented in a technical report of the survey. The report of sampling includes a definition and description of the sampling units used at each stage, a description of any stratification of the sampling frame, and a description of ways in which selection probabilities could have varied at each stage. Reports of response include the number of selected cases falling into each of the following categories: not eligible and why; no contact after four or more visits; personal refusal; proxy refusal; achieved interview (partial); achieved interview (full).

The national datasets from participating countries are sent within three months after completion of field work to the archive.

In Germany the ESS is drawn with help of the registers of local residents' registration offices. In these offices, all foreigners living in Germany as well as German citizens are listed. The registers are updated continuously. According to the 2002 ALLBUS[9] there are two independent target populations: one for West Germany (including West Berlin), and one for East Germany (including East Berlin) divided in regional categories: 1,085 strata in West Germany and 435 strata in East Germany.

A stratified two-stage probability sampling is applied. The design effect is derived from the first stage of the sampling design, i.e. because of clustering and from differing selection probabilities due to oversampling in East Germany. A net total sample size contains around 3,000 persons.

[9]German General Social Survey. cf. `https://download.za.gesis.org/`

2.3 The data manipulation process

2.3.1 The German Microcensus 2004

Base of operations is the `Scientific Use File` (SUF) 2004 which comprises 499,849 persons (data lines) and 331 variables (columns), beginning from EF1 (Federal state) and ending with EF756 (expansion factor for household and families). The data is encrypted in ASCII. The raw data file is read and converted by a SPSS syntax routine. The syntax provided by ZUMA

- converts alphanumeric variables into numeric variables
- substitutes blanks and @
- recodes and makes declaration of missing values
- reformats expansion factors
- adds variable names and variable labels
- saves the final data as SPSS file.

The data file is reduced by 290 variables which are not of interest for the simulation study. Further, 12,133 data rows are deleted by means of variables EF505 (population in principle residences) and EF506 (population in private households). These data rows concern i) persons in institutional accommodations (5,045 cases), e.g. mental institutions, and ii) population residing in additional residences (7,088 cases).

Furthermore, 22,705 data lines are deleted on the basis of variable EF539 (household net income). As household income is used as stratification variable in EU-SILC, all households which do not provide any income information are deleted completely (17,956 data lines). EU-SILC sampling design involves selection of farmer households. Due to non-availability of agricultural households for federal states `Berlin` and `Bremen` in the SUF, sampling of farmer households is not considered in the simulation study. Thus, households of self-employed farmers are deleted, too (4,749 data lines).

Afterward, the data is sorted by variables

- EF1: Federal state
- EF3: sample units (`Auswahlbezirk`)
- EF4: consecutive number of household in sample unit (`Auswahlbezirk`)
- head of the household

and saved as R-compatible data file (.dat format). It contains 465,011 individuals referring to 219,213 households.

Household income and personnel income are intended to be regressors in Logit models. In 2,497 households at least one household member does not provide any income information

(EF372, personnel net income). These households as well as 15 households, in which the eldest person is aged under 16, are deleted, as well. Finally, 216,701 households with 447,116 persons of which 385,866 individuals are aged 16 and over, remain in the population.

The head of household is identified by the following procedure: A binary vector indicating whether a person is householder (1) or not (0) was generated by variables EF507 (position in household) and EF509 (position in family). In all cases (7 households) in which a householder was not identifiable on the basis of these variables, the person drawing the highest salary is selected as head of household. A consecutive number is generated for the remaining households and its household members.

According to considerations of DESTATIS, the following federal states are pooled:

- Schleswig-Holstein and Mecklenburg-Western Pomerania (northern Germany)
- Hamburg, Bremen and Berlin (city states)
- Rhineland-Palatinate and Saarland (western Germany)
- Saxony, Saxony-Anhalt and Thuringia (eastern Germany).

Besides the Federal states, variables `type of household`, `social affiliation` and `household net income` are stratification variables in the German EU-SILC. Variables `type of household` and `social affiliation` have to be generated according DESTATIS classification by dint of the following GMC variables (cf. e.g. DESTATIS, 2004): EF127 (professional status), EF338 (predominant livelihood), EF539 (household net income), EF553 (marital status and cohabitation), EF627 (type of family/ cohabitation), EF637 (number of children aged ≥ 27) and EF640 (number of children aged < 18). The variables categories are as follows:

`type of household`:

- single-person household
- couple-household without children
- single parent
- couple-household with children
- other households

`social affiliation`:

- self-employed
- civil servant, employee, and apprentice
- pensioner
- other, or stay-at-home person

`household net income`:

- less than 900 Euro
- 900 Euro - 1300 Euro
- 1,300 Euro - 2,600 Euro
- 2,600 Euro - 3,600 Euro
- 3,600 Euro and over.

The composite indicator requires an `International Standard Classification of Education` (ISCED) which is generated according to classification of SCHROEDTER et al. (2006). Necessary GMC variables: EF74 (type of attended school), EF258 (graduation yes/no), EF259 (highest level of education), EF260 (training qualification yes/no), and EF261 (highest level of training qualification). The ISCED levels are as follows:

`International Standard Classification of Education (ISCED)`:

- level 1: primary education or first stage of basic education
- level 2: lower secondary or second stage of basic education
- level 3: upper secondary education
- level 4: post-secondary, non-tertiary education
- level 5: first stage of tertiary education
- level 6: second stage of tertiary education.

Several GMC variables, e.g. `household net income` or `age`, are recoded (new classification). This is to compute Logit models estimating EU-SILC and ESS variables of interest.

Programm 2.1: R syntax for generation of stratification variable `social affiliation`

```
# stratification variable social affiliation
# input variables:  EF127 (professional status)
#                   EF338 (What do you do for a living?)
#                   EF539 (household net income)

          # 1: self-employed with or without employees
szh     <- ifelse((EF127 == 1 | EF127 == 2 | EF127 == 3) & EF539!=50, 1,
          # 2: civil servant, employee, and apprentice
          ifelse((EF127 == 4 | EF127 == 5 | EF127 == 6 | EF127 == 7
               | EF127 == 8 | EF127 == 9 | EF127 == 10 | EF127 == 11) &
               EF539!=50, 2,
          # 3: pensioner
          ifelse(EF338 == 3, 3,
          # 4: other, or stay-at-home person
          4)))
```

Programm 2.2: R syntax for generation of stratification variable `type of household`

```
# stratification variable: type of household
# input variables:  EF553 (marital status and cohabitation)
#                   EF627 (type of family/ cohabition)
#                   EF637 (number of children aged >= 27)
#                   EF640 (number of children aged < 18)

          # 1: single-person household
hhtyp  <- ifelse(EF553==0, 1,
          # 2: couple-household without children
          ifelse((EF553==1 & EF627==1) | EF553==3, 2,
          # 3: single parent
          ifelse(EF553==7  & EF640 > 0 & EF637 < 1, 3,
          # 4: couple-household with children
          ifelse(((EF553==1 & EF627==2) | EF553==5) &
                   EF640 > 0 & EF637 < 1, 4,
          # 5: other households
          5))))
```

Programm 2.3: R syntax for generation of `ISCED`

```
# Initalization of ISCED variable
# ISCED level 0 (pre-primary education)
ISCED   <- rep(0,length(EF258))

# ISCED level 1 (primary education or first stage of basic education)
ISCED   <- ifelse((EF258==8 | EF258==9) & EF260==8,1,ISCED)
ISCED   <- ifelse(EF258==8 & EF260==9,1,ISCED)
ISCED   <- ifelse(EF74==1,1,ISCED)

# ISCED level 2 (lower secondary or second stage of basic education)
ISCED   <- ifelse((EF259==1 | EF74==2 | (EF258==1 & EF259==9)) &
                    (EF260==8 | EF261==1 | EF261==2 | EF260==9 |
                    (EF260==1 & EF261==99)),2,ISCED)
ISCED   <- ifelse((EF258==8 | EF258==9) &
                    (EF261==1 | EF261==2 | (EF260==1 & EF261==99)),
                    2,ISCED)
ISCED   <- ifelse((EF259==2 | EF259==3) &
                    (EF260==8 | EF261==1 | EF261==2 | EF260==9 |
                    (EF260==1 & EF261==99)),2,ISCED)

# ISCED level 3 (upper secondary education)
ISCED   <- ifelse(EF261==3 | EF261==4,3,ISCED)
ISCED   <- ifelse((EF259==4 | EF74==3) &
                    (EF260==8 | EF261==1 | EF260==9 |
                    (EF260==1 & EF261==99)),3,ISCED)

# ISCED level 4 (post-secondary, non-tertiary education)
ISCED   <- ifelse((EF259==4 | EF259==5 | EF74==3) &
                    (EF261==3 | EF261==4),4,ISCED)

# ISCED level 5 (first stage of tertiary education )
ISCED   <- ifelse(EF261==5 | EF261==6 | EF261==7 | EF261==8 |
                    EF261==9,5,ISCED)

# ISCED level 6 (second stage of tertiary education)
ISCED   <- ifelse(EF261==10,6,ISCED)
```

2.3.2 The European Community Statistics on Income and Living Conditions

The EU-SILC 2005 dataset consists of four separate data files:

	register information	collected data (survey)
households	`D-file`: Register of households which contains all households, including those, which did not participate in the survey. Dimension: 15,292 data lines, 29 variables.	`H-file`: Household data file containing all information which results from household questionnaires sent back. Dimension: 13,106 data lines, 138 variables.
individuals	`R-file`: Register of persons which contains all persons, including those, aged 15 and younger and those, which did not participate in the survey. Dimension: 31,276 data lines, 46 variables.	`P-File`: Personal data file containing all information which results from personal questionnaires sent back. Dimension: 24,982 data lines, 212 variables.

Table 2.3: EU-SILC data files: D-File, H-File, R-File, and P-File

The EU-SILC variable identification code includes the initial of the data file the variable belongs to (`D`, `H`, `R`, and `P`). Investigation of EU-SILC data is made by controlled remote data processing. A scientific use-file was not available before Spring 2008. The four EU-SILC data files are merged in order to generate

1. a person dataset and
2. a household dataset containing both, household and individual information.

Person data file

Register variables (`R`-File) are added to the `P`-File (questionnaires sent back). Matching key is the person identification number (variables RB030 and PB030). Household variables (`H`-File) are added to the enhanced person dataset. Key variables are RB030 and HB030 (household ID). The household ID is extracted from variable RB030 which is comprised of i) the household number and ii) the consecutive number of the household members. The NUTS level (variable DB040) is added to the combined person-household data file. The NUTS 2 information is reduced to NUTS 1 categories (Federal states) and the number of people per household is counted. Finally, income information which is equivalent to the income definition in the GMC is computed as sum of the following variables:

- HY020 (total disponibel household income)

- HY120 (regular taxes on wealth)
- HY130 (regular inter-household cash transfer paid)
- PY070 (value of goods produced by own-consumption)
- PY080 (pension from individual private plans).

Household data file

Household register information (D-File) is added to the household questionnaires (H-File). Key variable is the household identification number (variables DB030 and HB030). P-File variables are added to the extended household dataset. Matching variable is the person ID (HB070 and PB030). 45 persons from P-File cannot be matched to households. The reason is probably that theses persons have not filled out both, the household and the personal questionnaire. Finally, the GMC-equivalent income is generated, the NUTS 1 levels are computed, and the number of household members is counted.

Variables of interest not operable

EU-SILC variables are investigated with respect to i) their frequency distributions and ii) their correlation structure. The objective is to estimate best possible Logit models for those EU-SILC variables which are part of the ZUMA indicator in order to add them to the GMC database. Precondition for estimating Logit models is to adjust data for item non-response. Besides, categorical variables have to be transformed into binary data (dummy variables).

Some variables have been previously excluded due to a too high percentage of item non-response:

- HS030: arrears on hire purchase installments or other loan payments
- HS150: financial burden of the repayment of debts from hire purchases or loans
- PL060: number of hours usually worked per week in main job
- PL140: type of contract.

For further investigation, 35 EU-SILC variables (24 variables of interest and 11 regressor variables) are selected. Data lines containing one or more missing values are deleted. The person dataset comprises 3,800 (of 24,982) and the household dataset 1,215 (of 13,106) data rows with item non-response. Correlation analysis reveals some variables having a very weak correlation, in particular to potential regressors. Table 2.4 gives an overview on selected EU-SILC variables which are not possible to implement in the GMC due to a too poor correlation structure. Table elements which are not colorised stands for correlation coefficients less than ± 0.1. The table on the left half side shows correlation to potential regressors, the table on the right half side shows correlation coefficients among EU-SILC variables themselves. Among others, this concerns variables which are part of the ZUMA composite indicator:

variable	HS010	HS020	HS160	HS170	HS180	HS190	HH040	HH080	HH090	PH040	PH060
EF1.1	0.00	0.00	0.00	-0.01	-0.04	0.01	0.01	0.00	0.01	0.00	-0.01
EF1.2	0.00	0.00	0.03	0.05	0.05	0.11	-0.01	0.01	0.01	-0.01	0.00
EF1.3	0.00	-0.01	-0.01	-0.02	-0.03	-0.02	0.00	-0.01	0.01	0.00	0.01
EF1.4	0.00	-0.01	-0.01	0.00	0.06	0.05	-0.01	0.01	-0.01	-0.01	-0.01
EF1.5	-0.01	0.00	0.00	0.03	0.03	-0.01	-0.01	0.01	0.00	-0.01	0.01
EF1.6	0.00	0.01	0.01	-0.02	-0.02	-0.03	0.01	0.01	-0.01	0.02	0.01
EF1.7	-0.01	-0.01	0.00	0.00	-0.02	-0.05	0.00	0.01	0.00	-0.01	0.00
EF1.8	0.00	0.00	0.00	-0.02	-0.01	-0.05	0.00	-0.01	0.01	0.01	0.01
EF1.9	0.01	0.01	0.00	0.00	-0.03	0.01	0.02	0.01	0.01	0.00	0.00
EF1.10	0.02	0.01	-0.02	0.01	-0.01	-0.01	0.02	-0.03	-0.02	0.01	-0.01
EF30.1	0.04	0.06	0.03	0.02	0.02	0.03	0.03	0.01	0.00	0.04	0.04
EF30.2	0.02	0.04	0.03	0.04	0.01	0.01	0.06	-0.01	-0.01	0.06	0.04
EF30.3	0.04	0.02	0.01	0.00	0.03	0.03	0.05	0.01	0.02	0.09	0.03
EF30.4	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	-0.02	0.03	0.03
EF30.5	-0.03	-0.04	-0.01	-0.01	-0.01	-0.02	-0.04	0.01	0.00	-0.05	-0.02
EF30.6	-0.04	-0.04	-0.04	-0.03	-0.05	-0.04	-0.07	-0.02	0.01	-0.12	-0.08
EF32	0.00	0.00	-0.01	0.00	0.00	-0.01	-0.05	0.00	0.00	-0.07	-0.01
EF35.1	0.04	0.05	0.05	0.06	0.06	0.08	0.07	-0.04	0.00	0.07	0.05
EF35.2	-0.06	-0.06	-0.07	-0.06	-0.04	-0.08	-0.08	0.04	0.01	-0.09	-0.10
EF35.3	0.00	-0.01	-0.01	-0.03	-0.03	-0.02	-0.02	0.00	-0.01	-0.05	-0.03
EF35.4	0.04	0.03	0.05	0.04	0.02	0.04	0.05	-0.02	0.00	0.08	0.10
EF52	-0.02	-0.03	0.00	-0.01	0.00	0.00	-0.01	0.01	0.03	0.00	-0.01
EF110	-0.01	-0.02	0.00	-0.01	0.02	-0.02	0.00	0.02	0.01	0.05	0.01
EF127.1	0.01	0.00	0.00	-0.01	-0.02	-0.02	-0.01	-0.01	0.01	0.01	0.01
EF127.2	0.04	0.01	0.00	-0.01	0.01	-0.01	0.02	-0.02	-0.01	0.01	0.01
EF127.3	-0.03	0.00	0.00	0.02	0.01	0.02	-0.01	0.02	0.00	-0.01	-0.01
EF127.4	0.00	-0.01	-0.01	0.00	-0.01	-0.02	0.01	0.00	0.01	0.00	-0.01
EF358	0.12	0.12	0.06	0.08	0.04	0.07	0.09	-0.01	-0.02	0.14	0.14
EF359	0.09	0.13	0.03	0.07	0.05	0.05	0.07	0.00	-0.02	0.10	0.10
EF521.1	0.01	0.01	0.04	0.04	0.03	0.04	0.01	-0.05	-0.02	0.01	0.04
EF521.2	-0.03	-0.03	-0.01	0.00	-0.01	0.00	0.00	0.02	0.01	-0.06	-0.04
EF521.3	0.03	0.01	-0.01	0.00	0.00	-0.01	0.01	0.01	-0.01	0.02	0.01
EF521.4	-0.02	-0.01	-0.02	-0.02	-0.01	-0.03	-0.01	0.02	0.02	0.02	-0.02
EF521.5	0.02	0.04	-0.02	-0.04	-0.01	-0.01	0.00	0.01	0.00	0.03	0.01
EF539.1	0.04	0.07	0.00	0.03	0.01	0.02	0.02	-0.02	-0.03	0.04	0.04
EF539.2	0.06	0.03	0.03	0.03	0.01	0.05	0.05	-0.06	-0.04	0.08	0.08
EF539.3	0.03	0.04	0.03	0.04	0.02	0.03	0.04	0.00	0.01	0.04	0.07
EF539.4	0.00	0.01	0.02	0.02	0.01	0.01	0.02	-0.01	0.00	0.00	0.01
EF539.5	0.01	0.01	0.01	0.01	0.01	-0.01	-0.01	0.01	0.00	0.00	-0.01
EF539.6	-0.03	-0.04	-0.02	-0.01	0.00	-0.01	-0.01	0.02	0.01	-0.03	-0.03
EF539.7	-0.03	-0.03	-0.02	-0.05	-0.03	-0.05	-0.04	0.02	0.00	-0.03	-0.05
EF539.8	-0.03	-0.03	-0.05	-0.05	-0.03	-0.02	-0.05	0.01	0.02	-0.05	-0.05

variable	HS010	HS020	HS160	HS170	HS180	HS190	HH040	HH080	HH090	PH040	PH060
HS010	1.00	0.38	0.05	0.04	0.03	0.05	0.07	0.00	-0.01	0.12	0.11
HS020	0.38	1.00	0.06	0.04	0.05	0.06	0.10	0.00	-0.01	0.13	0.15
HS040	-0.16	-0.18	-0.07	-0.09	-0.05	-0.08	-0.13	0.01	0.02	-0.20	-0.20
HS050	-0.09	-0.10	-0.07	-0.09	-0.06	-0.07	-0.12	0.03	0.02	-0.14	-0.13
HS060	-0.19	-0.22	-0.08	-0.11	-0.07	-0.11	-0.16	0.01	0.02	-0.22	-0.23
HS070	-0.08	-0.09	-0.03	0.00	0.00	-0.02	-0.04	0.03	0.02	-0.04	-0.05
HS080	-0.01	-0.01	-0.04	-0.04	-0.03	-0.04	-0.04	0.04	0.03	-0.06	-0.05
HS090	-0.01	0.00	0.00	-0.02	0.01	0.01	0.01	0.03	0.00	0.04	0.00
HS100	-0.03	-0.05	-0.01	-0.03	-0.02	-0.03	-0.02	0.03	0.02	-0.04	-0.03
HS110	-0.08	-0.07	-0.04	-0.08	-0.06	-0.09	-0.05	0.06	0.02	-0.05	-0.08
HS120.1	0.24	0.23	0.10	0.06	0.05	0.08	0.09	0.00	-0.03	0.14	0.15
HS120.2	0.13	0.17	0.05	0.07	0.05	0.06	0.10	-0.02	-0.01	0.14	0.16
HS120.3	-0.02	-0.03	0.03	0.05	0.04	0.03	0.05	-0.02	0.00	0.08	0.06
HS120.4	-0.07	-0.08	-0.02	-0.02	-0.02	-0.02	-0.04	0.02	0.02	-0.07	-0.09
HS120.5	-0.07	-0.08	-0.07	-0.09	-0.07	-0.08	-0.11	0.01	0.01	-0.15	-0.13
HS140.1	0.09	0.10	0.05	0.09	0.07	0.07	0.11	0.00	-0.01	0.14	0.13
HS140.2	-0.05	-0.04	-0.02	-0.03	-0.02	-0.01	-0.04	0.02	0.02	-0.04	-0.05
HS140.3	-0.04	-0.05	-0.03	-0.05	-0.04	-0.05	-0.07	-0.02	-0.01	-0.10	-0.08
HS160	0.05	0.06	1.00	0.10	0.08	0.07	0.16	-0.02	-0.01	0.10	0.08
HS170	0.04	0.04	0.10	1.00	0.52	0.23	0.13	-0.02	-0.02	0.10	0.09
HS180	0.03	0.05	0.08	0.52	1.00	0.25	0.11	-0.01	-0.03	0.09	0.08
HS190	0.05	0.06	0.07	0.23	0.25	1.00	0.11	-0.01	0.00	0.10	0.08
HH030.1	0.03	0.02	0.04	0.04	0.03	0.07	0.00	0.00	-0.02	0.03	0.06
HH030.2	0.03	0.02	0.04	0.06	0.04	0.06	0.03	-0.05	-0.01	0.02	0.04
HH030.3	0.01	0.01	0.04	0.06	0.04	0.05	0.05	-0.01	-0.01	0.02	0.00
HH030.4	-0.02	-0.01	-0.03	-0.01	-0.01	-0.02	-0.01	0.02	0.02	0.00	0.00
HH030.5	-0.01	-0.02	-0.03	-0.07	-0.05	-0.05	-0.03	0.02	0.00	-0.03	-0.04
HH030.6	-0.01	-0.01	-0.03	-0.05	-0.02	-0.04	-0.04	0.02	0.00	-0.01	-0.02
HH030.7	-0.01	-0.01	-0.03	-0.04	-0.03	-0.04	-0.03	0.01	0.00	-0.01	-0.02
HH050	0.07	0.10	0.16	0.13	0.11	0.11	1.00	-0.04	-0.02	0.13	0.10
HH050	-0.11	-0.12	-0.07	-0.09	-0.07	-0.10	-0.15	0.02	0.01	-0.14	-0.13
HH080	0.00	0.00	-0.02	-0.02	-0.01	-0.01	-0.04	1.00	0.17	-0.01	-0.01
HH090	-0.01	-0.01	-0.01	-0.02	-0.03	0.00	-0.02	0.17	1.00	-0.02	-0.01
PH010.1	0.00	-0.01	-0.01	-0.04	-0.04	-0.02	-0.03	0.01	0.00	-0.07	-0.06
PH010.2	-0.01	-0.02	-0.04	-0.05	-0.03	-0.05	-0.04	0.01	0.00	-0.05	-0.05
PH010.3	0.00	0.00	0.01	0.05	0.03	0.03	0.02	0.00	0.00	0.06	0.04
PH010.4	0.03	0.04	0.06	0.05	0.05	0.07	0.07	-0.02	0.01	0.07	0.08
PH020	0.01	0.02	0.02	0.07	0.06	0.06	0.02	-0.02	0.00	0.03	0.05
PH030.1	0.04	0.04	0.05	0.05	0.04	0.05	0.05	0.00	0.01	0.06	0.08
PH030.2	0.01	0.01	0.03	0.04	0.04	0.05	0.04	-0.01	0.01	0.07	0.05
PH030.3	-0.03	-0.03	-0.05	-0.07	-0.07	-0.08	-0.06	0.01	-0.01	-0.10	-0.09
PH040	0.12	0.13	0.10	0.10	0.09	0.10	0.13	-0.01	-0.02	1.00	0.47
PH060	0.11	0.15	0.08	0.09	0.08	0.08	0.10	-0.01	-0.01	0.47	1.00

legend to correlation table:

value:	<= -0,4	<= -0,3	<= -0,2	<= -0,1	0	>= 0,1	>= 0,2	>= 0,3	>= 0,4
colour:									

Table 2.4: Correlation table of EU-SILC variables which were not operable

- HH040: leaking roof, damp walls/ floors/ foundation, or rot in window frames or floor
- HH080: bath or shower in dwelling
- HH090: indoor flushing toilet, for sole use of household
- HS160: problems with the dwelling: too dark, not enough light
- HS170: noise from neighbours or from the street
- HS180: pollution, crime or other environmental problems
- HS190: crime, violence or vandalism in the area.

This causes a lack of information in indices `housing` (variables HH040, HH080, HH090) and `housing area` (variables HS170, HS180, HS190). As mentioned on page 13, variables `new clothes`, `membership in club or organisation` and `frequency of meeting friends and relatives` co-opted to indicators `standard of living` and `social relations` are not covered by EU-SILC, too. Thus, the ESS is being investigated.

2.3.3 ESS variables of interest

ESS data is disposable from the ESS homepage after registration. Data available for download is of ESS round 1 (2002/03), round 2 (2004/05) and round 3 (2006/07). The latest dataset comprises 2,916 data lines for Germany. As done for EU-SILC, ESS variables are investigated with respect to their frequency distributions and to their correlation structure. Some interesting variables regarding indicator `work` have been previously excluded due to a too high percentage of item non-response (48.39% to 50.82%):

- E48, STFJB: how satisfied with job
- E49, STFJBOT: satisfied with balance between time on job and time on other aspects
- E51, JBSTRS: find job stressful
- E52, UEMPNYR: become unemployed in the next 12 months, how likely.

24 ESS variables are selected (12 variables of interest in addition to 12 regressor variables). Adjusting for item non-response, the data matrix reduces by 984 (33,74%) to 1,932 data rows.

2.3.4 Re-Definition of the Composite Index on Individual Living Conditions

The ESS is chosen in order to substitute the variables which are not available or manageable in EU-SILC, but which are required for the CIoILC. Due to a lack of exact equipollents, some missing variables are replaced by other variables. For this reason, the definition of the composite indicator could not be adopted as is. The following single indicators are enhanced by ESS variables:

- `standard of living`
  - NETUSE: personnel use of internet/ e-mail/ www
  - MBLTPH: personally have mobile telephone
  - STFSDLV: satisfied with standard of living
  - WRINCO: worried that income in old age will not be adequate to cover later years
- `housing`
  - AESFDRK: feeling of safety of walking alone in local area after dark
  - BRGHMEF: worry about home burgled has effect on quality of life
- `social relations`
  - WRKPRTY: worked in political party or action group in last twelve months

- WRKORG: worked in another organisation or association last twelve months
- WKVLORG: involved in work for voluntary or charitable organisations, how often past twelve months?
- SCLMEET: how often socially meet with friends, relatives or colleagues?
- FLCLPLA: feel close to the people in local area?
- FLTLNL: felt lonely, how often in the past week?

The authors divide index `income and standard of living` in single indicators `income` and `standard of living`. Further, more variables are added to indicator `standard of living` in order to statistically discriminate households to a greater extent. It should be noted, that sub-index `posession of durables` has to be updated from time to time with respect to technological change. Thus, variables `colour TV` and `telephone` are replaced by `mobile phone` and `internet access`.

Failing with any feasible substitutes for variables which specify housing area and repair status, indices `housing` and `housing area` are pooled to single indicator `housing`.

The allocation of score points and the calculation of the composite indicator are taken over as proposed by ZUMA. The modified composite indicator ranges from 7 to 35 points, whereas each of the seven single indicators ranges from 1 to 5 score points. Altogether, the re-defined composite indicator includes more information about population.

2.3.5 Evaluation of generated EU-SILC and ESS variables

As KEI is a policy-orientated research project, the aim within this deliverable is to work with a population which - even semi-synthetic - is to a large extend realistic in terms of reproducing the data structure as given in EU-SILC and ESS. Thus, the original and the generated variables are compared by relative frequencies, correlation structure and conditional distribution, e.g. distribution conditioned on age or income. This quality and plausibility check has benefit on the data generation process. In the following, the Logit model selection for EU-SILC and ESS variables of interest is presented.

Estimation of Logit models

Logit models are estimated under R[10] (GNU Licence) using R-packages `car` and `leaps`. As illustrated in FOX (2002), pages 220 et seqq, an optimal set of regressors is selected for each variable of interest via exhaustive search (forward-backward-selection). Categorical variables are estimated via ordered logistic regression. The model fit is evaluated by classical measures, such as residual sum of squares, `Akaike Information Criterion` (AIC), and `Bayesian Information Criterion` (BIC). The algorithm sequentially selects the best subset of regressors from a predefined set of regressor variables. The quality criteria are displayed for an increasing number of selected regressors.

[10]The R foundation for Statistical Computing, R version 2.6.2; cf. `www.r-project.org`

single indicator	variable: name /description	variable: source, code	variable: categories	score	transformation of score	calculation of single indicator
`income`	equivalised household net income as a percentage of the national median	MC: EF539	less than 60% of the national median	1		Single indicator `income` is equal to variable `equivalised household net income`.
			60% - < 90% of the national median	1.8		
			90% - < 110% of the national median	2.6		
			110% - < 140% of the national median	3.4		
			140% - < 170% of the national median	4.2		
			170% of the national median and above	5		
`standard of living`	affordability of:				transfer to	Indicator `standard of living` is an average scale across its five components. The indicator ranges from 1 to 5 points.
	keeping home adequately warm	EU-SILC: HH050	yes/no	1/0	range [1 ; 5]:	
	one week annual holiday trip	EU-SILC: HS040	yes/no	1/0		
	meal with meat every second day	EU-SILC: HS050	yes/no	1/0	0 = 1	
					1 = 1.66	
	capacity to face unexpected financial expenses	EU-SILC: HS060	yes/no	1/0	2 = 2.33	
					3 = 3	
	financial burden of total housing cost	EU-SILC: HS140	a heavy burden	0	4 = 3.66	
			somewhat a burden	1	5 = 4.33	
			not burden at all	2	6 = 5	
	possession of durables:				number of	
	car	EU-SILC: HS110	yes/no	1/0	affordable	
	washing maschine	EU-SILC: HS100	yes/no	1/0	items:	
	computer	EU-SILC: HS090	yes/no	1/0		
	internet access	ESS: A7, NETUSE	yes/no	1/0	0 = 1; 1 = 2;	
	mobile telephone	ESS: F72, MBLTPH	yes/no	1/0	2 = 3; 3 = 4;	
					4 = 5;	
	ability to make ends meet	EU-SILC: HS120	with great difficulty	1		
			with difficulty	2		
			with some difficulty	3		
			fairly easily	4		
			easily, very easily	5		
	satisfied with standard of living	ESS: E32, STFSDLV	(extremly) dissatisfied	1		
			neutral	3		
			(extremly) satisfied	5		
	worried that income in old age not be adequate to cover last years	ESS: D53, WRINCO	worried or extremely worried	1		
			neutral	3		
			not worried, not worried at all	5		
`housing`	number of rooms per person	EU-SILC: HH030	less than one room (excl. kitchen)	0	transfer to	Single indicator `housing` is defined as score sum of the three variables.
			one room	1	range [1 ; 5]:	
			more than one room	2		
					0 = 1	
	feeling of safety of walking alone in local area after dark	ESS: C6, AESFDRK	yes/no	1/0	1 = 2	
					2 = 3	
	worry about home burgled has effect on quality of life	ESS: C8, BRGHMEF	yes/no	1/0	3 = 4	
					4 = 5	

Table 2.5: Modified composite index of individual living conditions (Table 1)

single indicator	variable			score	transformation of scale	calculation of single indicator
	name /description	source, code	categories			
`education`	educational level	MC: EF258	ISCED 0/1: pre-primary or primary	1		Indicator `education` is equal to variable `educational level`.
		MC: EF259	ISCED 2: lower secondary	1.8		
		MC: EF260	ISCED 3: upper secondary education	2.6		
		MC: EF261	ISCED 4: post-secondary, non tertiary	3.4		
			ISCED 5: first stage of tertiary	4.2		
			ISCED 6: second stage of tertiary	5		
`health`	self-rated health status	EU-SILC: PH010	bad, very bad	0	transfer to range $[1\,;5]$:	Single indicator `health` is defined as score sum of the three variables.
			fair	1		
			good	2		
			very good	3		
					$0 = 1$	
					$1 = 1.66$	
	chronic health problem	EU-SILC: PH020	yes/no	1/0	$2 = 2.33$	
					$3 = 3$	
	hampered in daily activities	EU-SILC: PH030	yes, severely	0	$4 = 3.66$	
			yes, to some extend	1	$5 = 4.33$	
			no	2	$6 = 5$	
`social relations`	household size	MC: EF521	one person	0	transfer to range $[1\,;5]$:	Single indicator `social relations` is defined as score sum of the entered items.
			more than one person	1		
	membership in political party organisation/association	ESS: B14, WRKPRTY	yes/no	1/0	$0 = 1$	
		ESS: B15, WRKORG			$1 = 1.57$	
					$2 = 2.14$	
	work for voluntary or charitable organisation	ESS: E1, WKVLORG	yes/no	1/0	$3 = 2.71$	
					$4 = 3.29$	
					$5 = 3.86$	
	how often socially meet with friends, relatives or colleagues	ESS: C2, SCLMEET	less than once a month	0	$6 = 4.43$	
			once or serveral times a month	1	$7 = 5$	
			several times a week up to every day	2		
	feel close to the people in local area	ESS: E45, FLCLPLA	yes/no	1/0		
	felt lonely past week	ESS: E12, FLTLNL	yes/no	1/0		
`work`	activity status	MC: EF214	unemployed	1		The indicator is equal to variable `activity status`.
			employed	5		

Table 2.6: Modified composite index of individual living conditions (Table 2)

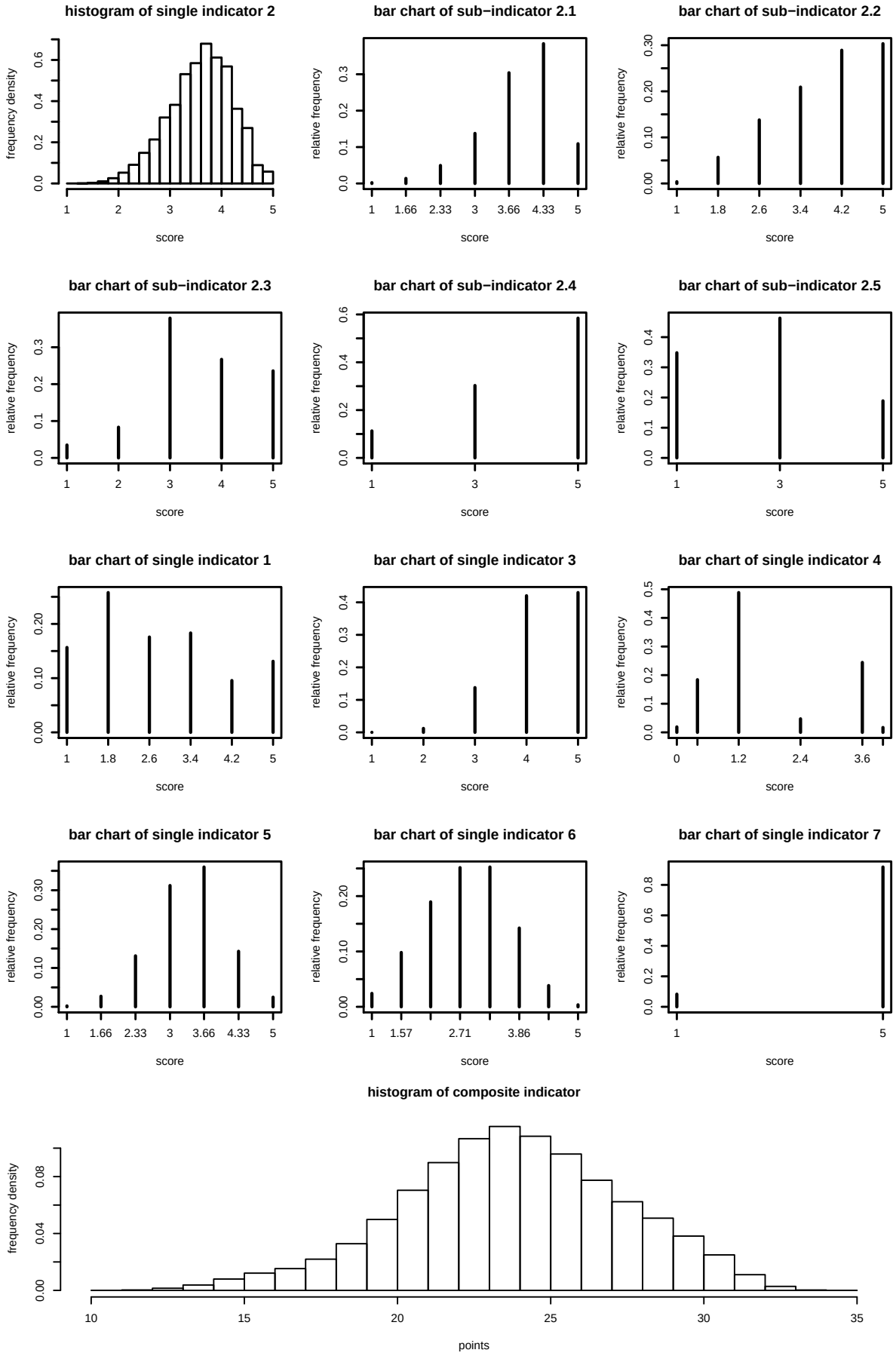

Figure 2.1: Relative frequency distribution of single indicators and density plot of CIoILC

In the following, two examples for the development of the adjusted R squared for an increasing number of regressors are given in Figures 2.2 and 2.3. Figure 2.2 shows the adjusted R squared for variable HH030 (number of rooms available to the household) and figure 2.3 for variable HS160 (problems with the dwelling: too dark, not enough light). Figure 2.2 show - by way of comparison - a high adjusted R squared for variable HH030, approaching towards the graphical asymptotic maximum located at 0.41. The adjusted R squared estimating variable HS160 is very low (maximum located at 0.0126). As mentioned in Subsection 2.3.2, HS160 could not have been operationalised. Besides, there is no improvement of the adjusted R squared beginning from 21 regressors upwards.

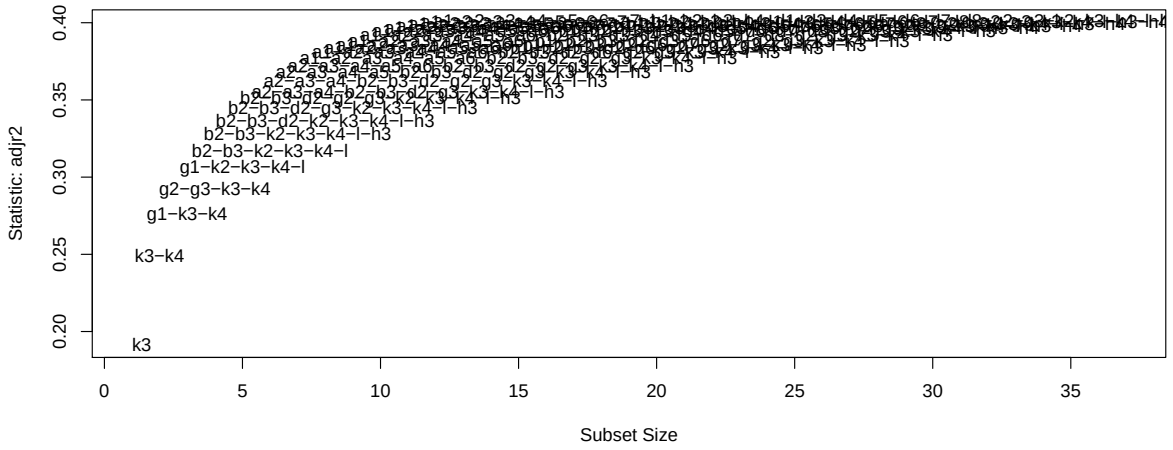

Figure 2.2: Development of adjusted R squared for variable HH030

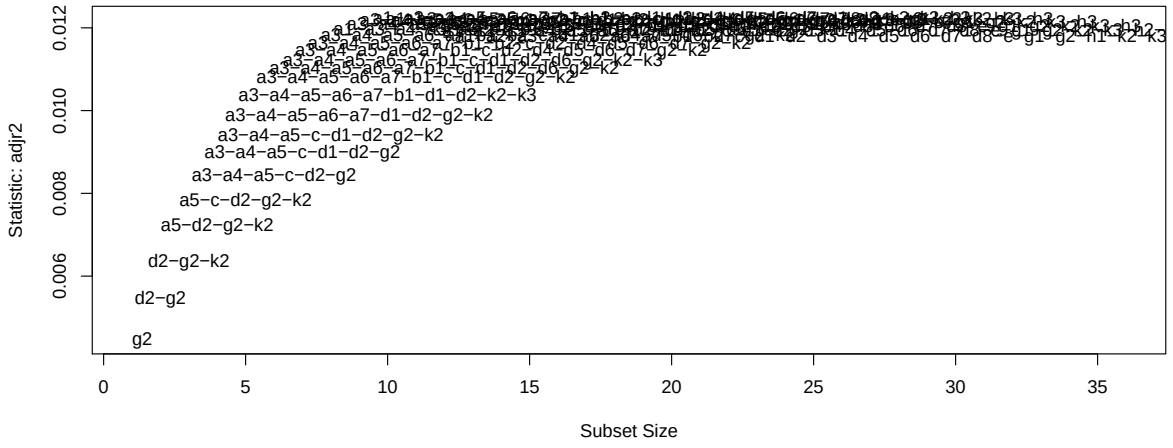

Figure 2.3: Development of adjusted R squared for variable HS160

The relative frequencies of original and generated EU-SILC and ESS variables are congruent to far extend. However, some higher relative deviations from the original frequency distribution are observed for several variables. In particular the difference is realized to a high degree for categorical variables, e.g. FLCLPLA and SCLMEET.

The Logit models are - from a statistical point of view - evidence-based. The regressors are statistically significant in the context of conventional distribution assumptions. Thus, the deviations in frequency distribution and correlation structures are due to different

structures of datasets (GMC, EU-SILC and ESS). The differences discovered are analysed in detail in Annex B.1 due to formal reasons. It is expressly pointed out that the extent of detail would go beyond the scope of this section.

The frequency distributions and the correlation tables of variables investigated in the simulation study are documented in detail in the Microsoft Excel file `Deliverable33b.xls` downloadable from the KEI server `www.kei.publicstatistics.net`. Frequency and correlation tables provided in this report always refer to datasets adjusted for non-response and persons aged 16 and over. The list of variables and some selected correlation tables are given in appendices A and B.

Chapter 3

Accuracy measurement of the Composite Indicator

Within the KEI project, a large set of single indicators were collected. Unfortunately, data quality reports of indicators including information to timeliness, coherence, comparability, accuracy and the underlying population are unlikely to be available on all indicators or on country specific disaggregation. In particular, precise point and variance estimates for most single indicators are not fully available to end-users. Furthermore, there is hardly any information about the interaction of single indicators. This may become evident for evaluating the accuracy of composite indicators. The aim of this chapter is to investigate data quality of the composite indicator on individual living conditions and to estimate its variance. Point and variance estimates are carried out using calibration estimators. Several sets of auxiliary information (regressor matrices) and survey designs are evaluated.

3.1 Quality of single and composite indicator estimates

One important task of quality measurement is dedicated to accuracy measurement and, to be precise, to variance estimation in the case of sample survey data. Assuming unbiased or approximately unbiased estimates of single indicators, the accuracy of a single indicator may be measured by its variance, its relative root mean square error or more general by confidence interval coverage. Once point and variance estimates are developed for single indicators, the major question arises whether these values suffice to gain also point and variance estimates for composite indicators. This can be formalized as follows.

Let $x^t_{i;c}$ be the outcome of a (single) indicator i ($i = 1, \ldots, \nu$, with ν variables of interest), for country c ($c = 1, \ldots, C$) at time t ($t = 1, \ldots, T$). Then, a composite indicator can be defined as a function

$$\Psi_{c,t} = \Psi_{c,t}\left(x^t_{1,c}, x^t_{2,c}, \ldots, x^t_{\nu,c}\right) \qquad : \qquad \mathbb{R}^\nu \to \mathbb{R}^1 \qquad . \tag{3.1}$$

In practice, the functional Ψ is reduced to a (convex) weighted sum of the single indicators where the normalized values of the single indicators $y^t_{1,c} := f(x^t_{1,c})$ with a suitable function

f. The function f can be either a normalizing function (cf. Nardo et al., 2005) or a (nonlinear) transformation, e.g. the quintile share ratio as a prominent representative of the Laeken indicators (cf. EUROSTAT, 2008).

Since the outcomes of the single indicators or at least parts of them are generally gained as estimates from survey data, and for simplicity reasons Equation (3.1) can be rewritten as

$$\widehat{\Psi}^*_{c,t} = \sum_{i=1}^{\nu} \gamma_i \cdot \widehat{y}^t_{i,c} \quad . \tag{3.2}$$

The $*$ denotes the transformed (normalized) single indicators. The weights γ_i are meant to be independent of the outcomes of the single indicators and sum up to one.

Two major goals may be connected with the use of composite indicators (for a thorough discussion of composite indicators we refer to workpackages 5 and 7), benchmarking and development. As an important measure variance estimates have to be derived in both cases. The less complicated case refers to the cross sectional view which will be used throughout this study. The variance of the composite indicator (3.2) is then

$$\mathrm{V}\left(\widehat{\Psi}^*_{c,t}\right) = \sum_{i=1}^{\nu} \gamma_i^2 \cdot \widehat{y}^{t\ 2}_{i,c} + 2 \cdot \sum_{\substack{i=1 \\ i<j}}^{\nu} \sum_{j=1}^{\nu} \gamma_i \cdot \gamma_j \cdot \mathrm{cov}\left(\widehat{y}^t_{i,c}; \widehat{y}^t_{j,c}\right) \quad . \tag{3.3}$$

The following sections are devoted improving estimates of the single indicators via calibration as well as of the derivation of the corresponding variance estimates. Further, the methodology will be exploited for applications to composite indicators. The main focus will be laid on the accuracy measurement of composite indicators. This, however, is strongly connected to the question whether composite indicators including accuracy measurement can be derived by end-users or only by NSIs. In fact, this question can be reduced to whether the covariances of Equation (3.3) vanish or are at least negligible.

3.2 Description of calibration estimators

The objective of calibration estimation is to adjust samples through re-weighting of sample units (households, individuals, companies etc.) to external data relating to the distribution of units of investigation in the target population, e.g. census data or administrative data files. The procedure of calibration aims at improving the accuracy of estimates by using auxiliary variables whose values are available both, for the responding sample units and for the whole population (population totals or category frequencies). These auxiliary variables are called calibration variables.

Consider a finite population $\mathcal{U} = \{1, \ldots, \kappa, \ldots, N\}$ from which a probability sample $\mathcal{S}$ ($\mathcal{S} \subseteq \mathcal{U}$) of size n is drawn. The first order inclusion probability of the kth sample unit is π_k. Let y be the variable of interest and $\mathbf{x}$ a p-vector of auxiliary variables. For all elements $k \in S$ the value y_k and $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kp})$ is observed. Further, the population total of $\mathbf{x}$, $\boldsymbol{\tau}_x$ is known. It is

$$\tau_{x_j} = \sum_{\kappa \in \mathcal{U}} x_{j\kappa} \quad .$$

An unbiased estimate of the population total t_y according to HORVITZ AND THOMPSON (1952) (Horvitz-Thompson estimator) is

$$\widehat{\tau}_{y,\mathrm{HT}} = \sum_{k\in\mathcal{S}} \tfrac{1}{\pi_k} y_k = \sum_{k\in\mathcal{S}} d_k y_k \quad .$$

The inverse inclusion probability of a sample unit $d_k = 1/\pi_k$ is called design weight. The target of the calibration procedure consists of finding new (calibrated) weights $w_k = g_k d_k$ so that the calibration constraints

$$\sum_{k\in\mathcal{S}} w_i \boldsymbol{x_i} = \boldsymbol{\tau_x} \tag{3.4}$$

are satisfied and that the calibrated weights w_k remain as close as possible to the design weights d_k. The final weights are a result of minimising - for any particular sample S - the quantity

$$\min_{\boldsymbol{g}\in\Omega_g} \sum_{k\in\mathcal{S}} d_k G_k\Big(\tfrac{w_k}{d_k}\Big)$$

subject to the constraints (3.4). Occasionally, the g-weights have to satisfy some additional boundary constraints $g \in \Omega_g$. The distance function $G(g)$ (cf. 3.2) with argument $g_k = w_k/d_k$ is - from a mathematical point of view - positive, strictly convex and twice continuously differentiable on the interior of its domain. It is $G(1) = 0$, $G^{'}(1) = 0$, and $G^{''}(1) = 1$. The g-weight g_k measures the difference between the basic weight d_k and the final weight w_k. In case $g_k = 1$ the design weight d_k remains unchanged, that is $w_k = d_k$.

The calibration problem presents as non-linear optimisation problem

$$\min_{\boldsymbol{g}\in\Omega_g} \sum_{k\in\mathcal{S}} d_k G_k\Big(\tfrac{w_k}{d_k}\Big) - \boldsymbol{\lambda}^T\Big(\sum_{k\in\mathcal{S}} w_k \mathbf{x_k} - \sum_{k\in\mathcal{U}} \mathbf{x_k}\Big)$$

leading to the Lagrange-function

$$L(\mathbf{w}, \boldsymbol{\lambda}) = \sum_{k\in\mathcal{S}} d_k G(g_k) - \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{w} - \boldsymbol{\tau}_x)$$

with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_j, \ldots \lambda_p)$ as Lagrange multiplier. Partial differentiating ($\frac{\partial L}{\partial \mathbf{w}} = 0$ and $\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0$), the resulting system of dimension $(n+p)\times(n+p)$ can be transformed to a $(p \times p)$ system in $\boldsymbol{\lambda}$:

$$\boldsymbol{\Phi}(\boldsymbol{\lambda}) = \sum_{k\in\mathcal{S}} d_k F(\mathbf{x}_k^T\boldsymbol{\lambda})\mathbf{x}_k - \boldsymbol{\tau}_x = \mathbf{X}^T\mathbf{w}(\boldsymbol{\lambda}) - \boldsymbol{\tau}_x = \widehat{\boldsymbol{\tau}}_x - \boldsymbol{\tau}_x = \mathbf{0} \quad . \tag{3.5}$$

$F(u)$ is the inverse function of $G(g)$ called calibration function. As G is positive and strictly convex, the existence of F is guaranteed. It is $F(0) = 1$ und $F^{'}(0) = 1$. Calculation of the calibrated weights

$$w_i = w_i(\lambda) = d_i F(\mathbf{x}_i^T\lambda) = d_i g_i(\lambda) = d_i g_i \quad \text{respectively} \quad \mathbf{w}(\boldsymbol{\lambda}) = \mathbf{D}F(\mathbf{X}\boldsymbol{\lambda})$$

require to determine the Langrange multiplier $\boldsymbol{\lambda}$ by solving equation 3.5 iteratively for the unknown vector. A general solution is given by the Newton-Raphson algorithm or the Fisher-Scoring algorithm.

$$\boldsymbol{\lambda}^{(l)} \;=\; \boldsymbol{\lambda}^{(l-1)} - \Big(\boldsymbol{\Phi}'(\boldsymbol{\lambda}^{(l-1)})\Big)^{-}\; \boldsymbol{\Phi}(\boldsymbol{\lambda}^{(l-1)})$$

$$= \boldsymbol{\lambda}^{(l-1)} - \left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\lambda}^{(l-1)})\mathbf{X}\right)^{-} \mathbf{X}^T\mathbf{w}(\boldsymbol{\lambda}^{(l-1)}) - \boldsymbol{\tau}_x \quad . \tag{3.6}$$

It should be noted, that $\boldsymbol{X}$ is the $(n \times p)$ calibration design matrix. Matrices $\mathbf{D}$ and $\mathbf{W}$ are diagonal weight matrices of dimension $(n \times n)$. Evaluation of equation 3.6 involves computation of a generalised inverse $\left(\boldsymbol{\Phi}'(\boldsymbol{\lambda}^{(l-1)})\right)^{-}$ as descripted in Vanderhoeft (2003). In case of redundant information matrix $\boldsymbol{X}$ need not to be of full rank. After determining $\boldsymbol{\lambda}$, the calibration estimator of τ_y can be computed by

$$\widehat{\tau}_{y,\mathrm{CAL}} = \sum_{k\in\mathcal{S}} d_k F(\boldsymbol{x}_k^T\boldsymbol{\lambda}) y_k = \sum_{k\in\mathcal{S}} w_k y_k \quad .$$

Deville and Särndal (1992) propose seven distance functions. As presented in Deville et al. (1993), four distance functions are retained by Sautory (1991) in constructing SAS macro CALMAR (CALage sur MARge)[1], a computer programme for calculation of calibrated weights:

1. the linear method
2. the multiplicative method
3. the logit method
4. the linear truncated method.

Table 3.2 heads the distance and calibration functions appendant to item 1 to 4. Further, figure 3.1 displays the functions. Distance functions 2 to 3 guarantee positive weights. Distance functions 3 and 4 restrict the range of calibriated weight by specifying a lower and an upper boundary L and U adding an additional constraint to the mathematical programming problem. Thus, in cases 3 and 4 a solution is not guaranteed. Calibration methods 1 and 2 always lead to a solution.

N^o	distance function $G(g)$	calibration function F(u)	Ω_B
1.	$\frac{1}{2}(g-1)^2$ for $g \in \mathbf{R}$	$1+u$ for $u \in \mathbf{R}$	$\mathbf{R}^n$
2.	$g\ln(g) - g + 1$ for $g \in \mathbf{R}_0^+$	e^u for $u \in \mathbf{R}$	$[0,+\infty)^n$
3.	$\frac{1}{A}\left((g-L)\log\frac{(g-L)}{1-L} + (U-g)\log\frac{U-g}{U-1}\right)$ for $g \in (L,U)$ and $0 \le L \le 1 < U$ where $A = \frac{U-L}{(U-1)(1-L)}$	$\frac{L(U-1)+U(1-L)e^{Au}}{(U-1)+(1-L)e^{Au}}$ for $u \in \mathbf{R}$ where $A = \frac{U-L}{(U-1)(1-L)}$	$[L,U]^n$
4.	$\frac{1}{2}(g-1)^2$ for $g \in \mathbf{R}$ and $L < 1 < U$	$1+u$ for $u \in [L-1, U-1]$ L for $u \ge L-1$ U for $u \ge U-1$	$[L,U]^n$

Table 3.1: Distance and calibration functions following Vanderhoeft (2003)

[1]CALMAR is an SAS driven macro developed at the National Institute for Statistics and Economic Studies (INSEE), Paris. It is available for download under `http://www.insee.fr/fr/methodes/outils/calmar/calmar.zip`

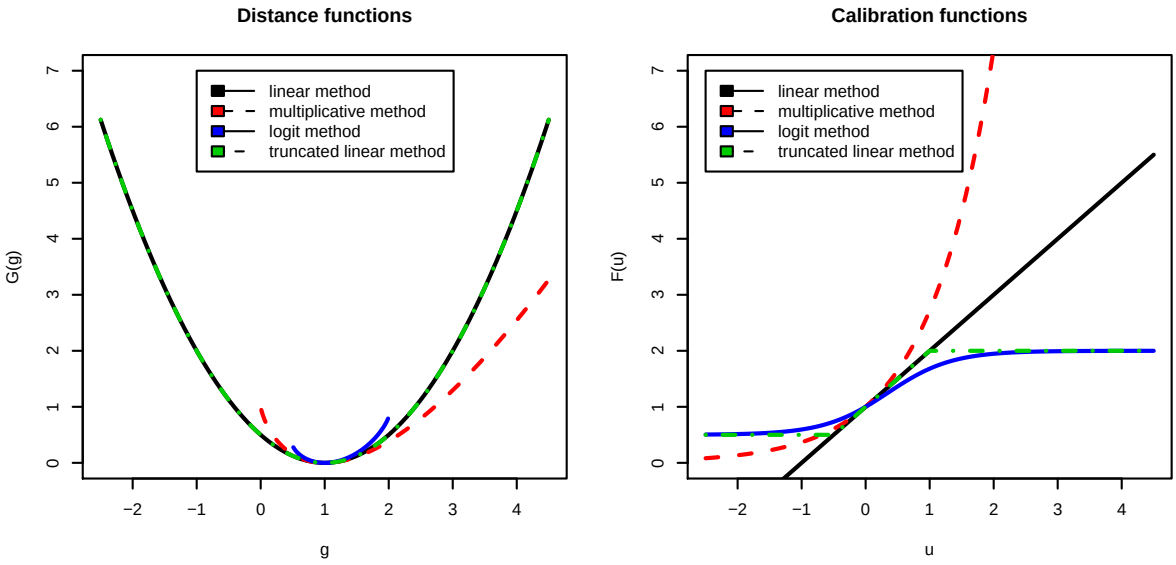

Figure 3.1: Distance and calibration functions with $L = 0.5$ and $U = 2$

Estimators based on the linear distance function of type $F(u) = (1 + \alpha u)^{1/\alpha}$, where $\alpha = 1$, are GREG estimators. The multiplicative method results for $\alpha \to 0$ and the minimum entropy distance for $\alpha = -1$. Weights derived from GREG estimator can be positive or negative. Negative weights may be unacceptable to some users, in particular when applying these weights to make estimates for geographic or socio-demographic subpopulations. An advantage in application is it's computational speed, especially in the context of Monte-Carlo-Simulation study. The linear calibration function is the fastest of all four methods.

In case 2, the calibrated weights range in $[0; \infty)$ due to the exponential calibration function $F_k(u) = \exp(u)$. The multiplicative method is also referred to as the raking ratio method, where all calibration variables are qualitative (complete or incomplete post-stratification). The generated weights then correspond to classical raking ration proposed by DEMING AND STEPHAN (1940). For some unlucky samples, the multiplicative method may yield some w_k which are extremely large compared to the design weights d_k (cf. figure 3.1).

The interest of methods 3 and 4 is to avoid negative and extreme large or small weights by providing a lower (L) and an upper (U) limit for g-weights, because extrem weights can affect the robustness of the estimates. The boundaries have to be chosen by the user, e.g. by successive trials in which L is increased towards 1 and U is decreased towards 1 until the optimisation problem is unsolvable. This is a time consuming process. In case 3 it is $F_u(-\infty) = L$, $F_u(\infty) = U$ and $F_u(0) = 1$. The domain of calibrated weights is $[Ld_k, Ud_k]$. The logit method is readily identifiable in figure 3.1; for $u \to +\infty$ ($u \to -\infty$) it converges towards $L = 0.5$ ($U = 2$). The truncated linear method is very similar to the logit method and was chosen for evaluation in the simulation study.

Although the distance functions lead to different calibration weights $\boldsymbol{w}$, DEVILLE/SÄRNDAL (1992) show, that the resulting calibration estimators are asymptotically equivalent to the generalised regression (GREG) estimator having the same asymptotic variance. In case of the linear method $F(u) = 1 + u$, the calibrated weights are given by

$w_k = d_k F(u) = d_k(1 + \mathbf{x}_k^T \boldsymbol{\lambda})$. The Lagrange multiplier $\boldsymbol{\lambda}$ is the solution of the system

$$\sum_{k \in \mathcal{S}} w_k \mathbf{x}_k - \boldsymbol{\tau}_x = \mathbf{X}^T \mathbf{d} + (\mathbf{X}^T \mathbf{D} \mathbf{X}) \boldsymbol{\lambda} - \boldsymbol{\tau}_x = \mathbf{0} \quad .$$

According to Rao (1965) and in the following notation of Deville and Särndal (1992)

$$\boldsymbol{\lambda}^* = -\Big(\mathbf{X}^T \mathbf{D} \mathbf{X}\Big)^- \Big(\mathbf{X}^T \mathbf{d} - \boldsymbol{\tau}_x\Big) = \mathbf{T}^- \Big(\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x\Big)$$

is a solution for generalised inverses $\mathbf{T} = (\mathbf{X}^T \mathbf{D} \mathbf{X})^-$. This solution is exactly $\boldsymbol{\lambda}^{(1)}$ of iterative solving algorithm with vector $\boldsymbol{\Phi}(\boldsymbol{\lambda}^{(0)}) = \mathbf{X}^T \mathbf{d} - \boldsymbol{\tau}_x$ and matrix $\boldsymbol{\Phi}'(\boldsymbol{\lambda}^{(0)}) = \mathbf{X}^T \mathbf{D} \mathbf{X}$. The calibration estimator can be written as

$$\begin{aligned}
\widehat{\boldsymbol{\tau}}_{y,\mathrm{GREG}} &= \widehat{\boldsymbol{\tau}}_{y,\mathrm{HT}} + (\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_{x,\mathrm{HT}})^T \widehat{\boldsymbol{\beta}} \\
&= \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k + \Big(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k\Big)^T \widehat{\boldsymbol{\beta}} \\
&= \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k + \Big(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k\Big)^T \Big(\sum_{k \in \mathcal{S}} d_k \mathbf{x}_k \mathbf{x}_k^T\Big)^{-1} \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k y_k \\
&= \sum_{k \in \mathcal{S}} d_k \mathbf{x}_k + \sum_{k \in \mathcal{S}} d_k \Big(\sum_{\kappa \in U} \mathbf{x}_\kappa - \sum_{\kappa \in \mathcal{S}} d_\kappa \mathbf{x}_\kappa\Big)^T \Big(\sum_{k \in \mathcal{S}} d_k \mathbf{x}_k \mathbf{x}_k^T\Big)^{-1} \mathbf{x}_k y_k \\
&= \sum_{k \in \mathcal{S}} d_k \left[1 + \Big(\sum_{\kappa \in U} \mathbf{x}_\kappa - \sum_{\kappa \in \mathcal{S}} d_\kappa \mathbf{x}_\kappa\Big)^T \Big(\sum_{k \in \mathcal{S}} d_k \mathbf{x}_k \mathbf{x}_k^T\Big)^{-1}\right] y_k \\
&= \sum_{k \in \mathcal{S}} d_k g_k y_k \\
&= \sum_{k \in \mathcal{S}} w_k y_k \quad ,
\end{aligned}$$

where $\boldsymbol{\beta}$ is the sample-based vector of regressors obtained by solving the normal equations. The GREG estimator can be interpreted as a Horvitz-Thompson estimator $\widehat{\boldsymbol{\tau}}_{y,\mathrm{HT}}$ for the variable of interest, adjusted by the difference between the known population totals $\boldsymbol{\tau}_x$ and the corresponding HT estimates $\widehat{\boldsymbol{\tau}}_{x,\mathrm{HT}}$, weighted by the multiple regression coefficient $\boldsymbol{\beta}$.

Deville and Särndal (1992) proof, that $\widehat{\tau}_{y,\mathrm{CAL}}$ is asymptotically equivalent to $\widehat{\boldsymbol{\tau}}_{y,\mathrm{GREG}}$. Thus, the asymptotic varianz (AV) of $\widehat{\tau}_{y,\mathrm{CAL}}$ is (see also Deville et al., 1993)

$$\mathrm{AV}\,(\widehat{\boldsymbol{\tau}}_{y,\mathrm{CAL}}) = \sum_{k \in \mathcal{S}} \sum_{j \in \mathcal{S}} \Big(\pi_{kl} - \pi_k \pi_l\Big)(w_k e_k)(w_l e_l) \quad ,$$

where π_{kl} is the second order inclusion probability, this is $P(k \& l \in \mathcal{S})$, and $e_k = y_k - \mathbf{x}_k^T \boldsymbol{\beta}$ is residual to be calculated from a sample-based weighted linear regression, as $\boldsymbol{\beta}$ is unknown for the population. The WLS estimate for $\mathbf{e} = (e_1, \ldots, e_n)^T$ is given by

$$\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{D} \mathbf{X})^- \mathbf{X}^T \mathbf{D} \mathbf{y} \quad .$$

According to VANDERHOEFT (2003) the asymptotic variance can be written in matrix notation as

$$\widehat{\mathrm{AV}}\left(\widehat{\boldsymbol{\tau}}_{y,\mathrm{CAL}}\right) = \widehat{\mathbf{e}}\mathbf{W}\breve{\Delta}\mathbf{W}\widehat{\mathbf{e}} \quad ,$$

where $\breve{\Delta} = \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right)$.

EU-SILC data is collected at different levels. As stipulated in Subsection 2.3.2, both, household and individual information is collected. All household members aged 16 and over are interviewed. Survey data can be adjusted either in independent calibrations for households and individuals or in an integrative calibration, as recommended by EUROSTAT. The idea of a simultaneous calibration as proposed by SAUTORY (1991) is to ensure consistency in the statistics obtained from the various information levels respectively datasets. It consists in performing a single calibration on household level. The following description is based on ESTEVAO AND SÄRNDAL (2003) and SAUTORY (2003).

Consider a finite population $\mathcal{U}$ of elements $\{1, \ldots, \kappa, \ldots, N\}$ which are clustered into households $\{1, \ldots, \iota, \ldots, M\}$. A two-stage cluster sample is drawn, involving two distinct populations of interest: The population of

1. households $\mathcal{U}_M = \{1, \ldots, \iota, \ldots, M\}$
2. individuals $\mathcal{U}_I = \{1, \ldots, \kappa, \ldots, N\}$.

The population $\mathcal{U}$ is the union of M household units $\mathcal{U}_\iota$ sized N_ι persons, where $\iota \in \mathcal{U}_M$. In addition, the sub-population $\mathcal{U}_K = \{1, \ldots, \kappa', \ldots, N'\}$ of individuals aged 16 and over is investigated.

At first stage, a probability sample $\mathcal{S}_M = \{1, \ldots, i, \ldots, m\}$ of households is drawn from $\mathcal{U}_M$. The first-stage inclusion probability of ith selected household is π_i and the first-stage design weight is $d_i = 1/\pi_i$. Suppose, all members of the selected households are surveyed and form sample $\mathcal{S}_I = \{1, \ldots, k, \ldots, n\}$. According to German EU-SILC, auxiliary information from the GMC is available for all household members. The design weight of any individual $k \in \mathcal{U}_i$ is $d_k = d_i$, when household i is completely observed.

At second stage, a sample of elements $\mathcal{S}_K = \{1, \ldots, k', \ldots, n'\}$ is drawn from $\mathcal{S}_M$, that is, a sample $\mathcal{S}_i$ of size n_i is drawn from each $\mathcal{U}_i$, where $\mathcal{S}_K = \bigcup_{i \in \mathcal{S}_M} \mathcal{S}_i$ and $\mathcal{S}_K \subseteq \mathcal{S}_M \subseteq \mathcal{U}$. The conditioned inclusion probability of element k' is $d_{k'|i} = 1/\pi_{k'|i}$. Its overall design weight is $d_{k'} = 1/(\pi_i \pi_{k'|i}) = d_i d_{k'|i}$. $\mathcal{S}_K$ consist of persons surveyed with the personal questionnaire. These persons can be referred to as Kish individuals. In the simulation study, the following auxiliary information is available:

- $\mathbf{x}_i$: vector of auxiliary information known for each household $i \in \mathcal{S}_M$
- $\boldsymbol{\tau}_x = \sum_{\iota=1}^{M} \mathbf{x}_\iota$: vector of totals known for $\mathcal{U}_M$
- $\mathbf{z}_{i,k}$: vector of auxiliary information known for each individual k in household i

- $\boldsymbol{\tau}_z = \sum_{\kappa=1}^{N} \mathbf{z}_\kappa$: vector of totals known for $\mathcal{U}_I$
- $\mathbf{v}_{i,k'}$: vector of auxiliary information known for each Kish individual k' in household i
- $\boldsymbol{\tau}_v = \sum_{\kappa'=1}^{N} \mathbf{v}_{\kappa'}$: vector of totals known for $\mathcal{U}_K$

As mentioned above, the objective is to perform a single calibration on household level. Thus, for each household $i \in \mathcal{S}_M$ the vector of total values of calibration variables for individuals k is calculated: $\boldsymbol{z}_i = \sum_{\kappa \in \mathcal{U}_i} \mathbf{z}_{i,k}$. Further, the vector of totals of calibration variables is estimated for Kish individuals k' in each household i: $\widehat{\boldsymbol{v}}_i = \sum_{\kappa' \in \mathcal{U}_i} d_{k'} \mathbf{v}_{i,k'}$. The known vector of calibration variables for household i is then $(\mathbf{x}_i, \mathbf{z}_i, \widehat{\boldsymbol{v}}_i)$. According to Sautory (2003) the calibration equation can be written as:

$$\boldsymbol{\Phi}(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\gamma}) = \sum_{i \in \mathcal{S}_M} d_i F\Big(\mathbf{x}_i^T \boldsymbol{\lambda} + \mathbf{z}_i^T \boldsymbol{\mu} + \widehat{\boldsymbol{v}}_i^T \boldsymbol{\gamma}\Big)\Big(\mathbf{x}_i, \mathbf{z}_i, \widehat{\boldsymbol{v}}_i\Big) - \Big(\boldsymbol{\tau}_x \quad \boldsymbol{\tau}_z \quad \boldsymbol{\tau}_v\Big) = \mathbf{0} \quad .$$

Calculation of the calibration weights

$$w_i = d_i F(\mathbf{x}_i^T \boldsymbol{\lambda} + \mathbf{z}_i^T \boldsymbol{\mu} + \widehat{\boldsymbol{v}}_i^T \boldsymbol{\gamma})$$

requires to determine the components $\boldsymbol{\lambda}, \boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ of the vector of Lagrange multipliers. The calibrated weight assigned to person k of household i is $w_{i,k} = w_i$ and for Kish individual $k' \in \mathcal{U}_i$ it is $w_{k'} = w_i d_{k'}$. Sautory (2003) verifies that the samples are correctly calibrated on the population totals. It is

$$\begin{aligned} \sum_{k \in \mathcal{S}_I} w_{i,k} \mathbf{z}_{i,k} &= \sum_{i \in \mathcal{S}_M} w_i \Big(\sum_{k \in \mathcal{U}_i} \mathbf{z}_{i,k} \Big) = \sum_{i \in \mathcal{S}_M} w_i \boldsymbol{z}_i = \boldsymbol{\tau}_z \quad \text{and} \\ \sum_{k' \in \mathcal{S}_K} w_{k'} \mathbf{v}_{i,k'} &= \sum_{k' \in \mathcal{S}_K} w_i d_i \mathbf{v}_{i,k'} = \sum_{k' \in \mathcal{S}_K} w_i \widehat{\mathbf{v}}_i = \boldsymbol{\tau}_v \quad . \end{aligned}$$

Within calibration routines further weighting schemes could be implemented easily. On example is the original German EU-SILC which has to consider propensities for the Access Panel which is behind EU-SILC. This however is not followed within the given study.

Assuming existence of second order inclusion probabilities, the variance of the HT estimator is given by

$$\mathrm{V}\left(\widehat{\tau}_{y,\mathrm{HT}}\right) = \sum_{i \in \mathcal{U}} \pi_i (1 - \pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \cdot \sum_{\substack{i,j \in \mathcal{U} \\ i<j}} \sum (\pi_{ij} - \pi_i \cdot \pi_j) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} \quad .$$

An unbiased estimate is (cf. Särndal et al. 1992, page 43) is

$$\mathrm{V}\left(\widehat{\tau}_{y,\mathrm{HT}}\right) = \sum_{i \in \mathcal{R}} (1 - \pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \cdot \sum_{\substack{i,j \in \mathcal{R} \\ i<j}} \sum \left(1 - \frac{\pi_i \cdot \pi_j}{\pi_{ij}}\right) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} \quad .$$

In case of a stratified random sampling with H strata ($\{1, \ldots, h, \ldots, H\}$) the variance of the HT estimator is given by

$$\mathrm{V}\left(\tau_{y,\mathrm{HT}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{\sigma_h^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1}$$

which can be estimated adequately by

$$\widehat{\mathrm{V}}\left(\widehat{\tau}_{y,\mathrm{HT}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{s_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \quad .$$

Stratified random sampling will be used throughout this study.

3.3 Frame of the simulation study

The variance estimators set out in Section 3.1 are applied to the semi-synthetic dataset and evaluated. The CIoILC set out in Sections 2.1 and 2.3 is estimated. Calculations are performed using the statistical programme package R.

In the following, the auxiliary variables used in the calibration estimates as regressor matrices are introduced. Further, the sample designs by means of which the EU-SILC sample is generated are explained. The statistical measures used to assess the estimations (accuracy measurement) are also explained. Further, pointers are given on the lay-out and interpretation of the graphics setting out our results.

3.3.1 Estimators

As stated in Section 3.1, DEVILLE et al. (1993) recommends using four distance or calibration functions. DEVILLE AND SÄRNDAL (1992) show that all calibration estimators yield asymptotically equivalent results. The results achieved within the framework of the simulation study for both the linear and the truncated linear method (and which are illustrated below) furnish proof of this asymptotic equivalence. Both the multiplicative and the Logit method yield similarly equivalent results with (invariably) non-negative weight vectors. The results were not additionally shown in the figures, in order to make way for more clearly represented and more detailed figures. The figures illustrate the results for the following estimators:

- Horvitz-Thompson estimator (HT)
- Calibration estimator with linear calibration function (GREG)
- Calibration estimator with truncated linear calibration function (CAL).

The GREG estimator can generate weights vectors with negative elements, frequently unwanted in official statistics. The CAL estimator yields non-negative weights in all cases. It is based, contrary to the GREG estimator, on an iterative calculation method, which usually involves comparatively long calculation times.

3.3.2 Auxiliary variable sets

Within the compass of the simulation study, six different auxiliary variable sets were tested in order to estimate the composite indicator. The auxiliary variable sets differ primarily in the number of regressors. Tables A.5 to A.8 illustrate the auxiliary variable sets.

`UT1`: 56 regressors. All auxiliary variables selected within the compass of the simulation study are included. The auxiliary variable set contains the greatest number of regressors and a maximum amount of auxiliary information. In order to avoid multi-collinearity, for each variable at least one category of characteristics is not included as a regressor.

`UT2`: 48 regressors. As in auxiliary variable set `UT1`, but minus the variables EF539 (income) and EF215 (collecting unemployment benefits). Many of the generated EU-SILC and/or ESS variables are strongly correlated with income. In addition, the income characteristic is a stratification variable used in surveying EU-SILC. Variable EF215 is entered into the work indicator. `UT2` studies the extent to which the estimators forfeit precision over and against `UT1`.

`R1`: 21 regressors. Using the R-routine `regsubsets`, an auxiliary variable set of 21 regressors is selected. The aim is to study whether, and if so by how much, the estimation loses precision over and against `UT1` and `UT2`. Estimations of the composite indicator on the basis of the underlying population have shown, that when more than 21 regressors are involved what we get is only a relatively small change in the statistical measures (e.g. in the adjusted R squared or the AIC criterion). The issue is pursued of whether this auxiliary variable set comprising only 21 selected regressors suffices for a precise projection of the EU-SILC samples.

`R2`: As in `R1`, but without the variables EF539 (income) and EF215 (collecting unemployment benefits). Compare the motivation here with `UT2`. `NSI`: 34 regressors. Used here is the auxiliary variable set proposed by DESTATIS. Over and against UT, the following information is not contained: EF110 (employment), EF130 (public sector employment), EF138 (part-time or full-time work), EF215 (collecting unemployment benefits), EF259 and EF261 (education level), EF338, EF358 and EF359 (state-provided social benefits)]. Another difference is that income groups are interpreted more broadly (the hne variable replaces EF539).

`EC`: 39 regressors. In its EU-SILC report, Description of Target Variables, page 36, the EC proposes the following auxiliary variable set: “Recommended calibration variables at the household level are household size, tenure status and region (NUTS 2) and, at the personal level, distribution of population by age (five years age group) and gender.” Since the tenure variable tenure is not available to the GMC, the auxiliary variable set implemented in the simulation study deviates from this scenario. Furthermore, only NUTS 1 information can be used.

The EU-SILC study in Germany is designed to draw on fairly comprehensive information by cross-linking with the GMC. A point to note is that the EU’s recommendation is aimed at member states with only scant register information available for estimation purposes. One of the aims of the simulation study is to make abundantly clear the value of having additional auxiliary information on call.

Estimation of the composite indicator is performed in the simulation study as follows:

- The single indicators are uniformly estimated using, in each case, one of the six above-mentioned sets of auxiliary variables, while the composite indicator is calculated as a linear combination of the single indicators.
- Alternatively, the single indicators are estimated using, in each case, an own auxiliary variable set.

As a rule of thumb, it can be assumed that the single indicators aggregated to form a composite indicator will stem from different surveys. Accordingly, the projection of each and every single indicator can be performed in different ways. This circumstance should be taken into account in respect of the auxiliary variable set. Using the R-routine `regsubsets` for each single indicator a special auxiliary variable set is determined, into which are entered, in each case, 10 of the total of 56 available regressors (48 if we exclude income and unemployment). The small number of, in each case, 10 regressors is chiefly to be explained by the fact that there may very well be, in practice, few auxiliary variables available. Further, it should again be noted that the statistical dimensions, as soon as we get above 10 regressors, only change marginally when the single indicators are estimated on the basis of the underlying population.

3.3.3 Sampling procedures

DESTATIS sampling design (`NSI`): The EU-SILC sample from 2005 consists of a 25% random sample and a 75% quota sample. In total, 14,100 are drawn from the population, 4,100 by random sampling and 10,000 by quota sampling.

The EC foresees for Germany a minimal sample size of 8,250 households. The German Statistics Office usually draws on a pool of 14,100 households, which is approximately 70% more than called for.

Planning by DESTATIS foresees splitting the sample size disproportionally for the characteristics of household type and social position and proportionally for the characteristic of net household income. Rounding out the sample sizes for the preceding characteristics is done using the Niemeyer method. The sample size for each of Germany's states is divided in terms of the individual stratification characteristics using Deming's iterative proportional fitting method.

Unbiased estimation requires that at least one element per stratum is sampled. The cell (stratum) specific inflation factor depends on the sampling fraction. Sparse cells imply extreme weights that may, in turn, influence the quality of the estimated values. The stratification plan DESTATIS uses was modified to let each cell include at least two households. It should be remembered that only a 70% sampling of the GMC (Scientific Use File) is available. This requires collapsing the cells together. Income classes and households (in that order) are lumped according to their social backgrounds. This results in a total of 581 strata, distributed among the 16 German states.

Simple Random Sampling with allocation proportional to size (StratRS): A stratified random sample running to some 14,100 households (`StratRS1`) and 8,250 households (`StratRS2`) is taken from the population. To ensure comparability of sample relationships, post-stratification as per DESTATIS (cf. `NSI`-design) is taken as a basis.

Simple Random Sampling (SRS): Drawn from the population are 14,100 households (`SRS1`) and 8,250 households (`SRS2`) without replacement.

3.3.4 Accuracy measures

The aim of this section is to present the accuracy measures applied in the simulation study in order to compare different point and variance estimators. The selection of measures and is based on Münnich et al. (2004).

One major measure of interest is the mean square error (MSE) of an estimator $\widehat{\theta}$ for the population parameter θ which is defined as

$$\mathrm{MSE}\,(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}-\theta)^2 = \mathrm{V}\,(\widehat{\theta}) + \Big(\mathrm{E}(\widehat{\theta})-\theta\Big)^2 = \mathrm{V}\,(\widehat{\theta}) + \mathrm{B}^2 \quad , \tag{3.7}$$

where $\Big(\mathrm{E}(\widehat{\theta})-\theta\Big)$ denotes the bias B of an estimator $\widehat{\theta}$.

According to Särndal et al. (1992), the MSE will appropriately compare between several different estimators $\widehat{\theta}_1, \widehat{\theta}_2, \ldots$ for one and the same parameter θ. From equation 3.7 follows, that large biases or variances will considerably influence the MSE. However, if the MSE of an estimator $\widehat{\theta}$ is small, the estimated value is likely to be close to the true value. Within the simulation study, the MSE is approximated by

$$\mathrm{MSE}\,(\widehat{\theta}) \doteq \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \frac{1}{R}\sum_{i=1}^{R}\widehat{\theta}_i\right)^2 + \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r - \theta\right)^2 \quad ,$$

where $\widehat{\theta}_r$ is the r-th estimate from R runs in the simulation.

The measures mentioned above can be transformed to relative measures with respect to the true value θ or its expected estimate. The coefficient of variation to estimators, is defined as

$$\mathrm{CV} = \frac{\sqrt{\mathrm{V}\,(\widehat{\theta})}}{\mathrm{E}(\widehat{\theta})} \quad ,$$

respectively

$$\mathrm{CV}^* := \frac{\sqrt{\mathrm{MSE}\,(\widehat{\theta})}}{\mathrm{E}(\widehat{\theta})} = \frac{\mathrm{RMSE}\,(\widehat{\theta})}{\mathrm{E}(\widehat{\theta})} \tag{3.8}$$

for biased estimators, where RMSE denotes the root mean square error. CV^* (cf. equation 3.8) is also called relative root mean square error (RRMSE). In case θ is an unbiased estimator CV is equal to CV^*. In the simulation study the coefficient of variation is estimated via

$$\widehat{\mathrm{CV}}^* = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r\right)^2 + \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r - \theta\right)^2}}{\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r} \quad .$$

Assuming that an estimator is at least approximately unbiased and normal, the $(1-\alpha)\cdot 100\%$ confidence intervals can be derived using

$$\left[\widehat{\theta} - z_{1-\alpha/2}\cdot\sqrt{\mathrm{V}\left(\widehat{\theta}\right)}\,;\,\widehat{\theta} + z_{1-\alpha/2}\cdot\sqrt{\mathrm{V}\left(\widehat{\theta}\right)}\right] \quad ,$$

where $z_{\cdot}$ denotes the normal distribution and α denotes the level of significance. In case of biased estimators, the standard deviation of the estimator can be substituted by the MSE which yields to

$$\mathrm{CI} = \left[\widehat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{MSE}\left(\widehat{\theta}\right)};\widehat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{MSE}\left(\widehat{\theta}\right)}\right] \quad ,$$

where the MSE is approximated by equation 3.7. As a measure of accuracy, the coverage rates of $(1-\alpha)\cdot 100\%$ confidence intervals can be evaluated within a simulation study. The theoretical value is $(1-\alpha)\cdot 100\%$ where lower rates are much more severe.

3.3.5 Figures

Qualitative analysis is based on four figures:

- boxplots of point estimates,
- boxplots of variance estimates
- a bar chart for the relative root mean squared error (RRMSE) and for the mean and the estimated variation coefficients (MeanCV and MeanEstCV); the measures are colour-coded as follows: RRMSE (red) MeanCV (blue) and MeanEstCV (black)
- bar chart of 90% and 95% coverage rates.

The point and variance estimators are evaluated after Horvitz-Thompson (HT) and after the calibration method (GREG, CAL) for six auxiliary variable sets (`UT1`, `UT2`, `R1`, `R2`, `NSI`, `EC`) and for five designs (`NSI`, `StratRS1`, `StratRS2`, `SRS1`, `SRS2`). This yields a total of 90 different estimates for the composite indicator.

In each case, 10,000 independent samples are taken. HT, GREG and CAL are estimated for each of the 30 different combinations of designs and auxiliary variable sets. The point and variance estimators are visualised by means of boxplots.

Accordingly, the evaluation figures 3.2 to 3.5 are designed as follows: For each estimator (HT, GREG, CAL) 30 boxplots are plotted line-by-line. Columns are sub-divided block-wise in terms of auxiliary variable (AV) sets (AV `NSI`, AV `UT1`, AV `UT2`, AV `R1`, AV `R2`, and AV `EC`). For each auxiliary variable set, the estimates for the five sample designs (`NSI`, `StratRS1`, `SRS1`, `StratRS2`, and `SRS2`) are plotted line by line. Here the samples in which 14,100 households were evaluated (`NSI`, `StratRS1`, `SRS1`) are ranged before those involving 8,250 households (`StratRS2`, `SRS2`).

To prevent this report from running to too many pages, the figures analysing the single indicators have been reduced in size and arrayed as follows:

distribution of point estimates	distribution of variance estimates
RRMSE, MeanCV, MeanEstCV	coverage rates

Table 3.2: Array of figures

3.4 Results of the composite indicator

From the perspective of sampling theory, some preliminary points are called for:

- Stratified sampling, provided the population can be divided into homogeneous subgroups by means of relevant stratification variables , generally improves the accuracy of estimation yielding results at least as good as those achieved by simple random sampling.
- The greater the sample size, the more precise the estimates will be. The effects are visible in the figures. The variance of the point estimators, which is elicited on the basis of `StratRS2`, `SRS2` is greater compared with `StratRS1` and `SRS1`. This applies for all three estimators (HT, GREG, CAL). The figures confirm these theoretical considerations.

The properties we wish our estimates to have are as follows: Unbiasedness, efficiency and consistency.

In the Monte Carlo simulation unbiasedness is deemed to be obtained if it is the case that, in the boxplot, the true value (true value of the composite indicator, red) corresponds to the mean estimate of the 10,000 replications (blue), i.e. the blue and red vertical lines overlap. For the boxplots of the variance estimators the following holds: the blue and red lines should overlap, i.e. the mean of the variance estimators should correspond to the variance of the point estimator. The lower the mean level of variance, the more precise the estimate is. Another desiderandum is that the variance estimators should not scatter over a wider range.

The third figure shows the values of RRMSE, the true, and the estimated average coefficient of variation, which should correspond. The interplay of point and variance estimates can be clearly gauged from 90% and 95% coverage rates (figure 4). It is important that these percentages be achieved as close as possible.

In the following, the four figures (cf. Figures 3.2 to 3.5) are analysed in more detail.

The results of HT vary with the design and sample size. HT does not use any auxiliary information. The group, consisting of 5 boxplots (each for one sampling design), posts identical results for HT across all auxiliary variable sets.

According to figure 3.2, the HT estimator leads to unbiased estimates for all designs. As was to be expected, the point estimators are widely scattered, especially when simple random sampling (SRS) is used. The difference between `StratRS1` and `StratRS2` as well as between `SRS1` and `SRS2` is solely due to the sample size (14,100 HH and 8,250).

The HT gains in precision when a suitable post-stratification is selected to estimate the composite indicator. Disproportional allocation of the sample size (NSI) has advantages over and against proportional allocation (StratRS), resulting in more precise estimates (smaller variance).

The precision of the GREG estimator depends primarily on the auxiliary information entered. All point estimates are asymptotically unbiased. The point estimators are less widely scattered, owing to the auxiliary information entered, than in HT. In the case of simple random sampling (SRS1, SRS2), inclusion of auxiliary information reduces variance considerably. Combining the auxiliary variable set UT1 and sample design StratRS turns out to yield, within the simulation study, the greatest precision in estimating the composite indicator. Point estimators are most widely scattered when auxiliary variable sets R1, R2, and EC are used, in particular when sampling 8.250 households. GREG and CAL (columns 2 and 3) yield equivalent point and variance estimates. Deviations only occur with the weights vector, which in CAL is invariably non-negative.

The variance estimators (cf. figure 3.3) show the differential precision of point estimations. What stands out in HT is the high variance of the point estimators achieved when simple random sampling (SRS) is used. HT tends, in comparison with GREG and CAL, to slightly over-estimate (SRS2) or under-estimate (SRS1) the true variance of the estimator. The percentages in the figure give the ratio between the estimated and the true variance (quotient of the mean of the variance estimates and the mean variance of the point estimates).

Compared with HT, the calibration estimators yield far more precise estimations, as is to be expected. Table 3.3 shows the variance ratios between HT and GREG. Of especial note here is that the performance improves under SRS: when the auxiliary variable sets UT1, UT2 and NSI are used, the variance is less than 5% of what it is for HT. The advantageousness of the NSI design, over and against SRS, varies with the explanatory content of the auxiliary variable set.

The extent to which the auxiliary variable sets affect the precision of estimations for a selected sampling design is set out in Figures 3.6 to 3.9. Over and against what was depicted previously, the columns are now sub-divided according to the five designs. For each design (colour-coded boxes) the six different auxiliary variable sets (AV NSI, AV UT1, AV UT2, AV R1, AV R2, and AV EC) are recorded line by line. Column 1 again makes clear that HT does not incorporate any auxiliary information. The estimator results obtained for a selected design are the same for all auxiliary variable sets (lines). Columns 2 and 3 show for the calibration estimators the differential efficacy of the auxiliary variable sets for a selected sampling design.

The auxiliary variable sets R1 and R2 show that, compared with HT estimation, even including a few available auxiliary variables yields greater precision. This is a reason to procure and incorporate population-related data. However, the explanatory content of the auxiliary variables should be sufficiently large.

The R2 auxiliary variable set yields the lowest precision gains, followed by R1 and EC. Concerning auxiliary variable set EC, the combined estimation based on the characteristics of age (five-year intervals) and gender performs well when using stratified sampling with proportional allocation (StratRS1). Disproportional allocation of sample size (NSI)

Auxiliary variable set	Sampling design	Mean variance estimates (*100,000)				HT est / HT true	GREG est / GREG true	Ratio GREG/ HT
		HT' est	HT true	GREG est	GREG true			
NSI	NSI	4.867	4.937	1.878	1.845	98,6%	101,8%	2,68
	StratRS1	5.678	5.517	1.689	1.689	102,9%	100,0%	3,27
	SRS1	40.052	41.138	1.852	1.838	97,4%	100,8%	22,38
	StratRS2	10.001	10.120	2.978	2.966	98,8%	100,4%	3,41
	SRS2	69.856	67.454	3.228	3.199	103,6%	100,9%	21,09
UT1	NSI	4.867	4.937	1.507	1.495	98,6%	100,8%	3,30
	StratRS1	5.678	5.517	1.366	1.365	102,9%	100,1%	4,04
	SRS1	40.052	41.138	1.469	1.469	97,4%	100,0%	28,01
	StratRS2	10.001	10.120	2.408	2.424	98,8%	99,3%	4,17
	SRS2	69.856	67.454	2.559	2.523	103,6%	101,5%	26,74
UT2	NSI	4.867	4.937	1.644	1.636	98,6%	100,5%	3,02
	StratRS1	5.678	5.517	1.519	1.504	102,9%	101,0%	3,67
	SRS1	40.052	41.138	1.748	1.755	97,4%	99,6%	23,43
	StratRS2	10.001	10.120	2.687	2.678	98,8%	100,3%	3,78
	SRS2	69.856	67.454	3.046	3.048	103,6%	99,9%	22,13
R1	NSI	4.867	4.937	4.077	4.127	98,6%	98,8%	1,20
	StratRS1	5.678	5.517	3.545	3.529	102,9%	100,4%	1,56
	SRS1	40.052	41.138	4.493	4.549	97,4%	98,8%	9,04
	StratRS2	10.001	10.120	6.245	6.202	98,8%	100,7%	1,63
	SRS2	69.856	67.454	7.833	7.860	103,6%	99,7%	8,58
R2	NSI	4.867	4.937	3.102	3.014	98,6%	102,9%	1,64
	StratRS1	5.678	5.517	2.885	2.842	102,9%	101,5%	1,94
	SRS1	40.052	41.138	3.642	3.788	97,4%	96,2%	10,86
	StratRS2	10.001	10.120	5.077	5.080	98,8%	99,9%	1,99
	SRS2	69.856	67.454	6.360	6.360	103,6%	100,0%	10,61
EC	NSI	4.867	4.937	2.280	2.100	98,6%	108,5%	2,35
	StratRS1	5.678	5.517	2.008	2.033	102,9%	98,8%	2,71
	SRS1	40.052	41.138	3.487	3.578	97,4%	97,5%	11,50
	StratRS2	10.001	10.120	3.599	3.542	98,8%	101,6%	2,86
	SRS2	69.856	67.454	6.078	6.170	103,6%	98,5%	10,93

Table 3.3: Mean estimated variances according to figure 3.3

yields an over-estimation of the true variance (+ 9%). Combined estimation allows for a gain in precision, compared to use of total values corresponding to one-dimensional frequencies. The auxiliary variable sets `R1`, `R2`, and `EC`, which comprise less information, supply significantly less precise projections for the composite indicator (which is in line with expectation), in particular when 8.250 households are sampled, and in combination with simple random sampling (`SRS`).

The values of the accuracy measures reflect the differential variation of the point estimators. For all estimators as well as combinations of the auxiliary variable set and the design, the measures correspond down to only small deviations. 90% and 95% coverage rates are almost achieved by almost all estimators. Excessive or insufficient coverage is rarely encountered. For HT estimators, absolute deviations ranging from +0.33% to -0.34% are found; for calibration estimators, the figures vary between +0.40% and -0.75%. As for the `NSI` design, the best coverage rates obtain for HT estimators. The - on the whole - very good coverage rates, even for SRS, cannot hide the fact that the lengths of the confidence intervals diverge as a result of differentially large standard deviations.

In the case of the calibration estimators, we find a preference for the `StratRS1` design. In sum, it can be stated that the precision of estimators depends on the design selected, the sample size and - in the case of calibration estimators - the auxiliary variable set. HT is design-sensitive, the calibration estimators are primarily sensitive in selecting the auxiliary variable set. GREG and CAL estimators (columns 2 and 3) yield asymptotically equivalent results and confirm theoretical expectations.

The GREG estimator only rarely yields weights vectors with negative entries (cf. 3.10). These few cases (at most 18 of 10.000 samples) are samples in which negative weights occur when the auxiliary variable sets `UT1` and `UT2` are used. Such negative weights may stem from the large number of auxiliary variables entered. Each auxiliary variable is a constraint for resolving the linear programming problem. The number of negative weights from over 10,000 replications of the Monte-Carlo study is below 0.02%. Since the GREG estimators

- yield asymptotically equivalent results, and
- only in few cases do negative weights occur,

CAL will not, in the following treatment, be additionally included in the figures. The calibration estimators are, as a result of adequate auxiliary variable sets, superior to the HT estimators (which is as expected). The calibration estimators yield more precise estimates, i.e. less scatter in the point and variance estimators as well as a lower level of variance. Correspondingly smaller are the expressions of RRMSE, i.e. the true and the estimated average coefficient of variation. But this presupposes the availability of up-to-date, population-related total values correlating with the composite indicator.

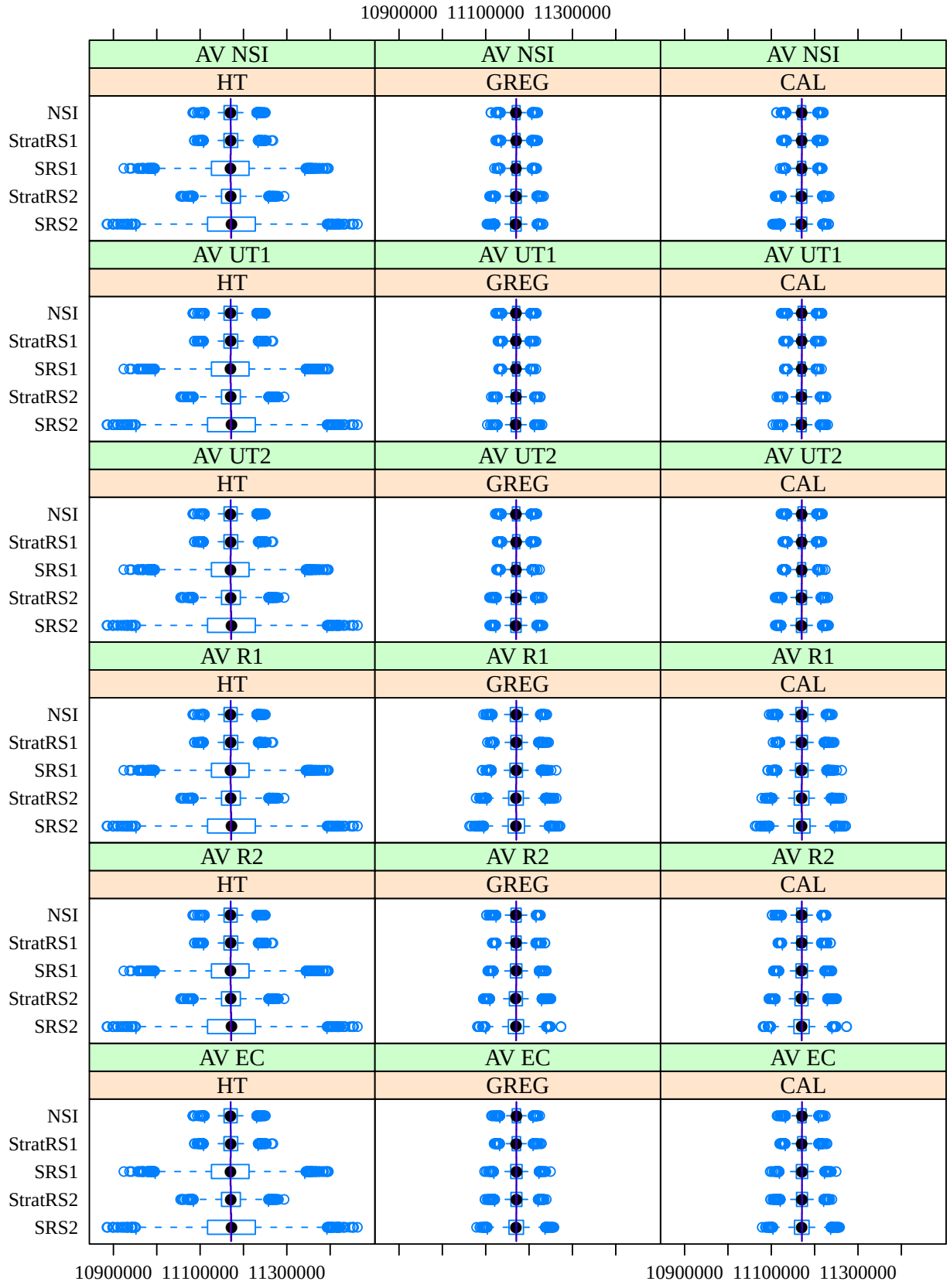

Figure 3.2: Point estimators for composite indicator

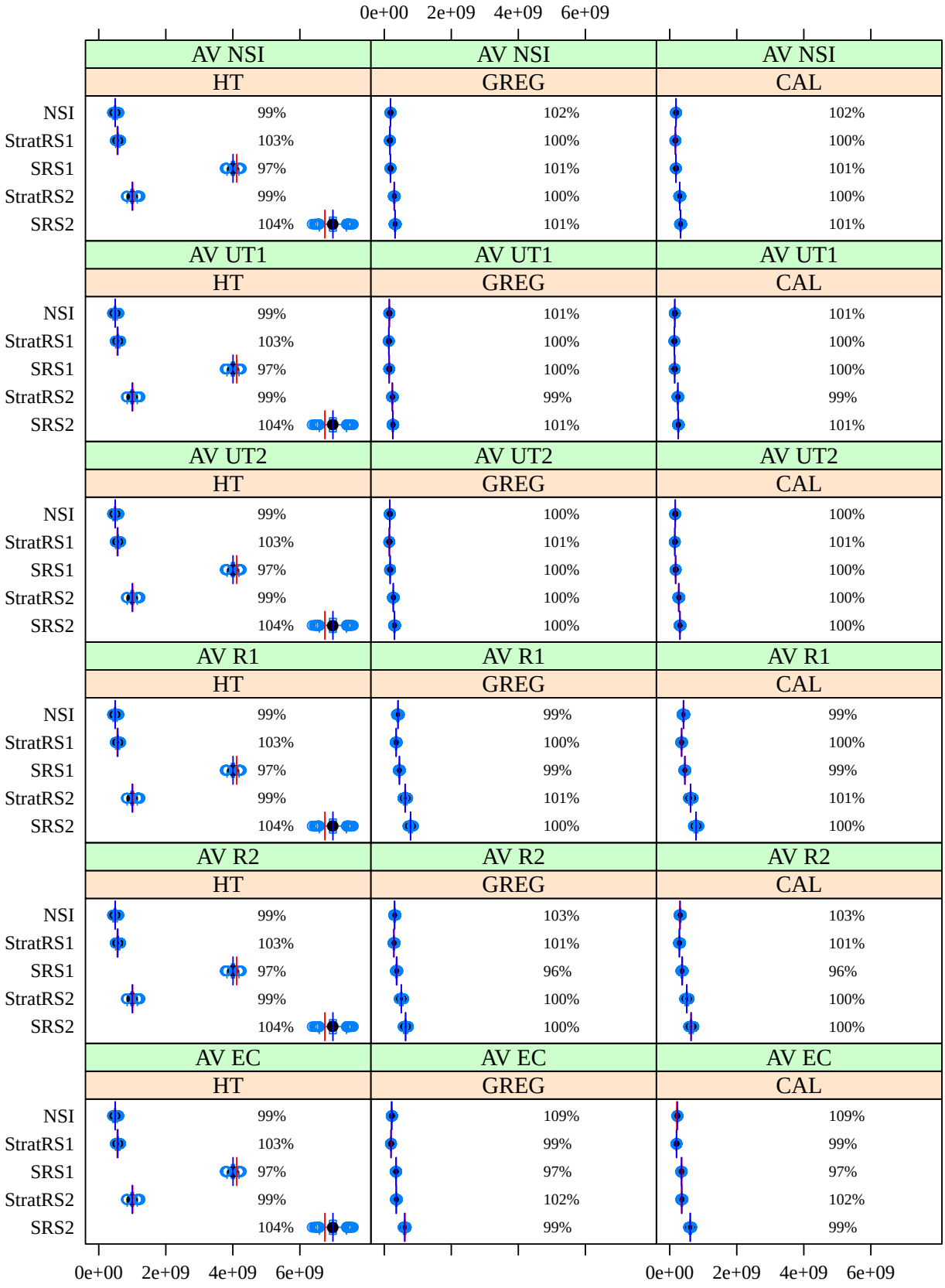

Figure 3.3: Variance estimators for composite indicator

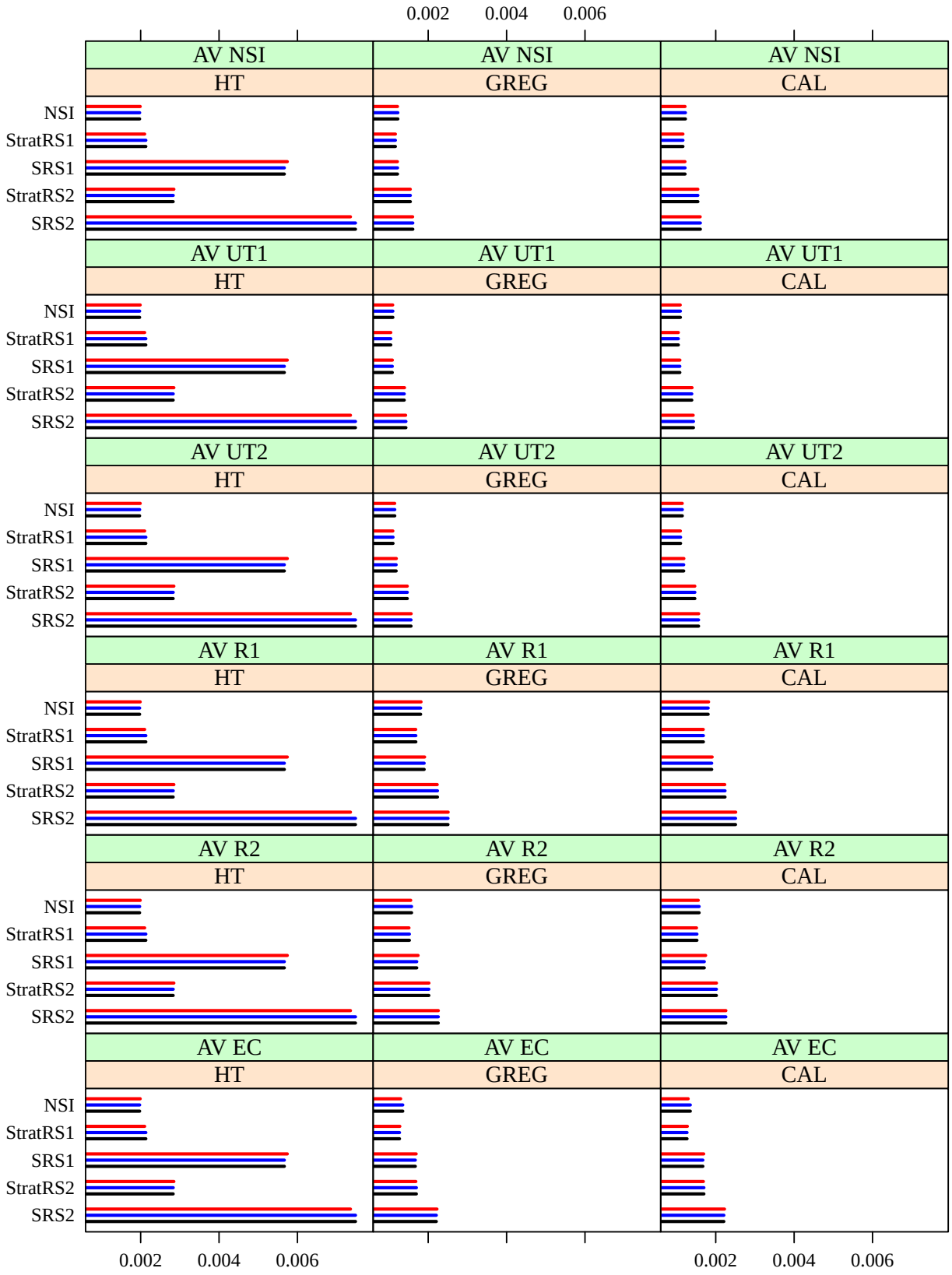

Figure 3.4: Statistical measures (RelRootMSE, MeanCV, and MeanEstCV) for CI

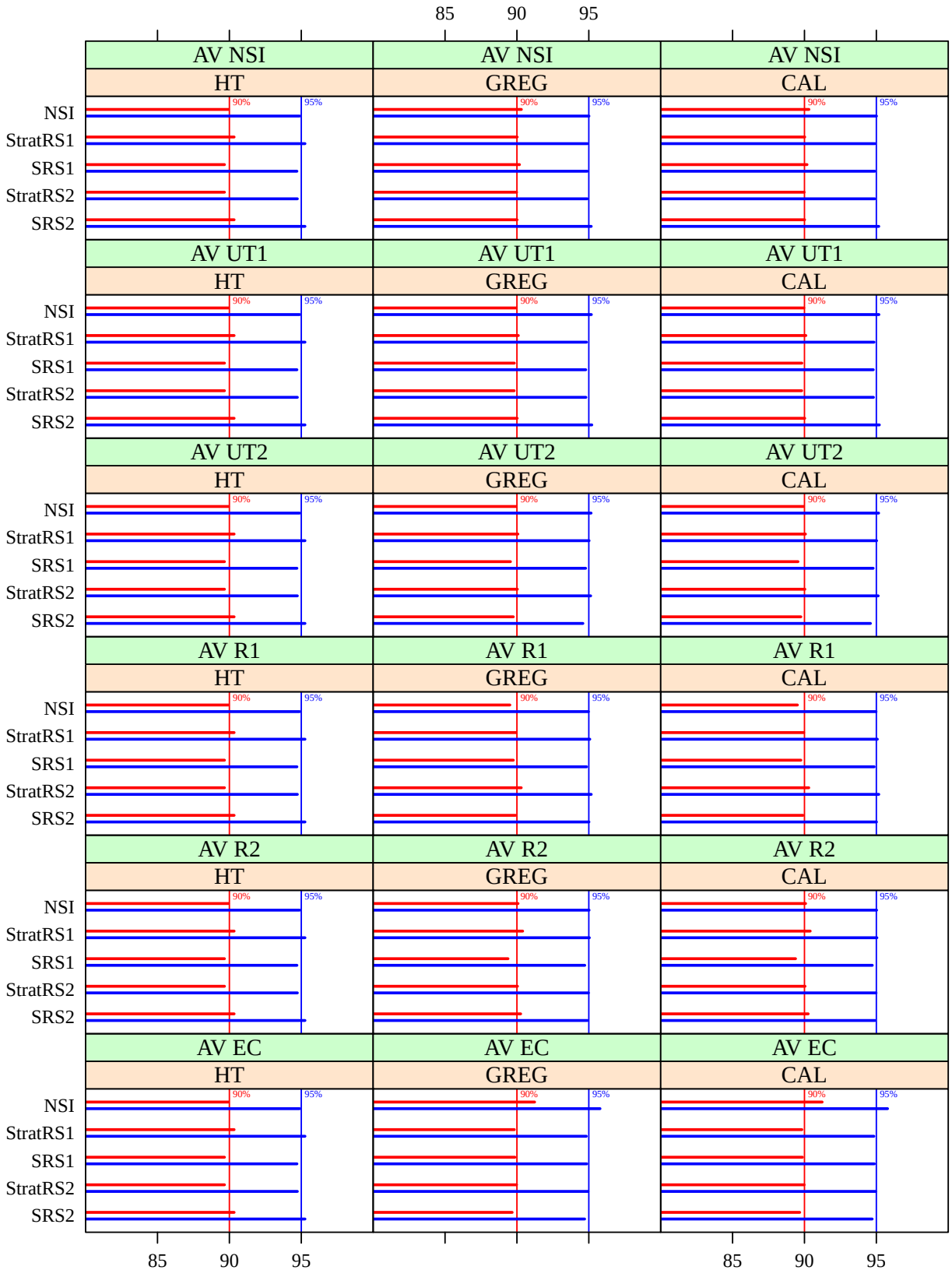

Figure 3.5: 90% and 95% confidence interval coverage rate for CI

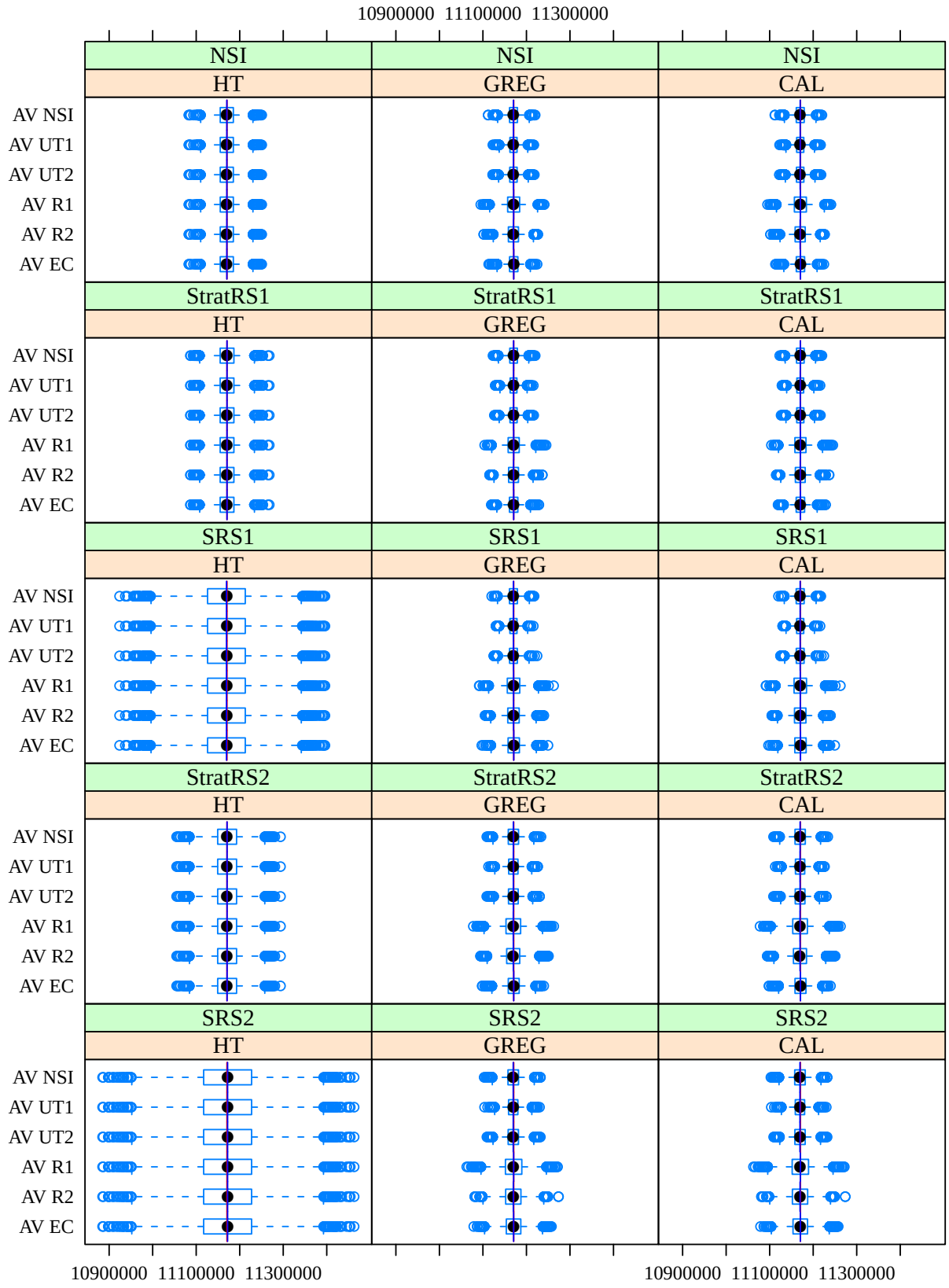

Figure 3.6: Point estimators for composite indicator

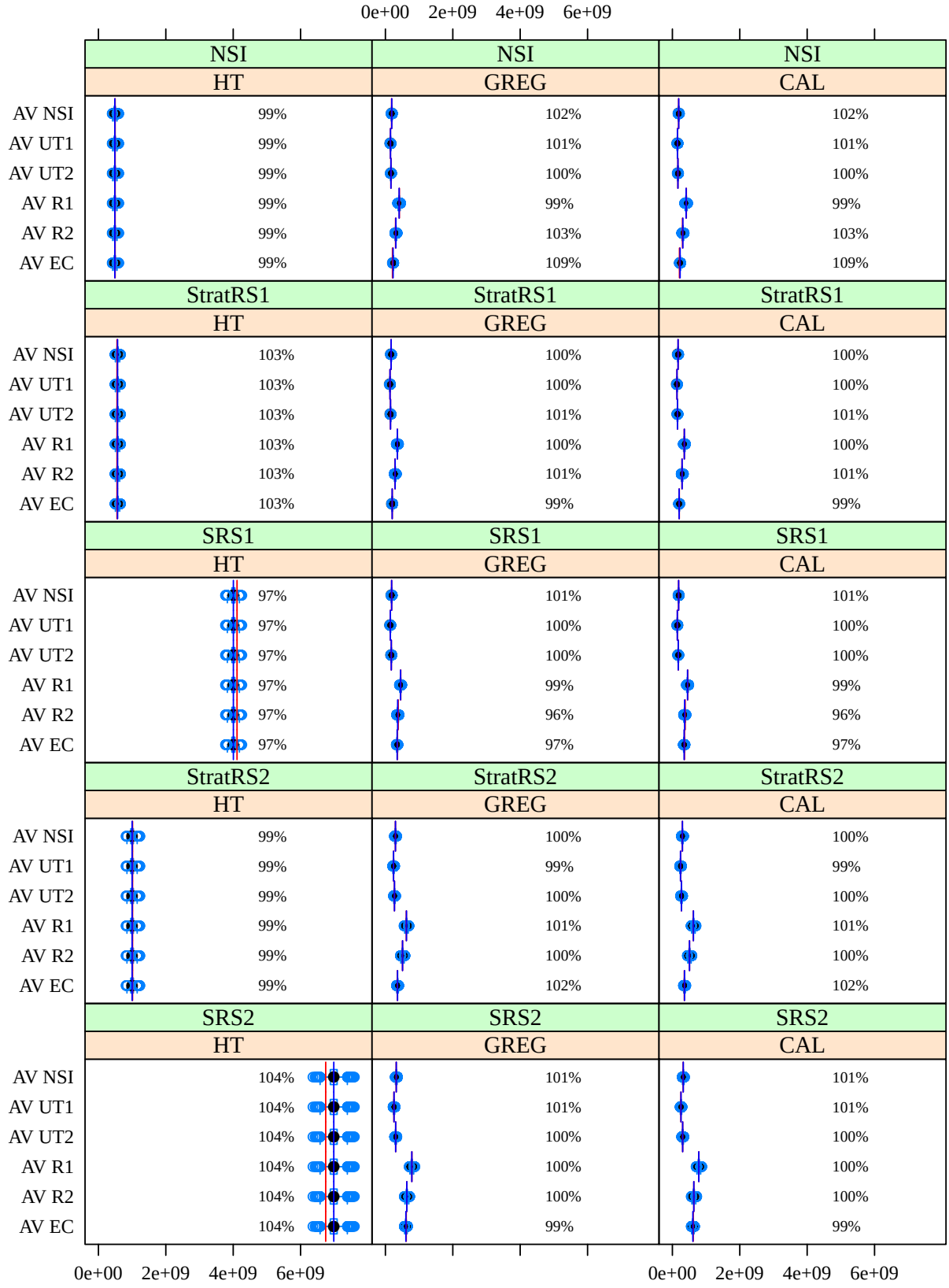

Figure 3.7: Variance estimators for composite indicator

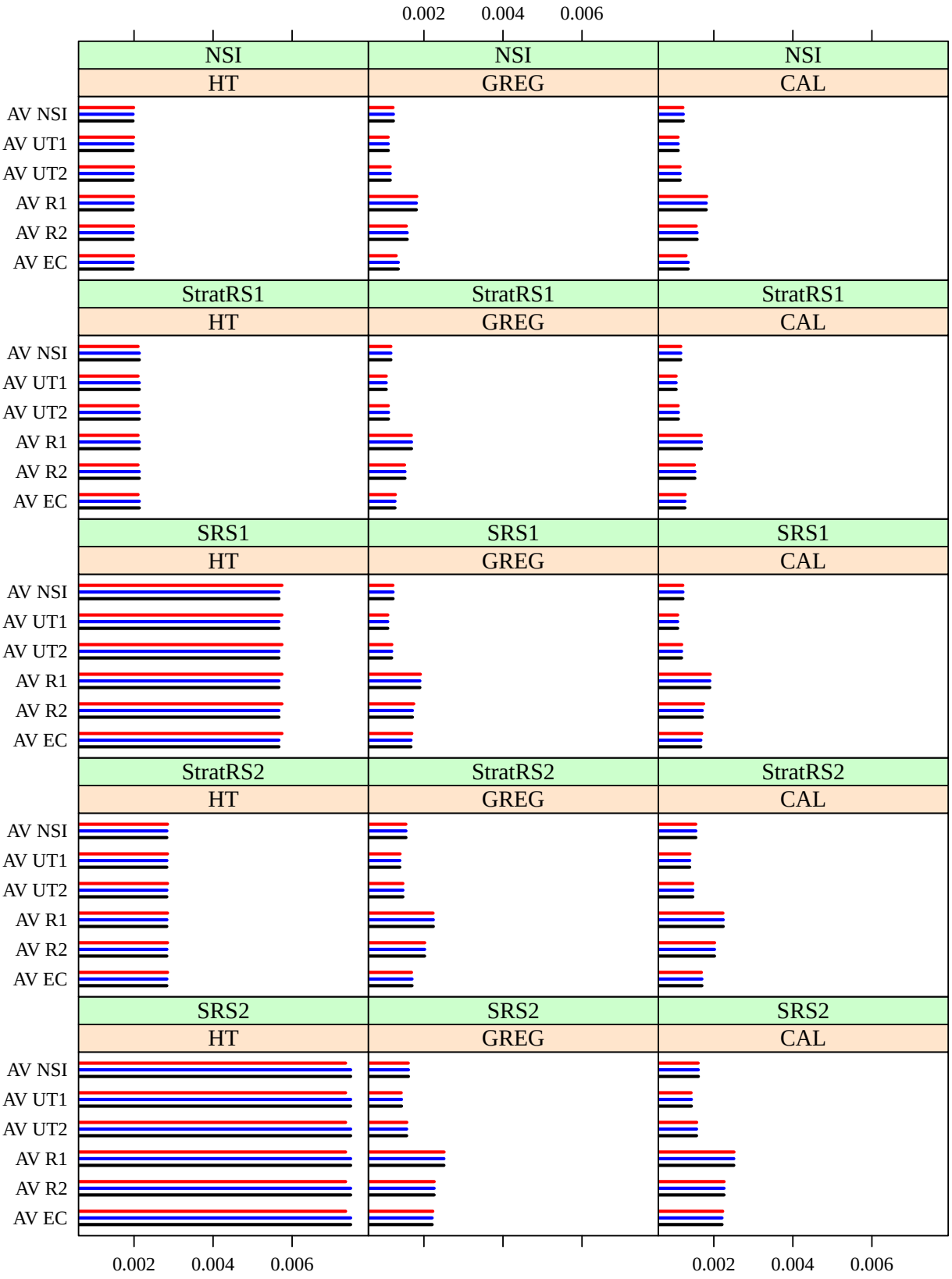

Figure 3.8: Statistical measures (RelRootMSE, MeanCV, and MeanEstCV) of CI

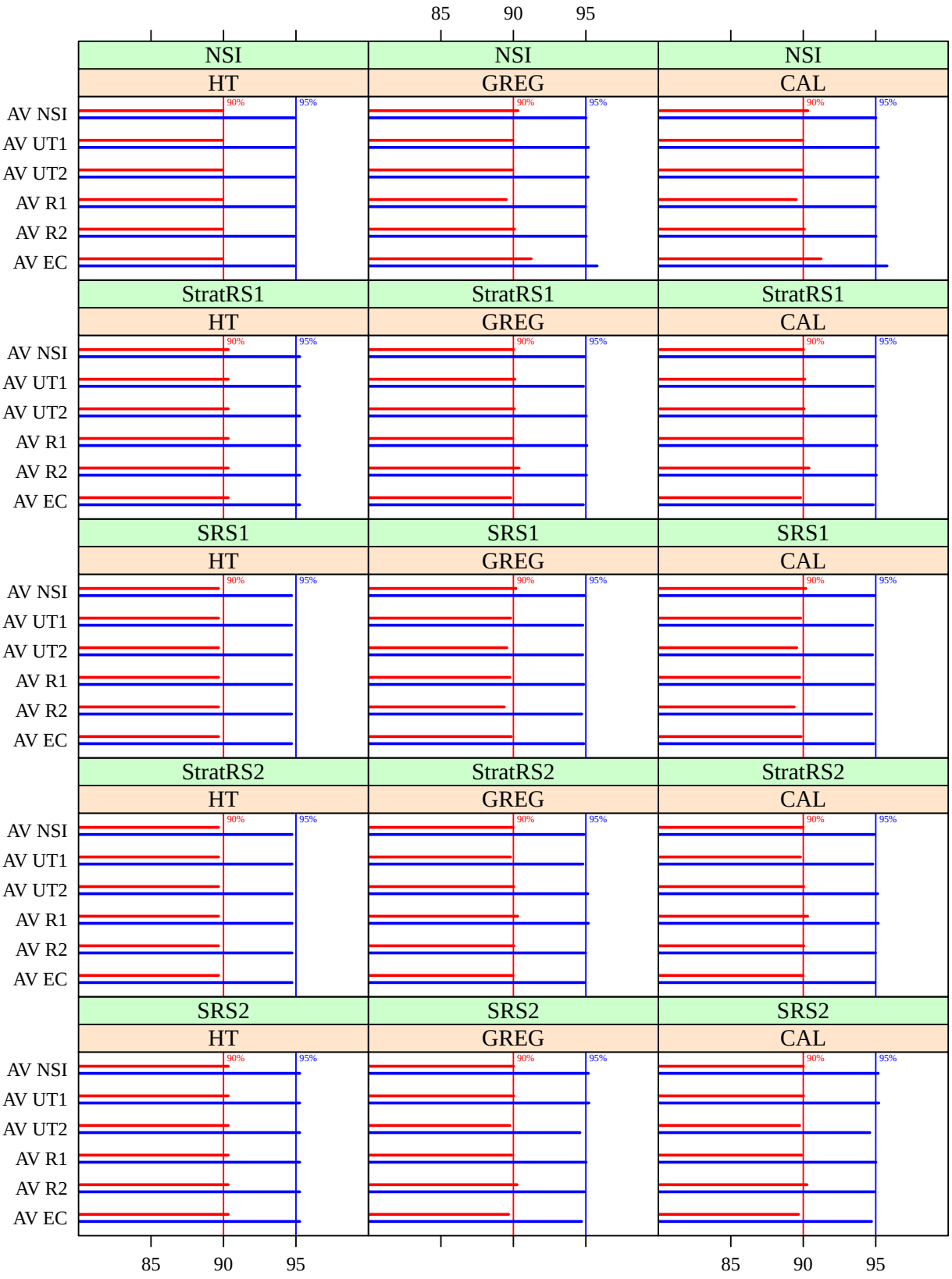

Figure 3.9: 90% and 95% confidence interval coverage rate for composite indicator

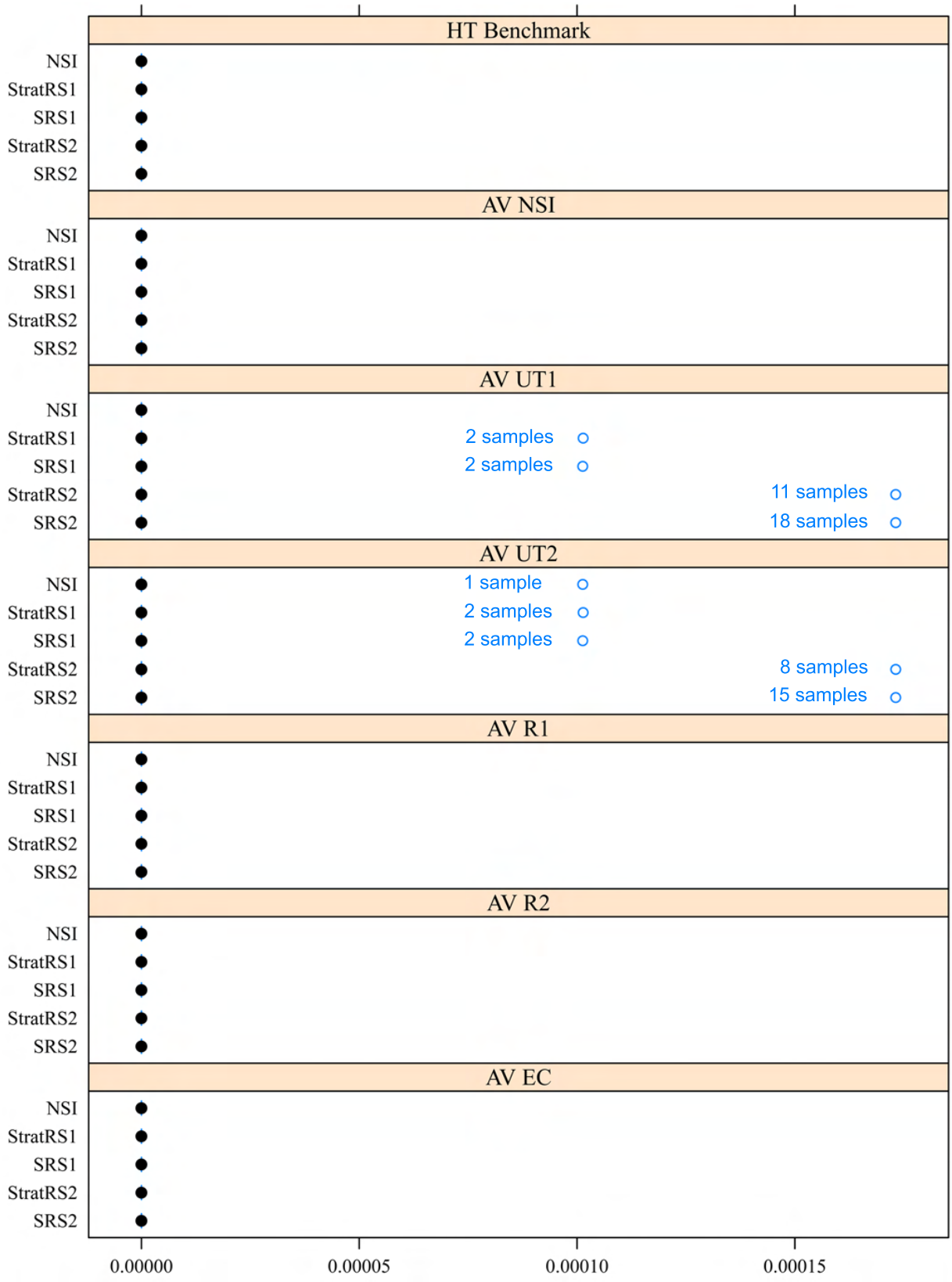

Figure 3.10: Negative weights ratio of GREG for composite indicator

3.5 Results of the single indicators

Within this paragraph results for the single indicators are given. Here, the effect of the auxiliary variable sets is analysed again and the results evaluated in light thereof. Analysis of the single indicators essentially confirms our provisional results.

Single indicator `income`

The point-scoring of the indicator is based on given quantile values. Originally these quantiles would have to be estimated from the sample which was omitted by complexity reasons. the interested reader is referred to the research done by the FP7 project AMELI[2], in which the according Laeken indicators are closely analysed.

The most precise estimate is obtained for the GREG estimator using the auxiliary variable set `UT1`. In the case of `UT2`, the lack of information on income (EF539) and on unemployment (EF215) results in correspondingly less precise estimation.

`R2` underlines, however, that for a more precise estimation of household equivalent income we require not only information on income, but on other socio-demographic characteristics too. `R1`, in which primarily information on income is entered, is found to perform worse than `R2`. `EC`, which permits more precise estimates for StratRS than does `R1` and `R2`, clearly shows that a person's income correlates with his/her age and gender. The household equivalent income depends, as well, on the household size.

It should be noted that the variables entered in `R1` and `R2` are selected from the population using forward-backward-selection. In the case of (slightly) biased samples, this choice of variables need not be optimal. MÜNNICH AND KNOBELSPIES (2008) show that more extensive analyses are required. Therefore, it is not absolutely necessary for `R1` to be better than `R2`.

In sum, it is apparent that `R1`, `R2`, and `EC` draw on too little information to keep the variation of the variance estimators low - especially when sampling is done using SRS and also when the sample size is reduced to 8,250 households. `UT1` yields, for these reasons, more stable estimates.

Single indicator `standard of living`

The auxiliary variable sets `UT1`, `UT2` and `NSI` generate precise variance estimates. Although the auxiliary variable set `UT2` does not contain variable EF215 or variable EF539, estimates result that are only slightly less precise than are those for the auxiliary variable set `UT1`. The lack of information on incomes can be compensated by co-opting other regressors (see Tables A.5 and A.6).

What `R1` and `R2` underline is that having only a few auxiliary variables available leads to considerable precision gains, as compared with HT estimation. Astonishingly, `R2` yields a slightly more precise projection than does `R1`. This result is conformant with the high precision of `UT2`.

The great variation in the variance estimators obtained for the `EC` auxiliary variable set in combination with `SRS` is due to - and points to the problem of - sparse cells. The statistical

[2]For more information on AMELI cf. `http://ameli.surveystatistics.net`

spread in the case of StratRS2 is much greater than in the case of StratRS1. Reducing the number of households from 14,100 to 8,250 particularly affects the weakly occupied classes.

Single indicator housing

The variables AESFDRK and BRGHMEF that are entered into the indicator are, compared with HH030, less dependent on the amount of income a household has at its disposal. UT2 and R2 therefore yield projections that are almost as precise as those generated by UT1 and R1. The variables co-opted to the single indicator can be explained well by his/her age and gender class. Thus, EC allows for variance estimates only a bit less precise than NSI. Altogether, UT1 supplies the most precise estimate, followed by UT2, NSI, and EC. Slightly more precise estimates result for R2 than for R1.

Single indicator education

The comparatively precise estimates generated by UT1 and UT2 can be attributed to the fact that the auxiliary information largely correlates with the indicator. No indications of a person's schooling and tertiary education are contained in the auxiliary variable set NSI, which explains why the estimates here are less precise.

R1 and R2 contain some few items of information on a person's schooling and tertiary education and also on household size and a person's age/gender and. But this information is insufficient to permit stable estimates. The variance estimators are strongly scattered as a result. texttttEC generates a little more precise estimates. The range of variation of the variance estimators is narrower as compared with R1 and R2.

Single indicator health

Of note here is the astonishingly high degree of precision achieved with auxiliary variable sets R1 and R2. The regressors contained in the other auxiliary variable set result in only small precision gains vis-à-vis the number of auxiliary variables entered. This is an indication of redundancy in the auxiliary variables. For single indicator 6, socialrelations, a similar picture emerges as for the health indicator.

Single indicator work

The great precision achieved by UT1 and R1 is due to the highly correlated information on variable EF215 contained in each auxiliary variable set. The auxiliary variable sets UT2 and NSI show that, even in the absence of this special variable, precise estimates can be generated. In the case of UT2, the other regressors entered compensate, in like manner, for the income information not contained. The case of R2 again underlines the effect of having fewer variables correlated with the indicator.

In sum, it should be noted that both the single indicators and the composite indicator can be estimated with high precision. The main thing is that adequate auxiliary information should be available that is correlated with the indicator of interest. The calibration estimators allow reliable estimates to be achieved, even for a sample size of 8,250 households, as per EU regulations. These are superior to those obtained by HT estimation. For an estimate using HT estimator a sample size of 14,100 households is recommended, as is also opting for an NSI design in order to avoid widely scattering estimates.

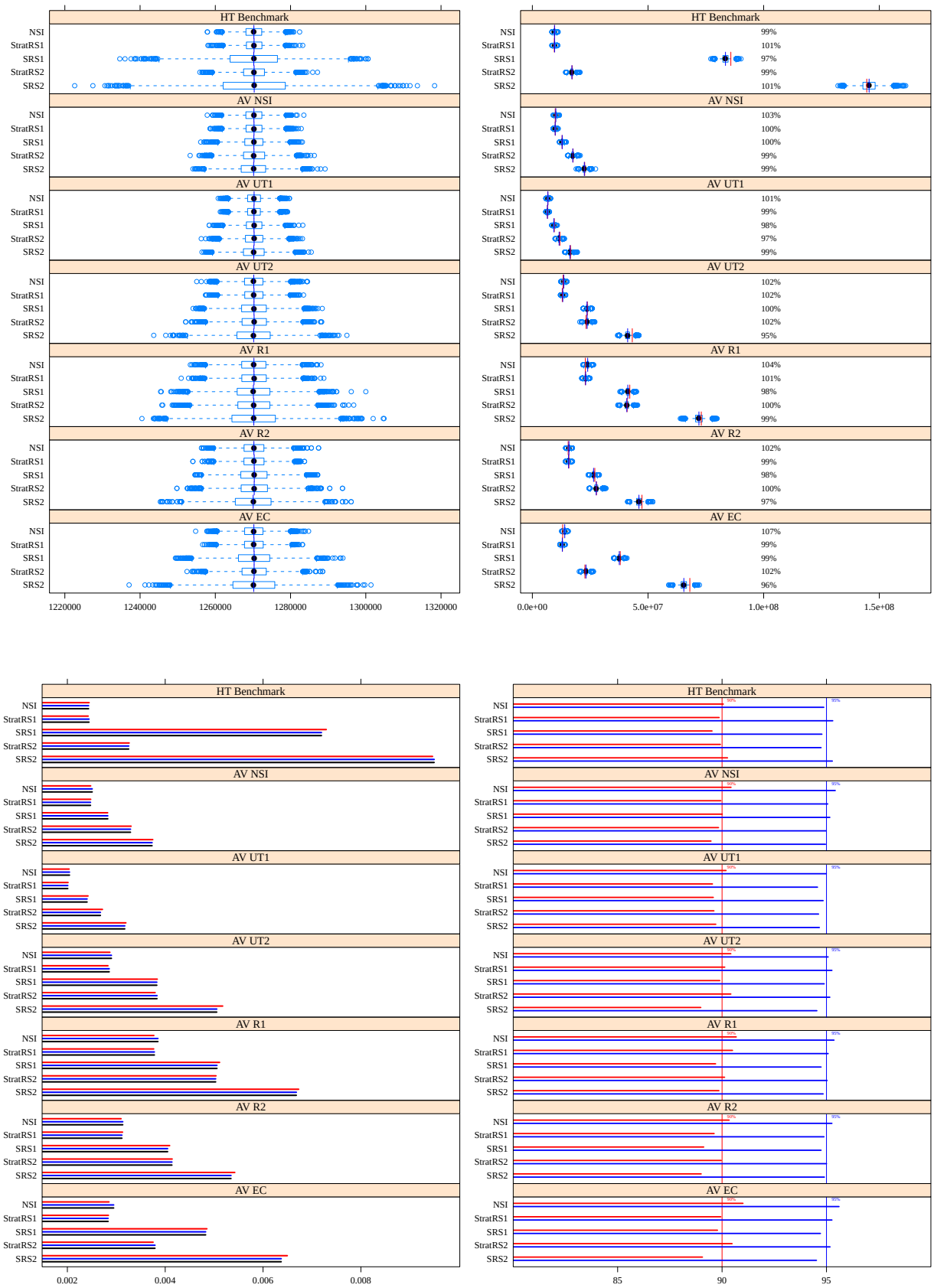

Figure 3.11: SI 1 - `income`; point- and variance estimators, measures, and coverage rates

Figure 3.12: SI 2 - `standard of living`; point- and variance estimators, measures, and coverage rates

Figure 3.13: SI 3 - `Housing`; point- and variance estimators, measures, and coverage rates

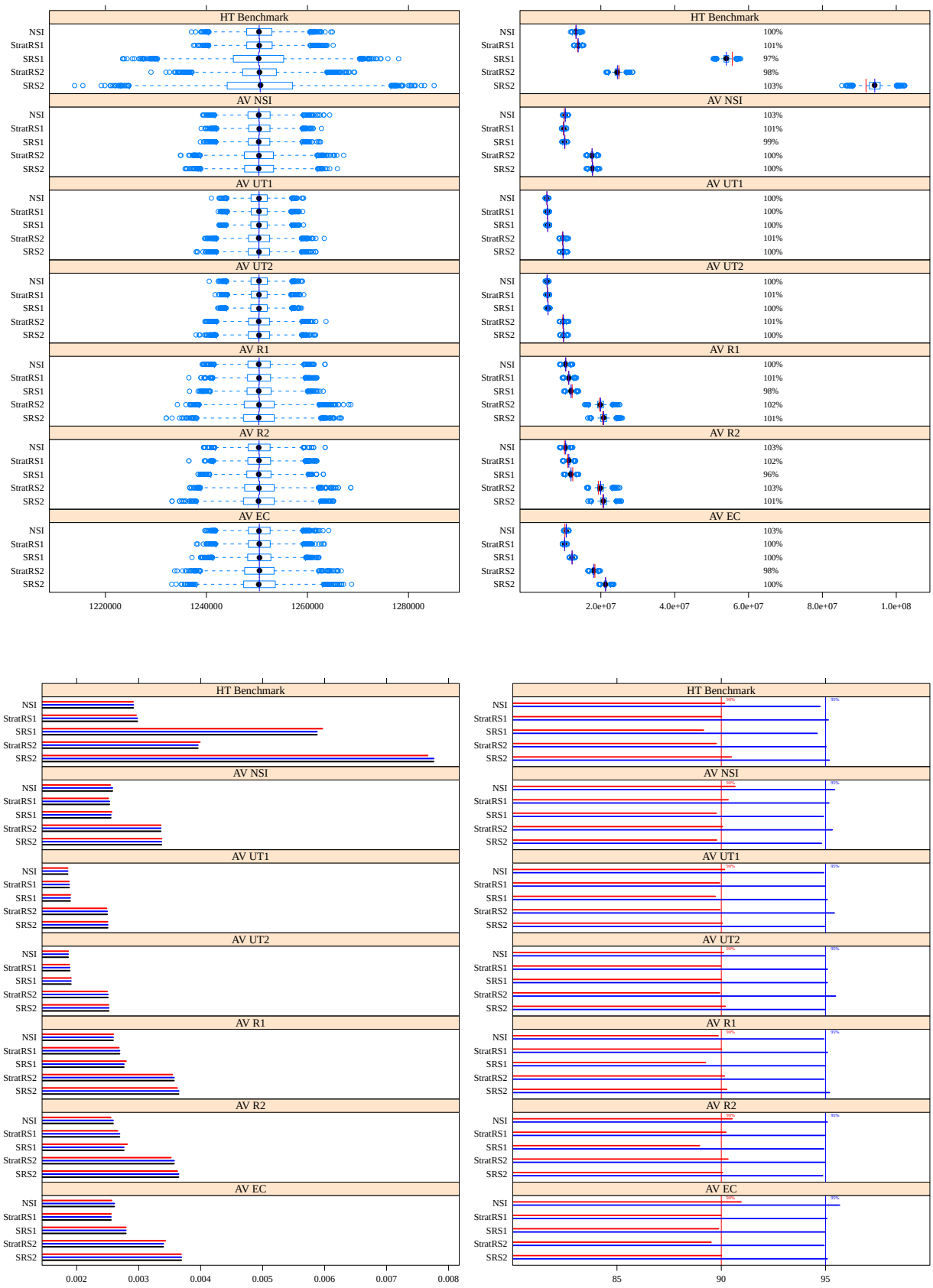

Figure 3.14: SI 4 - `education`; point- and variance estimators, measures, and coverage rates

Figure 3.15: SI 5 - `health`; point- and variance estimators, measures, and coverage rates

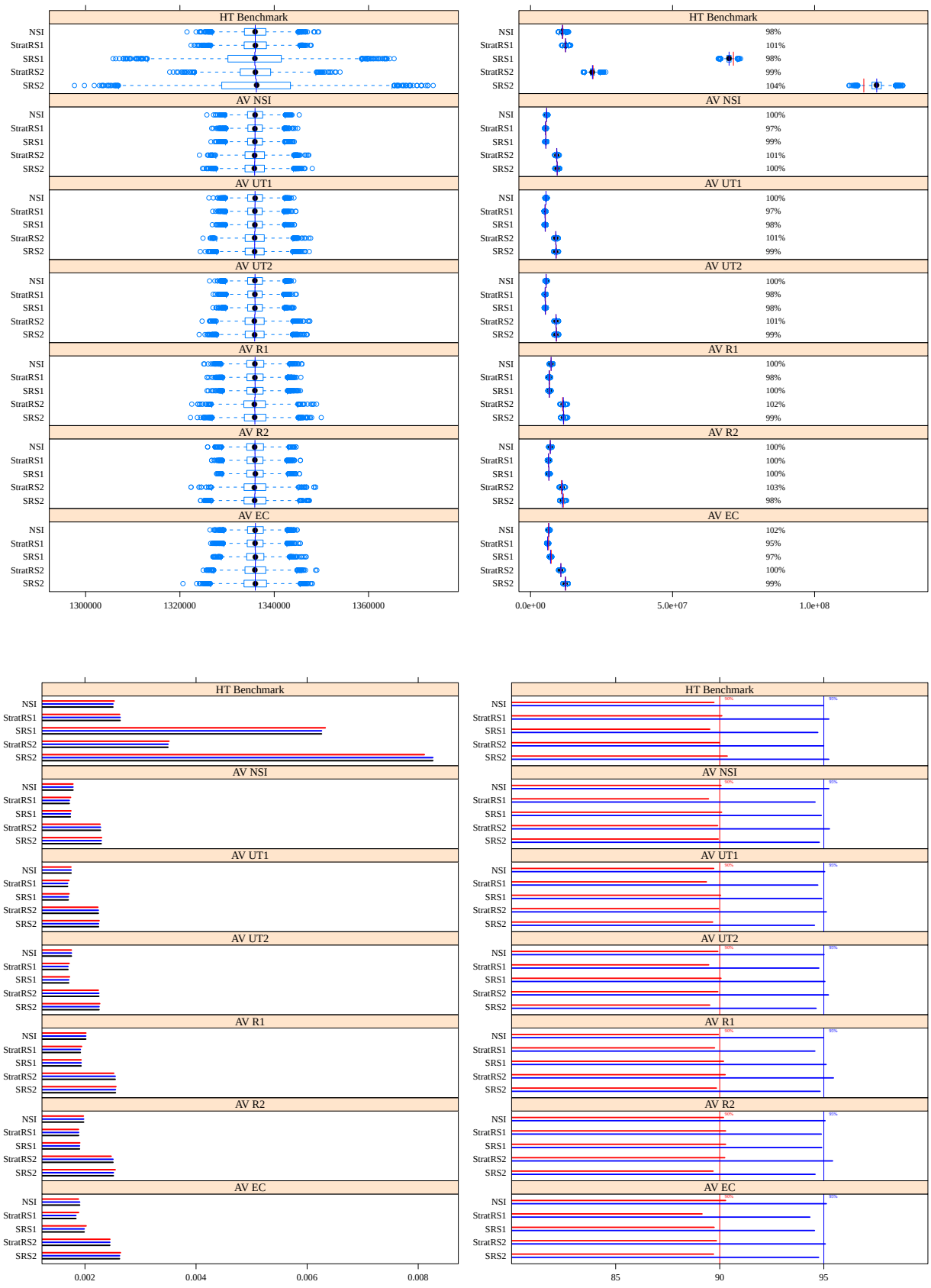

Figure 3.16: SI 6 - `social relations`; point- and variance estimators, measures, and coverage rates

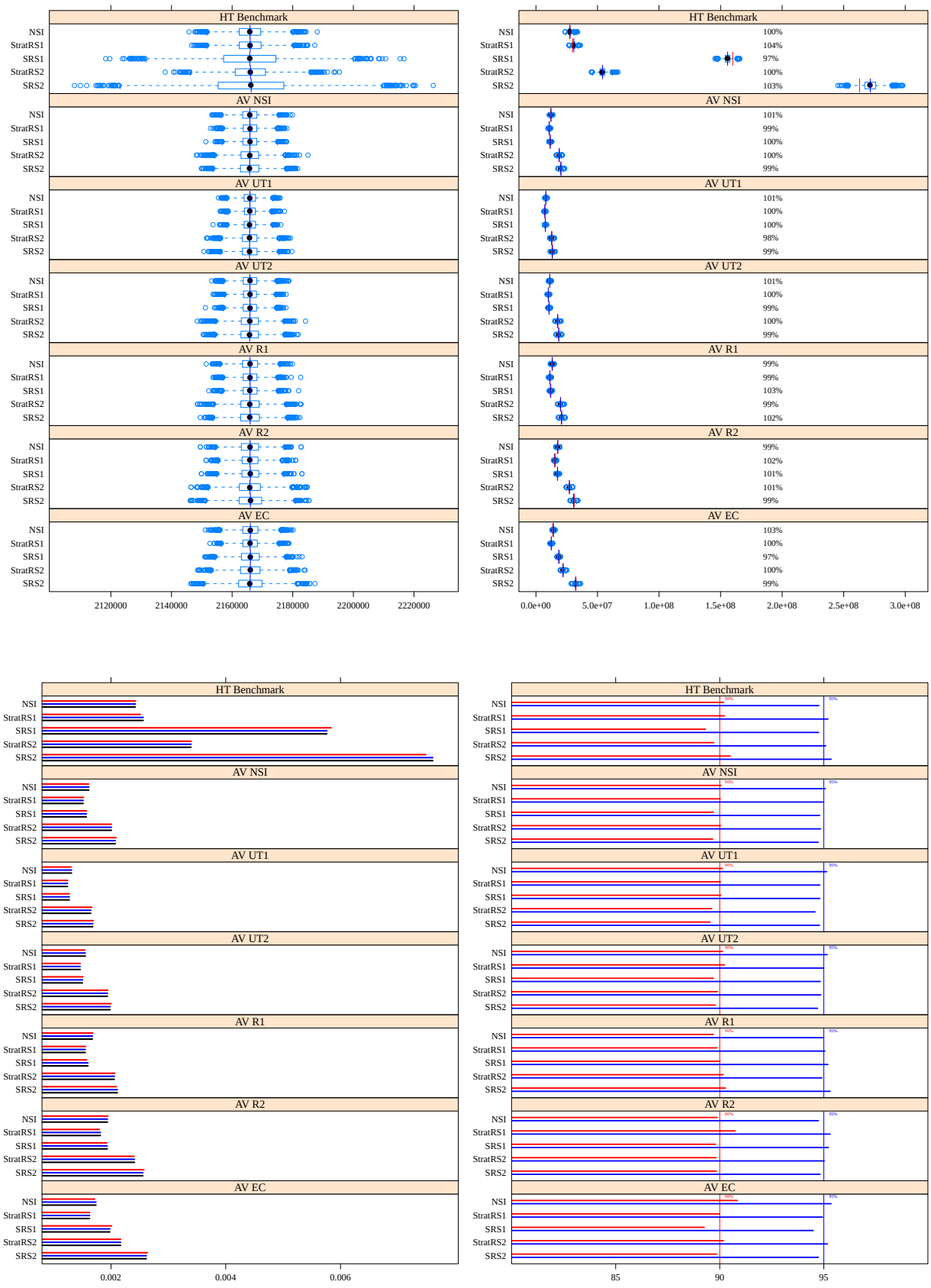

Figure 3.17: SI 7 - `work`; point- and variance estimators, measures, and coverage rates

3.6 Variance estimation of the composite indicator

In practice, macrodata are available for a multiplicity of single indicators (NewCronos or OECD database). For single indicators, though, mostly little in the way of statistical information is available. This holds especially for information on how the indicator was obtained, on the estimation method used or on its accuracy measurement. Available single indicators often stem from very different samples, and no international consensus exists on how they should be estimated. Only to a small extent can a simulation study reflect, and compensate for, these practical restrictions. But the simulation study, and its results, depends on doing so.

A general point worth noting is that official statistics (in the form of register data and/or basic populations) are only to a limited degree available for scientific study. Numerous items of indicator data, e.g. those collected by companies and other economic actors, can only be accessed for a fee. On the whole, there is a severe lack of available microdata on single indicators for KBE.

Defining composite indicators on the basis of available macrodata is not recommended, given the statistical properties of a composite indicator. Especially in view of the fact that single indicators stem from different surveys, no information is available as to how single indicators interact statistically. It cannot be generally assumed, that indicators, statistically speaking, are independent of each other, i.e. the co-variances of any two single indicators are zero or, alternatively, the co-variances of the incorporated single indicators add up to zero.

Using such macrodata as are available at the time of writing, the co-variances cannot be calculated or estimated. In view of these caveats, only the sum of the variances of the single indicators incorporated into a composite indicator can be used as a surrogate for the composite indicator's true variance. The extent to which this surrogate can reflect the size of the true variance was tested within the simulation study for ZUMA's modified composite indicator.

Figures 3.18 by 3.21 show the variance estimates resulting for HT and GREG. Figures 3.18 and 3.19 represents the results when auxiliary variable sets AV `NSI`, AV `UT1`, AV `UT2` und AV `EC` are applied. The boxplots of range 0 to 80,000,000 are drawn on a larger scale in figure 3.19. The following variance estimates are graphically reproduced:

- the sum of the estimated variances of the single indicators (boxplot with a black point in the middle);
- the variance of the composite indicator as sum of the estimated variances plus co-variances of the single indicators (boxplots with a red point in the middle).

This refers to the question of the negligibility of the covariances in Equation (3.3) which obviously is violated. The vertical red line indicates the true variance of the composite indicator. The vertical blue line marks the mean estimated variance (over 10,000 independent samples drawn from the population) based on single indicators.

The percentage data in figure 3.18 indicate the ratio between the estimated and the true variance of the composite indicator. The quotient of the estimated `sum of the`

`variances of the single indicators` (SVSI) and of the estimated `sum of variance-covariances of the single indicators` (SVCSI) varies. The percentage data shown in black indicate that the SVSI cannot - as was expected - serve as a surrogate for the (true) variance of the composite indicator. For HT estimator the quotient lies between 16 and 23 percent, for the GREG estimator - depending in the auxiliary variable set selected - between 19 and 40.

Such under-estimation (by more than 60 percent) of the true variance is too great for a reliable confidence interval to be assigned for the composite indicator. The coverage rates collapse accordingly (see figure 3.22). The 90 and 95 percent marks are clearly undercut.

The percentage data shown in red in figure 3.19 indicate that SVCSI corresponds to the (true) variance of the composite indicator. The percentages are in the order of 100%, with only small deviations.

Figures 3.20 and 3.21 show the results using auxiliary variable sets `R1` and `R2`. Figure 3.21 gives the range from 0 to 80,000,000 on a larger scale. Compared to Figures 3.18 and 3.19 three boxplots are plotted for each combination of auxiliary variable sets and survey designs:

1. SVSI (boxplot with a black point in the middle);
2. SVCSI (boxplots with a red point in the middle);
3. Variance estimates of composite indicator (boxplots with a green point in the middle).

In the cases of items 1 and 2, each single indicator is estimated using its own auxiliary variable set (see Tables A.7 and A.8), consisting of any 10 selected regressors. The regressors are specified using R-Routine regsubsets. An estimate of the composite indicator results as the weighted sum of the (estimated) single indicators.

Item 3 is when a dataset includes all variables co-opted to the composite indicator. Hence, the score of the composite indicator can be calculated for each sample unit. The population total of the composite indicator is estimated on the basis of individual scores. Concerning the auxiliary variable set, the entered regressors are selected for best possible up-sampling of the afore-mentioned individual scores.

As the composite indicator is a linear statistic, estimation according to items 2 and 3 is congruent when applying i) HT or ii) GREG using an identical set of auxiliary variables applied to all single indicators. According to figure 3.21 GREG leads to different estimates as a function of varying explanatory content on the part of the auxiliary variable sets.

The data in figure 3.21 indicate the mean variance estimate (marked by the blue vertical lines in the boxplots). The (true) variance is considerably under-estimated, while the SVSI only amounts to some 52 to 66 percent of the SVCSI. However, the quotient of SVSI and SVCSI using `R1` and `R2` is higher as compared to estimates in figure 3.19 (auxiliary variable sets AV `NSI`, AV `UT1`, AV `UT2` und AV `EC`).

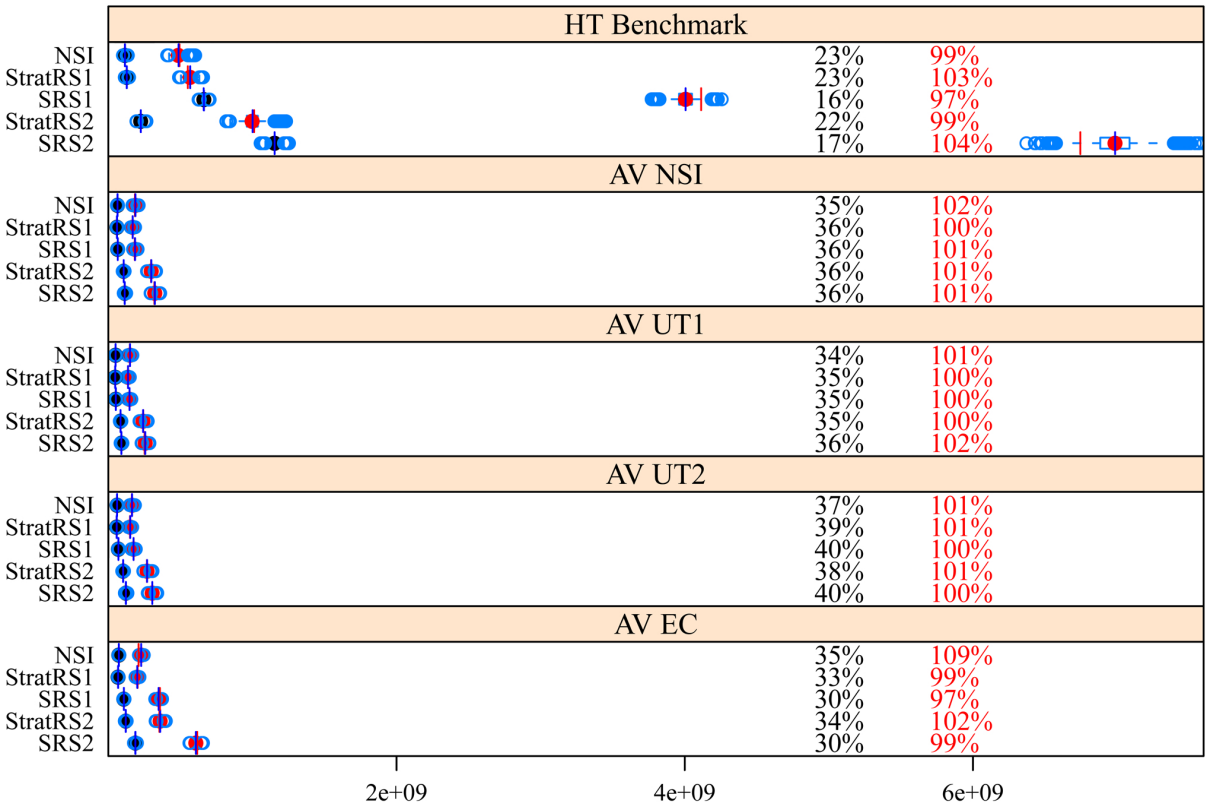

Figure 3.18: Variance estimates of CI with and without co-variances

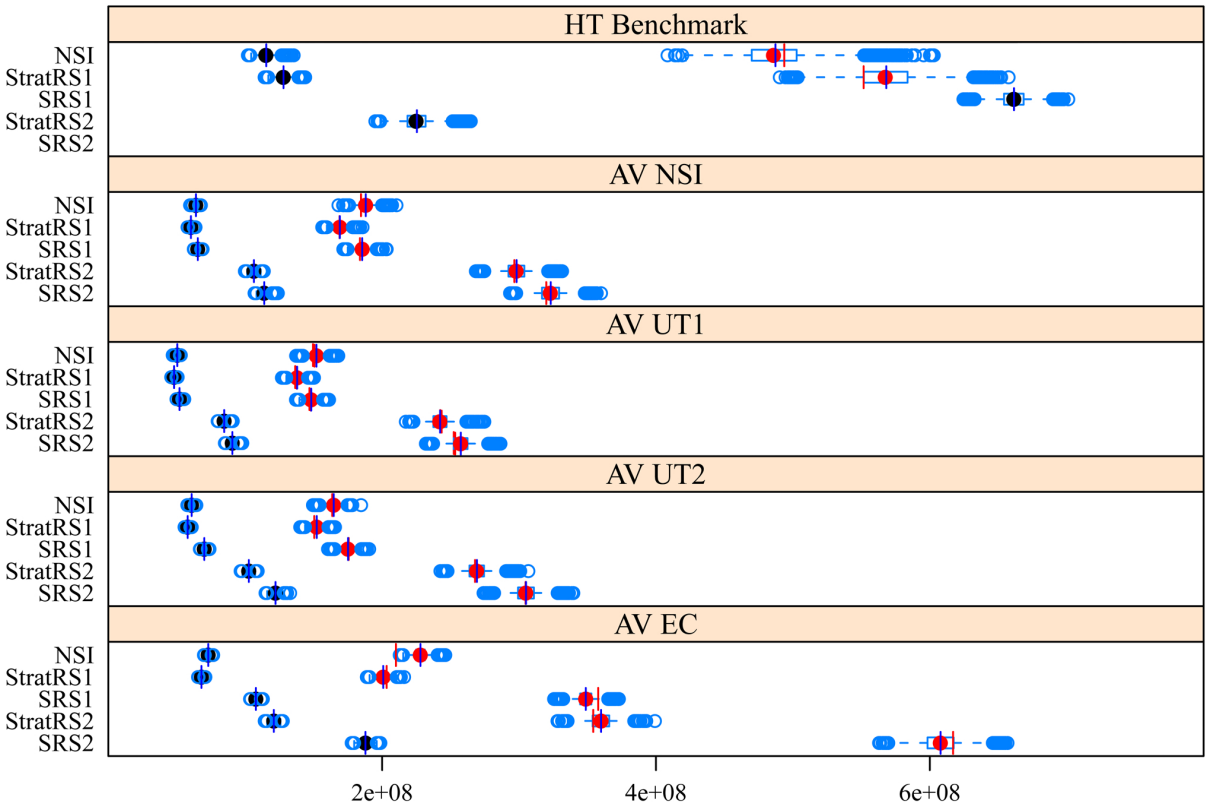

Figure 3.19: Variance estimates of CI drawn on larger scale

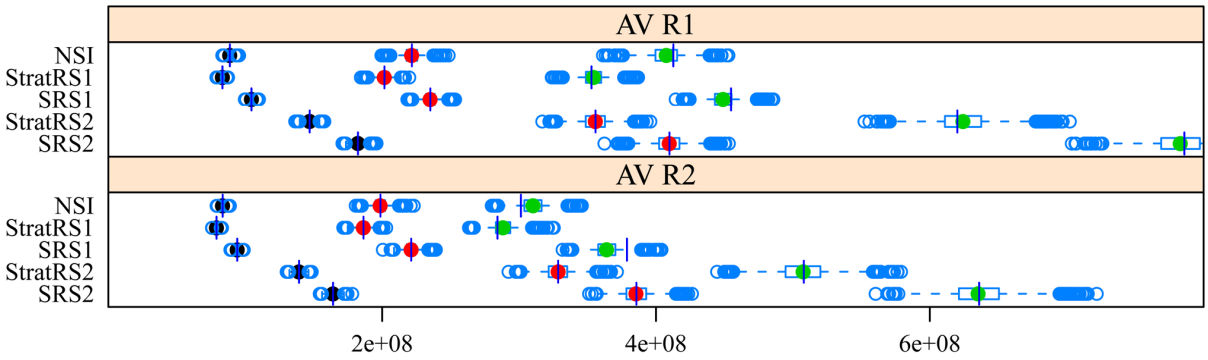

Figure 3.20: Variance estimates of CI for `R1` and `R2`

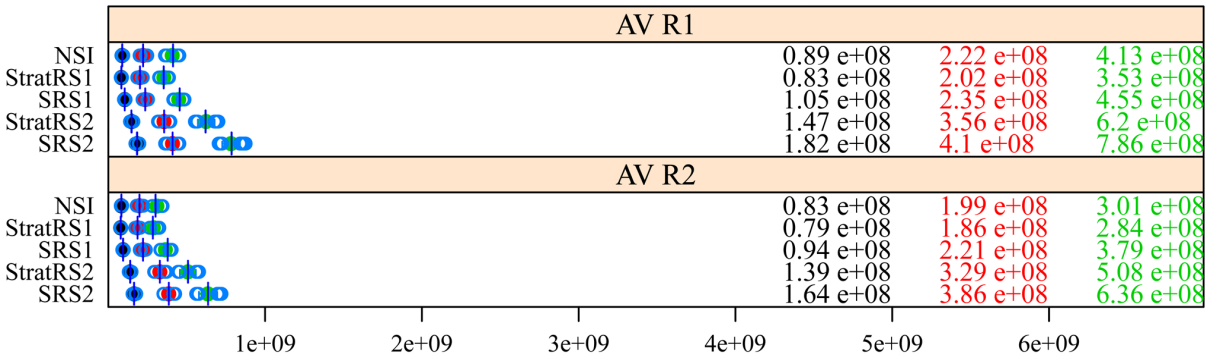

Figure 3.21: Variance estimates of CI for `R1` and `R2` drawn on larger scale

In sum, it can be said that when tabling estimates of a composite indicator, information on both the sampling scheme and the estimation technique (including details of the auxiliary variables used) is necessary to ensure enhanced comparability of results and, in consequence thereof, enhanced transparency.

3.7 Summary

The simulation study demonstrates that the sum of estimated variances of single indicators is not a suitable surrogate for the (true) variance of the composite indicator. The (true) variance of the composite indicator in general is severely under-estimated. Also, the postulate of the independence of single indicators jointly entered into a composite indicator is problematic. Confidence intervals, as calculated assuming independence on the part of samples and/or indicators, cannot represent the true range of variation of a composite indicator or can do so only to a very limited extent.

In case composite indicators are of great interest and the need to assess the quality of the composite indicators, access to microdata seems to be necessary in order to avoid severely biased accuracy measures. If truly reliable results are being sought, the best course would be to provide optimally uniform statistical surveys, i.e. with uniform standards, for purposes of achieving data quality and precision in estimating variables of interest. With

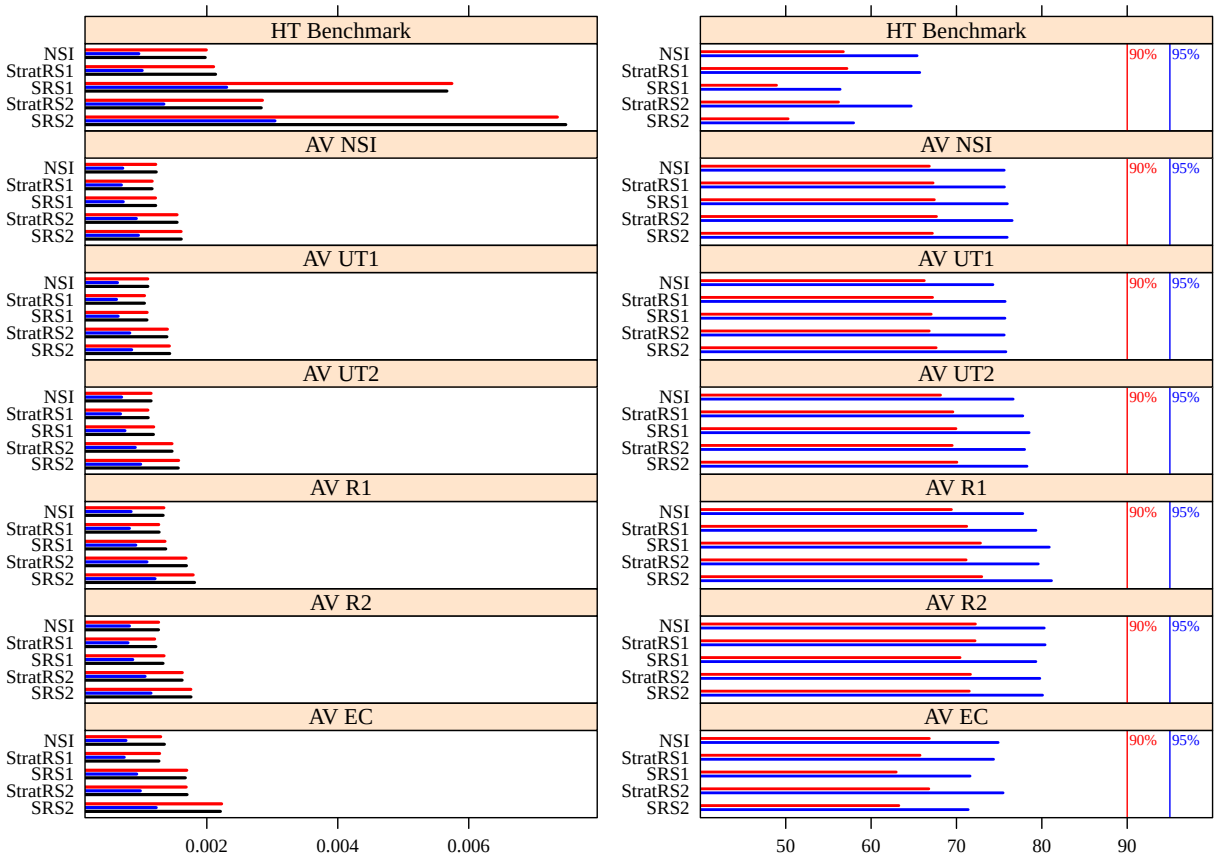

Figure 3.22: Accuracy measures for CI without consideration of co-variances

a view to curtail costs, especially in the matter of official statistics, exploring the options of expansion and linkage is an urgent priority. When it comes to defining and calculating composite indicators, the interfaces between surveys are of great importance, especially for estimating variances correctly. It is worth pointing out that for many indicators still nowadays no precise variance estimates are available, at least not for end-users which are not from Official Statistics. Additionally and not only for composite indicators, it may be worth to consider providing measures of dependency between single indicators or survey variables (e.g. co-variances) in quality reports.

It is appropriate, in the present context, to think about putting in place an efficient data system of available statistics, able to cater to the needs of politics and science alike. In order to answer the question of whether we do in fact measure what we want to measure (or are able to with the statistics currently available), it is indispensable that we have transparent access to informative metadata on surveys and indicators; such metadata should, in addition, be cross-linked to official research data centres and databases. Especially in defining new composite indicators, no surrogate can be found for the efficient exchange of information between scientists, official statisticians and, if necessary, economic researchers.

In terms of the single indicators drawn on, including how they are weighted and scored, the composite indicators that have been presented here are certainly limited and the methodology has to be improved and extended to more complicated cases. This should enable interested researchers to include all steps of generating composite indicators as

described in workpackages 5 and 7.

Using data across time allows conducting longitudinal analyses for which variance estimation over time should be applied. As a special case, keeping track of development of composite indicators leads to the demand of point and variance estimation for change (cf. workpackage 9 of the DACSEIS project: `http://www.dacseis.de`). The way EU-SILC and ICT are collected in Germany may possess exemplary value for other countries. By cross-linking with the GMC, numerous auxiliary variables can be included. What this means is considerably improved precision in estimating characteristics of interest, provided the data comply with the prerequisites.

Chapter 4

Estimation of composite indicators in small areas

4.1 Introduction

In general, sample surveys are designed to provide reliable estimates for *considerably* large areas of a population. Hence, the accuracy of survey variables or even composite indicators in this context should be sufficiently high. Nowadays, the necessity to derive better estimates for smaller sub-populations becomes more and more evident not only for providing policy figures on smaller regions or different subpopulations. The features that identify an area as `small` can be interpreted in different ways as discussed later. Regardless of whether the characteristic small is attached to the absolute or relative size of a subpopulation, the related sample size in the considered area may occur to be random, and hence small or even zero in some cases.

Questions regarding allocation of subsidies, in particular European subsidies, fiscal equalisation schemes (e.g. between Federal government and the Federal states), income and poverty (e.g. Lisbon target in decreasing poverty and their level of target achievement), or generating reliable estimates in the context of an European census based on register data, have obvious needs for establishing small area statistics which refers to regional disaggregation techniques.

Small areas, referred to as subpopulations of a large population, may be defined as either `small areas`, in case of geographic areas, or as `small domains`, in case of socio-demographic sub-groups. Examples of a geographic area include a federal states, administrative districts, counties, a city or commune, and a metropolitan area. A socio-demographic domain may refer e.g. to a specific age-gender-ethnicity group within a (larger) geographic area.

Direct estimators, such as the Horvitz-Thompson or calibration estimators, only use data of sample units corresponding with the area of interest. Due to the most likely small sample sizes in the small areas or domains, direct estimators yield standard errors which may be unacceptably large. Thus, indirect estimators are constructed that `borrow strength` from related or similar areas, increasing the effective sample size and therewith the estimation precision. Indirect estimators are based on either explicit or implicit models

providing a link between the small area considered and the related areas through auxiliary information. These auxiliary variables can be miscellaneous, cross-sectional as well as across time, for example information from neighbouring or next higher populations, data from a previous census or administrative records. Due to the growing demand for reliable statistics on the level of small areas, small area estimation (SAE) is already becoming an important field in survey statistics. A thorough overview of small area estimation methods is given in RAO (2003) or JIANG AND LAHIRI (2006).

The European Commission shows also great interest in small area applications. Within the 5th Framework Programme the projects EURAREA (`http://www.statistics.gov.uk/eurarea/`) and DACSEIS (`http://www.dacseis.de`) were investigating small area methods. Within the 7th Framework Programme small area estimation methods will be applied to poverty measurement within the projects SAMPLE (`http://www.sample-project.eu/`) and AMELI (`http://ameli.surveystatistics.net`).

Nowadays, the methodology of SAE is already applied in Official Statistics. Its demand is still increasing and attracts interest of policy makers, e.g. regarding allocation of governmental funds and in regional planning. As for reasons of economy and acceptability, there are usually no full population data available, the statistician has to use survey data for tabling actual key data. Thus, studies in regional concerns and analyses of socio-demographic key data require specific estimation techniques.

The aim of this Section is to elaborate applications of small area statistics to composite indicators. The given example will be conducted in close connection to the calibration approach from the last Chapter. The focus will be on point estimation in order to show possible applications of small area statistics to regional disaggregation.

The following Section gives a short overview of the classical estimators which are called the EURAREA standard estimators. After giving a short overview of the simulation study and its setting some selected results are presented and summarized.

4.2 Estimators of interest

The given study uses the so-called EURAREA standard estimators. These comprise the naive national sample mean, two design-based estimators as well as the two sets of standard small area estimators which are based on unit-level and area-level information. Further details, especially on MSE estimation, can be found on the EURAREA homepage. SAS macros can be found there as well, whereas R programmes are available at the DACSEIS homepage.

4.2.1 National sample mean: `NSM`

The `NSM` for variable Y is a fixed value for every area. The mean value μ is calculated using the following formula:

$$\widehat{\mu}_Y = \sum_{i \in s} \omega_i \, y_i / \widehat{N} \quad ,$$

where the estimated population size is

$$\widehat{N} = \sum_{i \in s} \omega_i$$

and ω_i is the inverse of the sample inclusion probability π_i of individual i ($\omega_i = 1/\pi_i$). The sum is taken over sample s. The `NSM` is not a small area estimator in terms of the definition, because it contains no area specific information. This estimator was included for aims of comparison with other estimators which take into account the differences between areas. The `NSM` is a very poor estimator for small areas, as it will produce large errors for areas which true population value differs severely from the national mean. Otherwise, the `NSM` yields a good estimate in case of a congruent structure in the small area as well as in the whole data set. In general, the estimators, which will be described below performs better in terms of bias and efficiency. In the context of estimating indicators the `NSM` is used as a total estimator τ for variable Y in area d as follows:

$$\widehat{\tau}_{Y_d} = 1/\widehat{N} \cdot \sum_{i \in s} \omega_i \, y_i \cdot N_d \qquad ,$$

when N_d is known in area d. As the `NSM` is multiplied by the population size in area d (N_d, expanded to a total), τ_{Y_d} yields to different values for each area d.

4.2.2 Direct estimator: `HT`

The direct estimator of the total in area d is defined as the design-weighted Horvitz-Thompson estimator for each area:

$$\widehat{\tau}^{\mathrm{HT}}_{Y_d} = \frac{N_d}{\widehat{N}_d} \sum_{i \in s_d} \omega_i \, y_i \qquad ,$$

where

$$\widehat{N}_d = \sum_{i \in s_d} \omega_i \qquad .$$

The sums are taken over sample s_d from area d and the design weights are also the inverses of the inclusion probabilities. The direct estimator of the domain total is approximately unbiased (Särndal et al. (1992), p. 185).

4.2.3 GREG estimator: `GREG`

For generating more precise estimates the direct estimates will be added by some auxiliary information. The auxiliary variables have to be known as totals for each area d as well as they have to be measured in sample s. To allow for the differences between the sample and population area totals of the auxiliary variable X, the direct estimator is adjusted, usually applying the standard linear model. This yields the generalised regression estimator

$$\widehat{\tau}^{\mathrm{GREG}}_{Y_d} = \widehat{\tau}^{\mathrm{HT}}_{Y_d} + \left(\boldsymbol{\tau}_{X_d} - \widehat{\boldsymbol{\tau}}^{\mathrm{HT}}_{X_d} \right)^T \widehat{\boldsymbol{\beta}} \qquad ,$$

where

$$\widehat{\boldsymbol{\tau}}_{X_d}^{\mathrm{HT}} = \sum_{i \in s_d} \omega_i \, \boldsymbol{x}_i \qquad .$$

$\boldsymbol{\tau}_{X_d} = \left(\tau_{X_{d,1}}, ..., \tau_{X_{d,p}}\right)^T$ is the vector of true totals of p covariates in the area d and $\widehat{\boldsymbol{\beta}}$ is the least squares regression estimate assuming a standard linear model $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$ with independent errors $\varepsilon_i \sim N(0, \sigma^2)$ for each unit i in the sample:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in s} \omega_i \boldsymbol{x}_i y_i \qquad .$$

Note that $\boldsymbol{\beta}$ is estimated using the entire sample s. An alternative presentation for the GREG estimator is through g-weights:

$$\widehat{\tau}_{Y_d}^{\mathrm{GREG}} = \sum_{i \in s} \omega_i \; g_{di} \; y_i \qquad ,$$

where g-weights depend on the domain d, element i and whole sample s:

$$g_{di} = \frac{N_d}{\widehat{N}_d} \cdot z_{di} + \left(\boldsymbol{\tau}_{X_d} - \frac{N_d}{\widehat{N}_d} \widehat{\boldsymbol{\tau}}_{X_d} \right)^T \left(\sum_{i \in s} \omega_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i \qquad ,$$

with domain indicators $z_{di} = 1$, if $i \in s_d$ and $z_{di} = 0$, otherwise.

The main property of the g-weights is that g-weighted sample sum of the auxiliary values equals the known domain total of these values (SÄRNDAL et al. (1992), p. 401):

$$\sum_{i \in s} \omega_i \; g_{di} \; \boldsymbol{x}_i = \sum_{i \in U} z_{di} \; \boldsymbol{x}_i = \sum_{i \in U_d} \boldsymbol{x}_i \qquad .$$

4.2.4 Synthetic estimator: SYNTH

An estimator is called a synthetic if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for at least one of these small areas. The underlying assumption is that the small areas have the same sample characteristics as the large area, mentioned above. If this assumption is satisfied, the synthetic estimator is very efficient as its MSE is small. For areas with strong individual effects the synthetic estimator can be heavily biased (RAO, 2003, p. 46).

Assuming that unit-level auxiliary data $\boldsymbol{x}_{di} = (x_{di1}, ..., x_{dip})^T$ is available for each population element i in small area d, the variable of interest y_{di} is related to $\boldsymbol{x}_{di}$ through a nested error linear regression model:

$$y_{di} = \boldsymbol{x}_{di}^T\boldsymbol{\beta} + u_d + \varepsilon_{di} \qquad ,$$

where $\boldsymbol{\beta}$ is vector of regression coefficients, u_d is the area-specific effect with $\mathrm{E}(u_d) = 0$, $\mathrm{var}(u_d) = \sigma_u^2$ and ε_{di} is the independent random error with $\mathrm{E}(\varepsilon_{di}) = 0$ and $\mathrm{var}(\varepsilon_{di}) = \sigma_\varepsilon^2$. The synthetic estimator is given by the formula

$$\widehat{\tau}_{Y_d}^{\mathrm{SYNTH}} = \boldsymbol{\tau}_{X_d}^T \widehat{\boldsymbol{\beta}} \qquad ,$$

with known area-level covariates vector $\boldsymbol{\tau}_{X_d} = (\tau_{X_{d1}}, ..., \tau_{X_{dp}})^T$.

4.2.5 Composite estimator: EBLUP

Composite estimators attempt to balance the potential bias of synthetic estimators and the instability of direct estimators. In KEI, one composite estimator was used combining a GREG estimator with a synthetic estimator, as described in the sections above. The composite estimator is BLUP (best linear unbiased predictor) for small areas. The literature presents some more estimators which differ on one hand in the model the estimator is based on and on the other hand in the kind of the direct estimator. Rao (2003) shows a broad overview on the variations of small area estimators.

The following EBLUP (empirical best linear unbiased predictor) estimator is a weighted combination of the synthetic and the GREG estimator. It is given by the formula

$$\widehat{\tau}_{Y_d}^{\text{EBLUP}} = \gamma_d \widehat{\tau}_{Y_d}^{\text{GREG}} + (1-\gamma_d)\widehat{\tau}_{Y_d}^{\text{SYNTH}} = \gamma_d(\widehat{\tau}_{Y_d} - \widehat{\boldsymbol{\tau}}_{X_d}^T \widehat{\boldsymbol{\beta}}) + \boldsymbol{\tau}_{X_d}^T \widehat{\boldsymbol{\beta}} \quad ,$$

where

$$\gamma_d = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2 / n_d} \quad .$$

$\widehat{\tau}_{Y_d}$ and $\widehat{\boldsymbol{\tau}}_{X_d}$ are the sample totals of the target variable y and the vector of auxiliary variables $\boldsymbol{x}$ for area d respectively; $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\sigma}_u^2$ are the parameter estimates of the two-level linear model. The weight γ_d ($0 \leq \gamma_d \leq 1$) measures the model variance $\widehat{\sigma}_u^2$ relative to the total variance $\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2/n_d$. If the model variance is relatively small, more weight is attached to the synthetic component. On the other hand, more weight is attached to GREG estimator if the domain sample size n_d increases (Rao (2003), p. 136).

4.3 Object of the simulation study

Within the following simulation study the standard estimators are analysed for their applicability to composite indicators, i.e. to CIoILC. The data handling process is exactly the same as presented in Chapter 2.

In order to gain distributions of the small area estimates, 10,000 replications were performed for each task on the small area estimators which were evaluated and compared to classical direct estimators. In order to avoid presenting too much information different kinds of plots and statistical measures are applied (cf. section 4.3).

As small area categorisation, the analysis uses a mixture between geographical and socio-demographic classification. The $D = 581$ small areas coincide with the strata of the stratified random sample (details can be found in Section 3.3.3 and Appendix C).

The focus is to show the applicability and elaborate the performance of given design-based and small area estimators considered above. The estimates are taken on the composite indicator CIoILC and its sub-indicators.

The results of the simulation study will be compared with respect to the impact of the following factors of possible influence:

- the type of target variable,
- the set of auxiliary variables,
- the sampling method.

The following paragraphs give an overview of estimators used and the auxiliary variable sets applied to estimate target variables. Furthermore, the implemented sampling procedures, the consulted accuracy measures, and the types of graphics used for characterising the simulation results are described. Basically, the subsequent sections describe the specifics of the SAE part in the simulation study. The study is based the objective described in the Subsections 3.3.1 to 3.3.5.

4.3.1 Target and auxiliary variables

Within the simulation study there are two variables of interest:

- the composite indicator `CIoILC`
- the sub-indicator `standard of living` (cf. tables 2.1 and 2.5 in general chapter 2).

The choice is based on the research work and its results. The estimation results presented in the following sections are conducted using the following auxiliary variable sets:

- `UT1`: Most of available variables (szh, hhtyp, EF1, EF30, EF32, EF35, EF52, EF110, EF130, EF138, EF215, EF259, EF261, EF338, EF359, EF521, EF539).
- `UT2`: `UT1` excluding variables EF539 (income) and EF215 (collecting unemployment benefits), both highly correlated with the indicators of interest.
- `NSI`: Auxiliary variables recommended by DESTATIS (szh, hhtyp, EF1, EF30, EF32, EF35, EF52, EF521, EF539).
- `EC`: Auxiliary variables recommended by EUROSTAT (EF1, EF30 & EF32, EF521).
- `R1`: Auxiliary variables selected with a R-routine `regsubsets` (EF35, EF52, EF110, EF138, EF215, EF261, EF521, EF539).
- `R2`: As `R1` excluding variables EF539 (income) and EF215 (unemployment benefits).

A detailed description of the variables listed above is given in subsection 3.3.2. The reader is also referred to tables A.5 and A.6 in appendix A.2. The object of this section is to analyse (ceteris paribus) the influence of auxiliary variable sets on the estimates of the composite indicator.

4.3.2 Accuracy measures

In accordance to Subsection 3.3.4 the following accuracy measures are applied to the small area simulation study for the purpose of comparing the simulation results. The selection of the measures is based on Münnich et al. (2004) and Office for National Statistics et al. (ed., 2003). In addition to the `RRMSE`, already described in section 3.3.4, two further measures will be applied:

`Relative bias`

A basic accuracy measure is given by the bias of a statistical expansion, defined as:

$$B(\widehat{\theta}) = E(\widehat{\theta} - \theta) \quad ,$$

where $\widehat{\theta}$ is an estimator of the population parameter θ. For a better interpretation of this measure, the bias is related to the true value of the population parameter θ:

$$RB = \frac{B}{\theta} \quad ,$$

called `relative bias`. Alternatively, the true value θ can substituted by $E(\theta)$.

`Relative dispersion`

The relative dispersion is a measure of variation, which is suitable for evaluating small area estimators. The dispersion is a difference between quantiles of the distribution of the parameter of interest $\widehat{\theta}$. Commonly the ratio between the dispersion and the true value or - in the present case - the median of the distribution of variable $\widehat{\theta}$ is used as a level of correction:

$$RD = \frac{\widehat{\theta}_d^{0,95} - \widehat{\theta}_d^{0,05}}{\widehat{\theta}_d^{0,50}} \quad ,$$

where:

$\widehat{\theta}_d^{0,95}$: 95^{th}-percentile of the distribution of $\widehat{\theta}_d$ in area d,

$\widehat{\theta}_d^{0,05}$: 5^{th}-percentile of the distribution of $\widehat{\theta}_d$ in area d, and

$\widehat{\theta}_d^{0,50}$: 50^{th}-percentile of the distribution of $\widehat{\theta}_d$ in area d (median).

4.3.3 Figures

The qualitative analysis is based on three types of figures:

- scatterplots of the point estimates
- Lorenz curves of the true and estimated composite indicator
- boxplots of the accuracy measures (`relative root MSE`, `relative bias`, and `relative dispersion`).

The figures are set up in the following way:

Scatterplots (cf. subsection 4.4.1):

For each area d the figures contrast the true (on x-axis) and the estimated (on y-axis) mean value of the composite indicator obtained over 10,000 independent drawn samples. Each point represents a small area. Ideally - in case of an unbiased estimation - the bisectrix should join all points, i.e. in each area d the true value and the estimated value of the composite indicator are congruent.

Lorenz Curves (cf. subsection 4.4.2):

Evaluation of point estimates by means of Lorenz Curves considers the background of the SAE methodology especially concerning disparities. As pointed out in previous sections, the estimation technique is commonly used for tabling indicators addressing e.g. poverty, income and other social key data in small geographical areas or socio-demographic groups. In this case, the statistician is highly interested whether estimates reproduce the disparity of the target variable in the base population.

The blue line in figures 4.4, 4.5, and 4.6 indicates the true disparity of the composite indicator (Lorenz Curve of the base population). It consists of $D = 581$ values (dots) each referring to a domain (small area) in the population. In order to illustrate the differences in true and estimated inequality of individual living conditions in detail, Lorenz Curves are plotted for 15 (of $D = 581$) randomly selected small areas, as well.

The single diagrams in figures 4.4 to 4.6 draw a comparison of Lorenz Curves obtained by the five small area estimators and the true disparity (blue line). Each diagram shows a green coloured bundle of 10,000 estimated Lorenz curves corresponding with one of the 10,000 independent drawn samples. Ideally, the bundle of green lines is i) small, indicating estimation with low variation and ii) congruent with the true value (blue coloured).

Boxplots (cf. subsection 4.4.3):

Accuracy measures - figured out for each estimator - are visualised using boxplots (cf. figures 4.8 to 4.9). An estimation of high quality is characterised by small values of RRMSE, RB, and RD. Basically, the graphics present the theoretical properties of the estimators applied. The direct estimators (HT and GREG) show - compared to other estimators - small relative bias but a much higher relative dispersion and relative root MSE. The indirect estimators demonstrate a higher RB but relatively small RD and RRMSE. Figures 4.8 to 4.9 show some selected results obtained by applying different auxiliary variable sets.

The following sections focus on graphical analysis of simulation results.

4.4 Selected results

4.4.1 Evaluation by means of scatterplots

A comparison of the true and the estimated values obtained by `NSM` (cf. figures 4.1 and 4.2) shows obviously strong discrepancies. The `NSM` turns out to be a poor estimator.

Due to its estimation technique, the NSM does not represent the area-specific pattern. The NSM is based on information related to the underlying population as a whole and not to small areas. However, one can see, that some small areas are estimated fairly well. In cases where congruence of the data structure can be assumed, that is, the small area is representative for the population, the NSM is a very simple and well working estimator. In this case, the mean (over 10.000 samples) population-based estimate of the target variable equals to the corresponding value of the small domain. In all other cases, the NSM yields unacceptable results: a high bias and a large variance (RRMSE cf. subsection 4.4.3).

The direct estimators HT and GREG as well as the indirect estimators SYNTH and EBLUP lead to results (dots) quite exactly on the angle bisector. So, figure 4.1 graphically confirms the statistical property of design unbiasdness of the direct estimators. Based on the graphical analysis even the indirect estimators seem to be quite unbiased. Obviously, the underlying model of the SYNTH is well chosen. As EBLUP is a method, which borrows unbiasdness by combining a direct estimator and a potential biased, but stable indirect estimator, the estimator performs well, too.

Some differences in estimation quality appear, when other auxiliary variable sets are applied (cf. figures 4.1 and 4.2). Estimates obtained using UT2 and R2 are worse compared to UT1 and R1, due to exclusion of variables `income` and `collecting unemployment benefits`. Auxiliary variable set EC performs worst. These results correspond to outcome of calibration estimation in subsections 3.4 et seq.

Estimation of sub-indicator `standard of living` (cf. figure 4.3) shows a similar result. Selection of sub-indicator `standard of living` is due to discussion on poverty and wealth in subsection 2.2.2. In addition, investigation of other sub-indicators provide no further insights in terms of estimation results, that is with respect to estimators and auxiliary variable sets applied. For this reason, results of other indicators are not illustrated. Furthermore, the influence of survey designs on the precision of point estimates is (almost) as discovered in chapter 3 (calibration estimation). In order to avoid lengthening of the report, these results are not illustrated, too.

As shown in the figures, the differences between the estimators applied turn out to be (at least graphically) small, except for NSM. This result may be affected by

- the underlying semi-synthetic dataset, which is free of non-response and which displays a homogenous data structure,
- the multitude of different auxiliary variables enhancing estimation quality, and
- the (maybe) relatively large sample sizes of $n = 14,100$ and $n = 8,250$ households.

Further insights into estimators' precision will be obtained by investigation of accuracy measures in subsection 4.4.3.

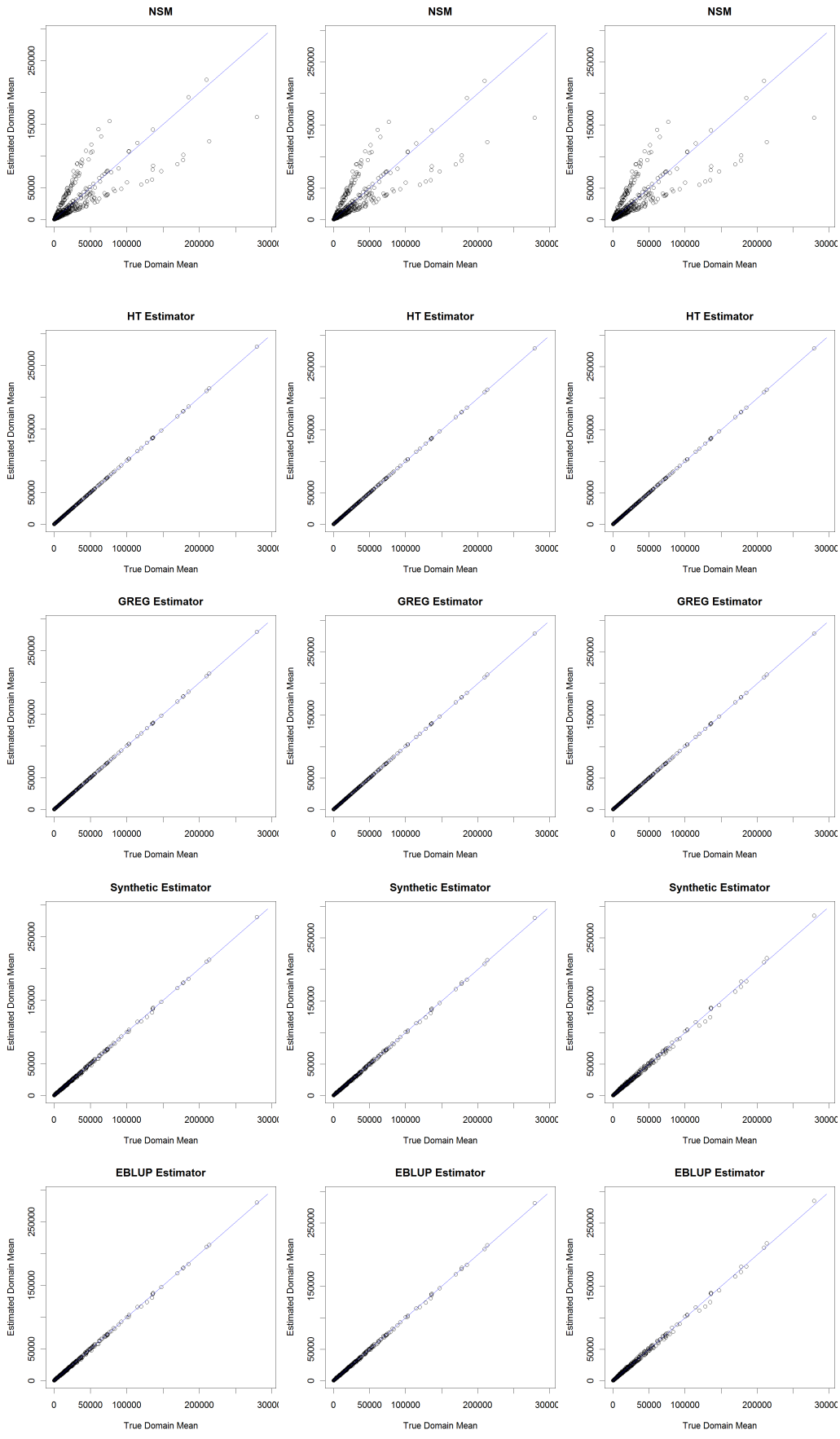

Figure 4.1: Scatterplots of `CIoILC`, sampled under `StratRS1` using auxiliary variable sets `NSI` (column 1), `UT1` (column 2), and `UT2` (column 3)

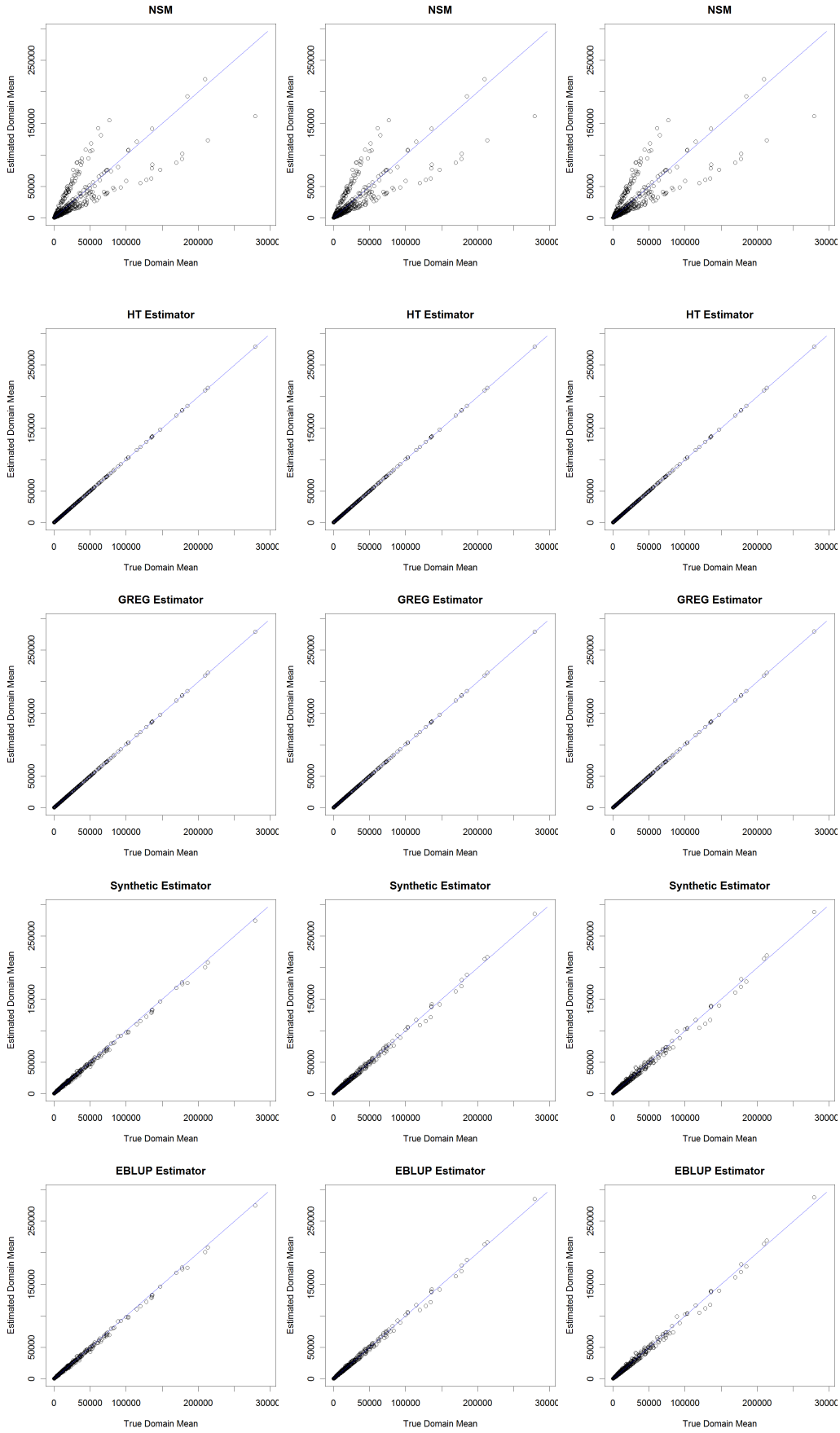

Figure 4.2: Scatterplots of `CIoILC`, sampled under `StratRS1` using auxiliary variable sets `R1` (column 1), `R2` (column 2), and `EC` (column 3)

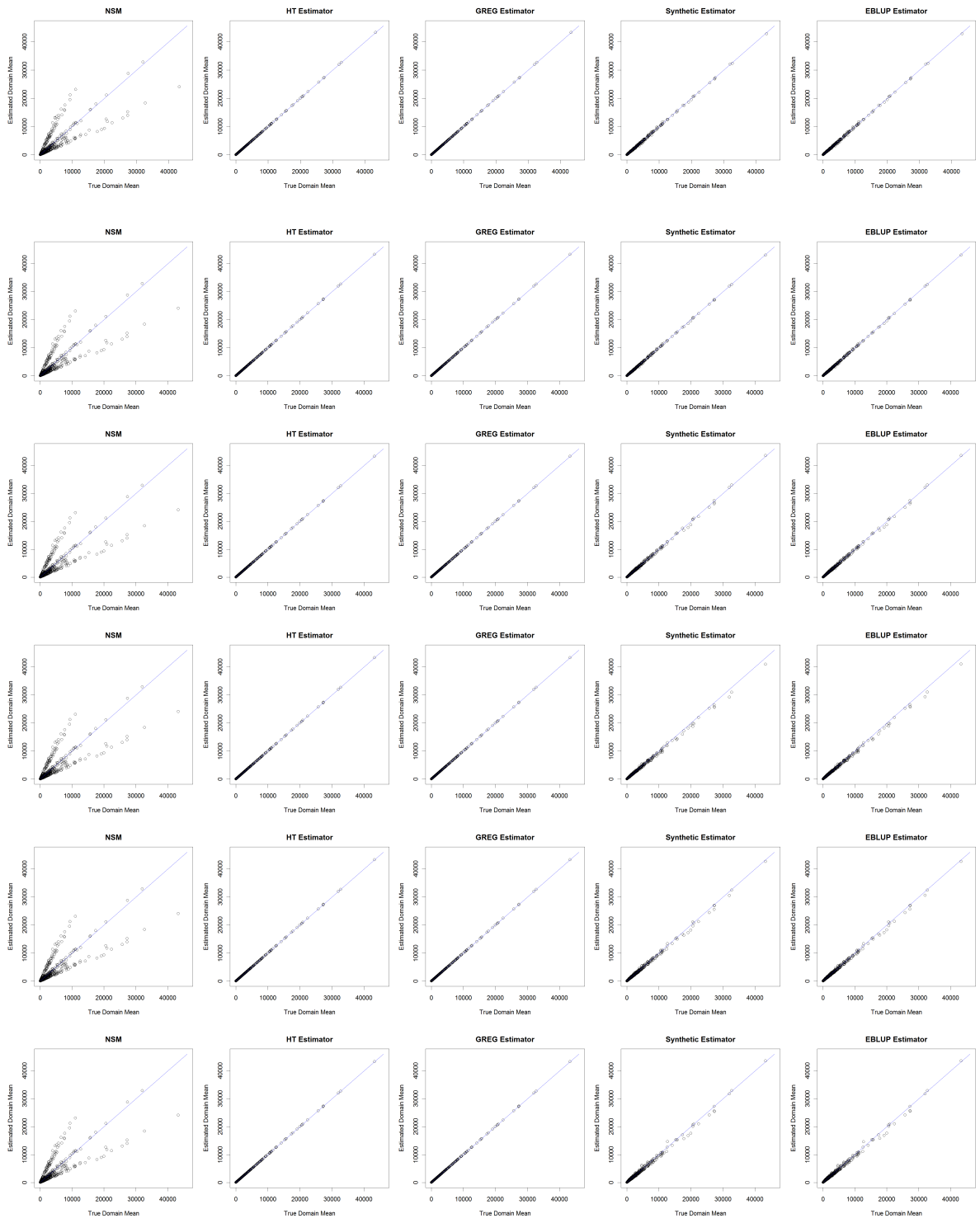

Figure 4.3: Scatterplots of `SISoL`, sampled under `StratRS1` using auxiliary variables sets `NSI`, `UT1`, `UT2`, `R1`, `R2`, and `EC` (line-by-line)

4.4.2 Evaluation by means of Lorenz Curves

Application of SAE is often associated with analysis of area-specific disparity, e.g. poverty or wealth, like `CIoILC` or `SISoL` in the present report. The graphical instrument of

Lorenz Curves is particularly suitable for analysing the grade of concentration. Based on estimation results presented in the previous subsection, the following figures show a selection of Lorenz Curves.

As pointed out in subsection 4.3.3, each diagram represents i) the true Lorenz Curve based on the population (blue line) and ii) the resulting bundle of 10,000 estimated Lorenz Curves (green coloured). Figure 4.4 displays the true and the estimated disparity of individual living conditions (CIoILC) observed for all $D = 581$ small areas. Estimated Lorenz Curves are contrasted for auxiliary variable sets `UT1` and `EC`. The estimated bundle of Lorenz Curves shown in the left column scatter less, due to the multitude of auxiliary variables entered in `UT1`.

Figures 4.5 and 4.6 display the discrepancies between the true and the estimated values in more detail, taking into account a random selection of 15 (out of 581) small areas. As was expected, `NSM` is not suitable for a precise estimation of the composite indicator's disparity. Using indirect estimators `SYNTH` and `EBLUP`, the bundle of Lorenz Curves is - due to a relative stable estimation - close to the true Lorenz Curve. Estimated Lorenz Curves disperse more when HT or GREG is applied. The accuracy of the resulting Lorenz Curves depends on the interaction of bias and efficiency of the estimator used.

Comparing figure 4.6 with 4.5, one can recognise an increasing width of the green bundle of Lorenz Curves, in particular for the indirect estimators. Furthermore, figure 4.8 shows an increasing bias using auxiliary variable set `EC`.

4.4.3 Evaluation by means of measures

The previous analysis dealt with graphical investigation of point estimators and shed light on estimation bias. This subsection focuses on a closer analysis of estimates by statistical measures presented in subsection 4.4.3: `Relative Dispersion`, `Relative Bias` and `Relative Root MSE`. The measures figured out for all estimators provide further insights to the interaction of target variable and auxiliary variable sets.

An overview on accuracy measures obtained for estimation of CIoILC is given in figure 4.8. `RB`, `RRMSE`, and `RD` are plotted column by column. Columns are subdivided block by block in terms of auxiliary variable sets (in order of UT1, UT2, R1, R2, and EC). For each auxiliary variable set measures are displayed as boxplots for the five small area estimators. Accuracy measures resulting for auxiliary variable set `NSI` are illustrated in figure 4.7.

As was expected, `UT1` is the most suitable set of ancillary information. Values of RB and RRMSE are obviously less high and scatter to a minor degree in case of GREG, `SYNTH` and `EBLUP`. Compared to `UT1`, `UT2` and `NSI`, the selection of top ten regressors - in terms of explaining the target variable - performs quite well. Estimates using auxiliary variable set `EC` are least precise.

For the `NSM` the resulting relative dispersion is low, but both, the relative bias and the relative root MSE, are the highest of all estimators used in the simulation study. That characteristic is due to the simple model structure of the `NSM` causing - amongst others -

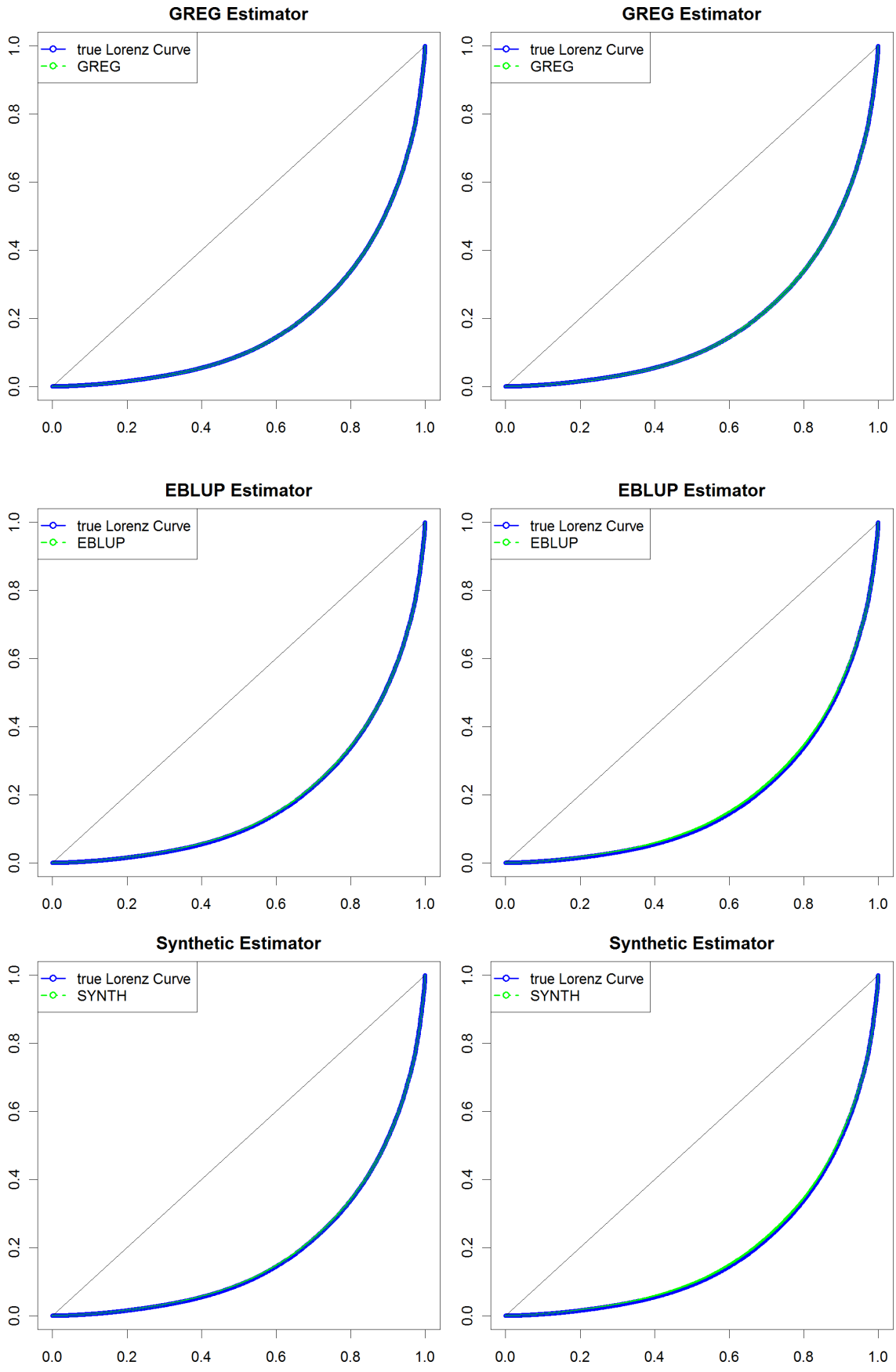

Figure 4.4: True and estimated Lorenz Curves of `CIoILC` for $D = 581$ areas, sampled under `StratRS1` using auxiliary variable sets `UT1` and `EC` (columns)

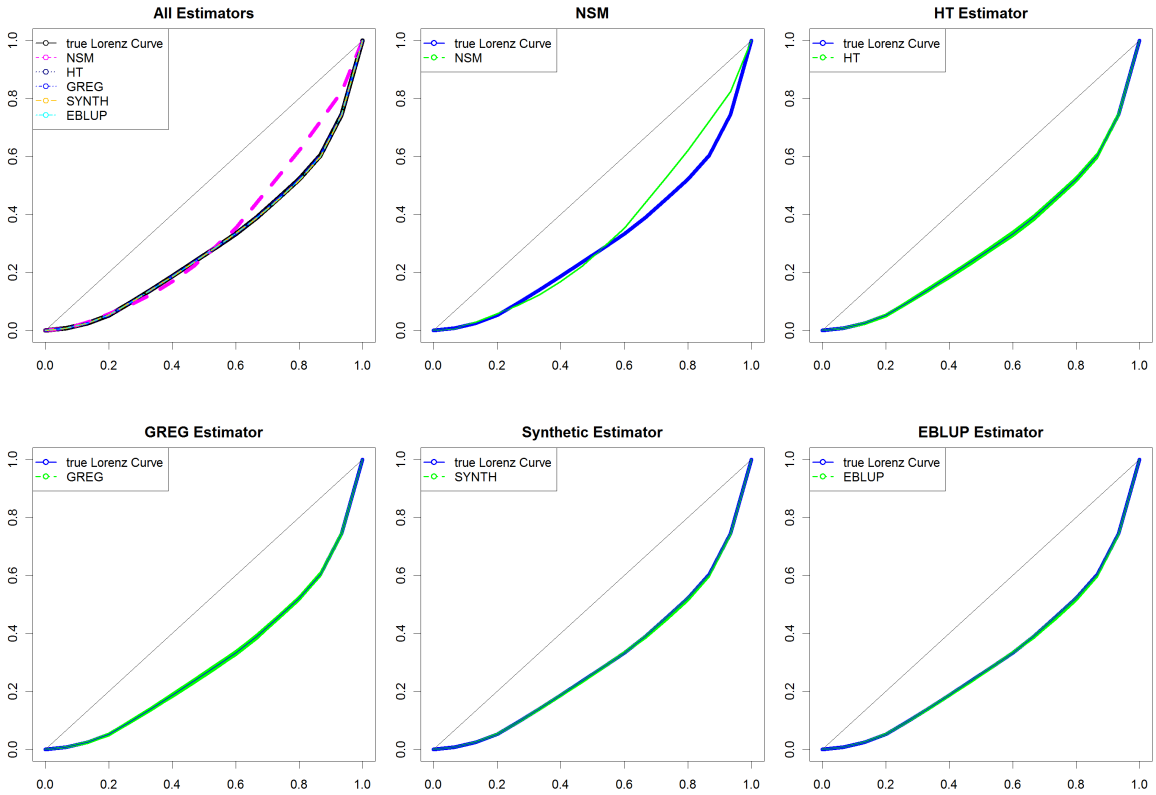

Figure 4.5: True and estimated Lorenz Curves of `CIoILC` for 15 selected areas, sampled under `StratRS1`, and estimated with `UT1`

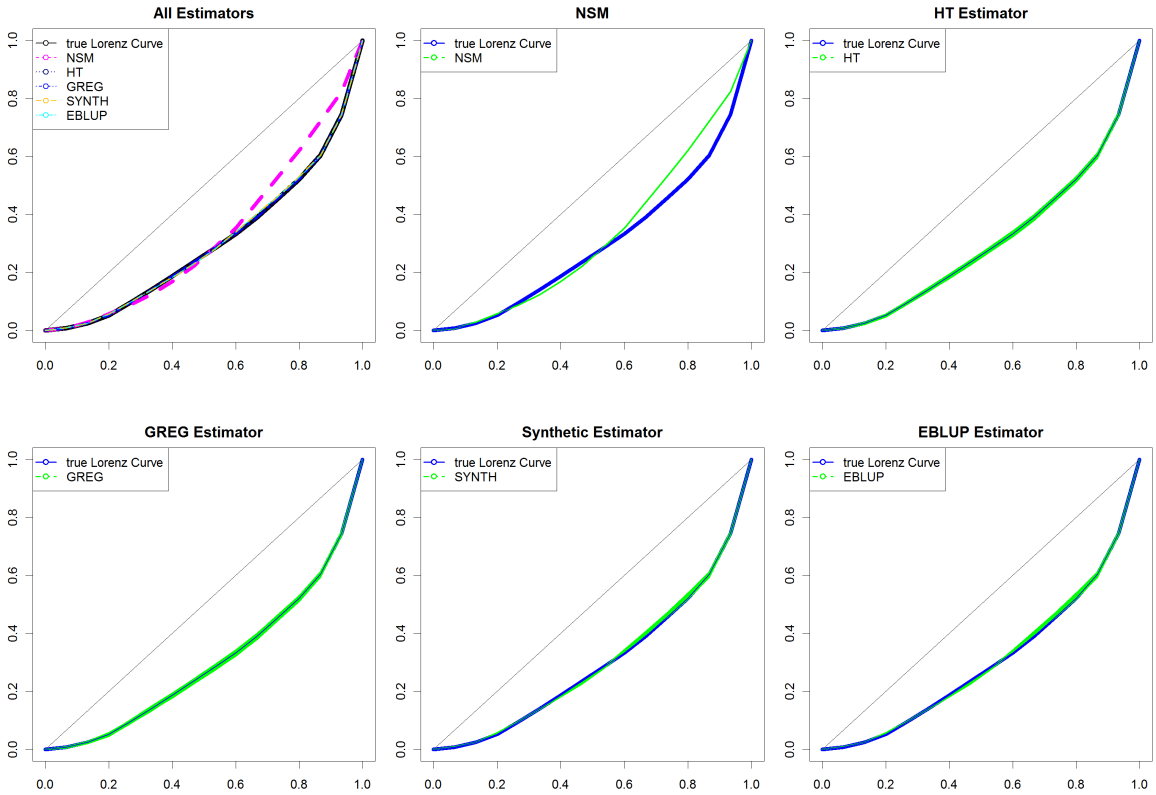

Figure 4.6: True and estimated Lorenz Curves of `CIoILC` for 15 selected areas, sampled under `StratRS1`, and estimated with `EC`

poor estimates. HT and GREG show a typical pattern: both of them are almost unbiased. Compared to the HT, the GREG being a model assisted estimator (use of ancillary information) has smaller variance and thus, a smaller relative rootMSE. The indirect estimators `SYNTH` and `EBLUP` are characterised by less dispersion, but - in comparison to direct estimators - an increasing bias. This result is consistent to estimation theory: Due to the reduction in variance of point estimates, `SYNTH` and `EBLUP` cause lower dispersion compared to direct estimators.

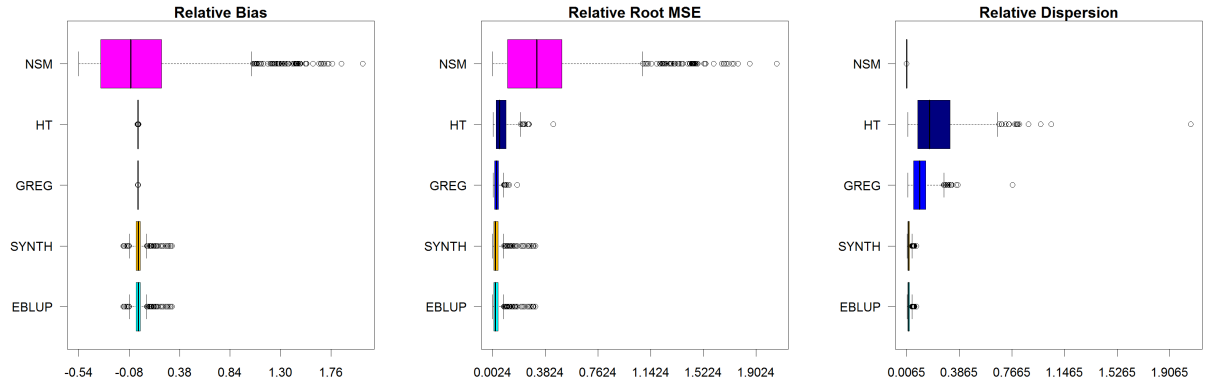

Figure 4.7: Measures of `CIoILC`, sampled under `StratRS1` using auxiliary variable set `NSI`

4.5 Summary

Based on the simulation study and on the results presented in the previous section, SAE is proved to be a suitable estimation technique and is applicable to estimating composite indicators. Estimating both, the composite indicator and sub-indicator `standard of living`, the theoretical characteristics of SAE are verified within the simulation study. As expected the direct estimator (`HT`) is design unbiased and accompanied by mostly higher variation, whereas the indirect estimators (`SYNTH` and `EBLUP`) yield biased estimates with relatively small variation. `GREG` - as model assisted estimator - performs quite well, taking both aspects into consideration.

Altogether, and in terms of the estimators applied, the methods except the NSM show relatively small differences. Reduction of sample size (less than 8,250 households) or investigation of only a few selected small areas would have yielded higher discrepancies, as indicated by figures 4.5 and 4.6. The `NSM` only yield good estimates, if the data pattern observed for the small area equals to that of the base population.

In the present simulation study, the strata defined by stratification variables `Federal state`, `type of household`, `social affiliation`, and `income` are used as small area categories ($D = 581$). That assumption and the restriction, that at least two sample units are selected per stratum ($n_d \geq 2$), can be modified, probably leading to higher differences between direct and indirect estimates, in particular if $n_d = 0$. Furthermore, investigation based on a (complete) real data set or a composite indicator based on single indicators which stem from different surveys, will probably have an impact on the estimators' performance.

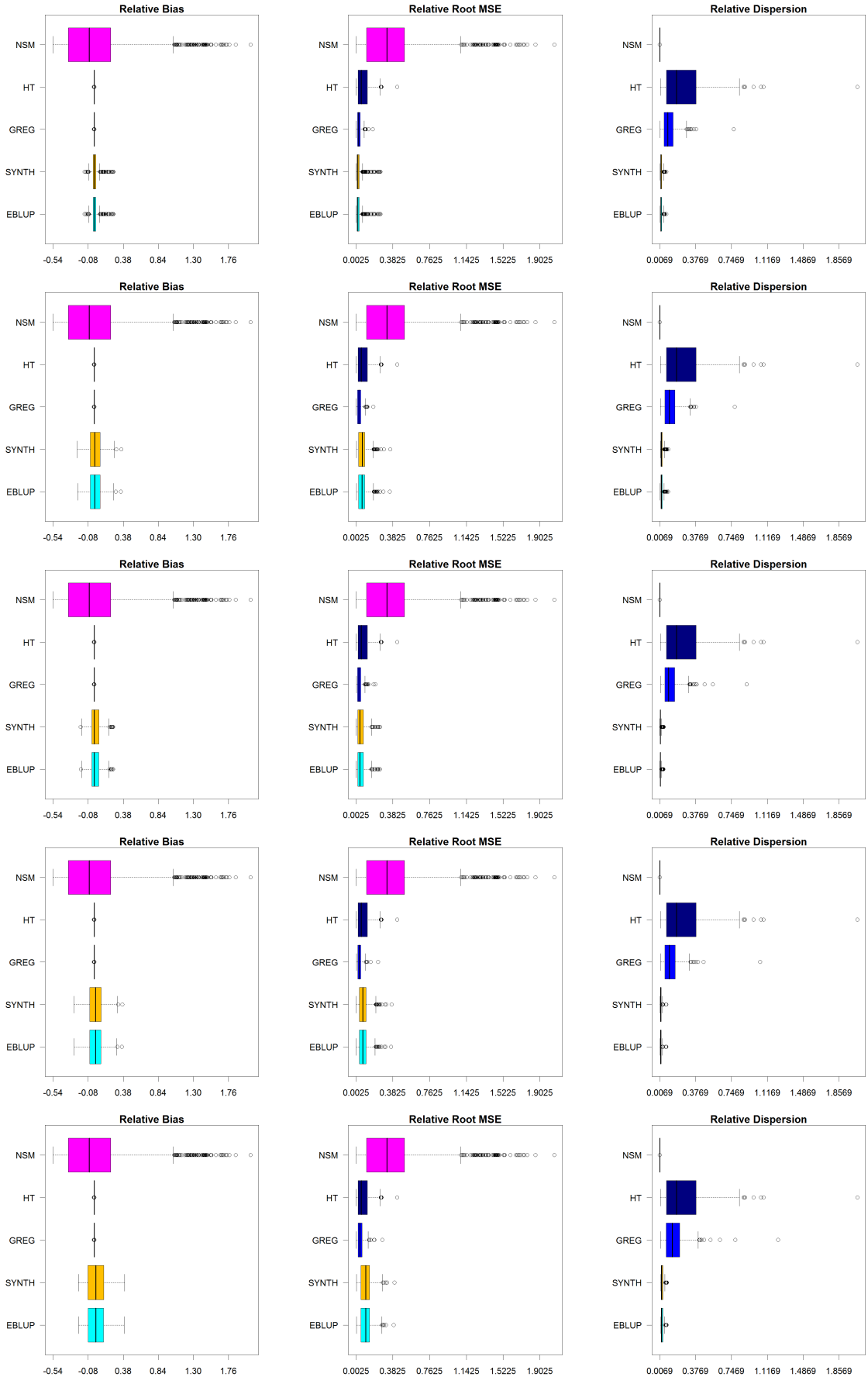

Figure 4.8: Measures of `CIoILC`, sampled under `StratRS1` using auxiliary variable sets `UT1` (top row), `UT2`, `R1`, `R2`, and `EC` (bottom row)

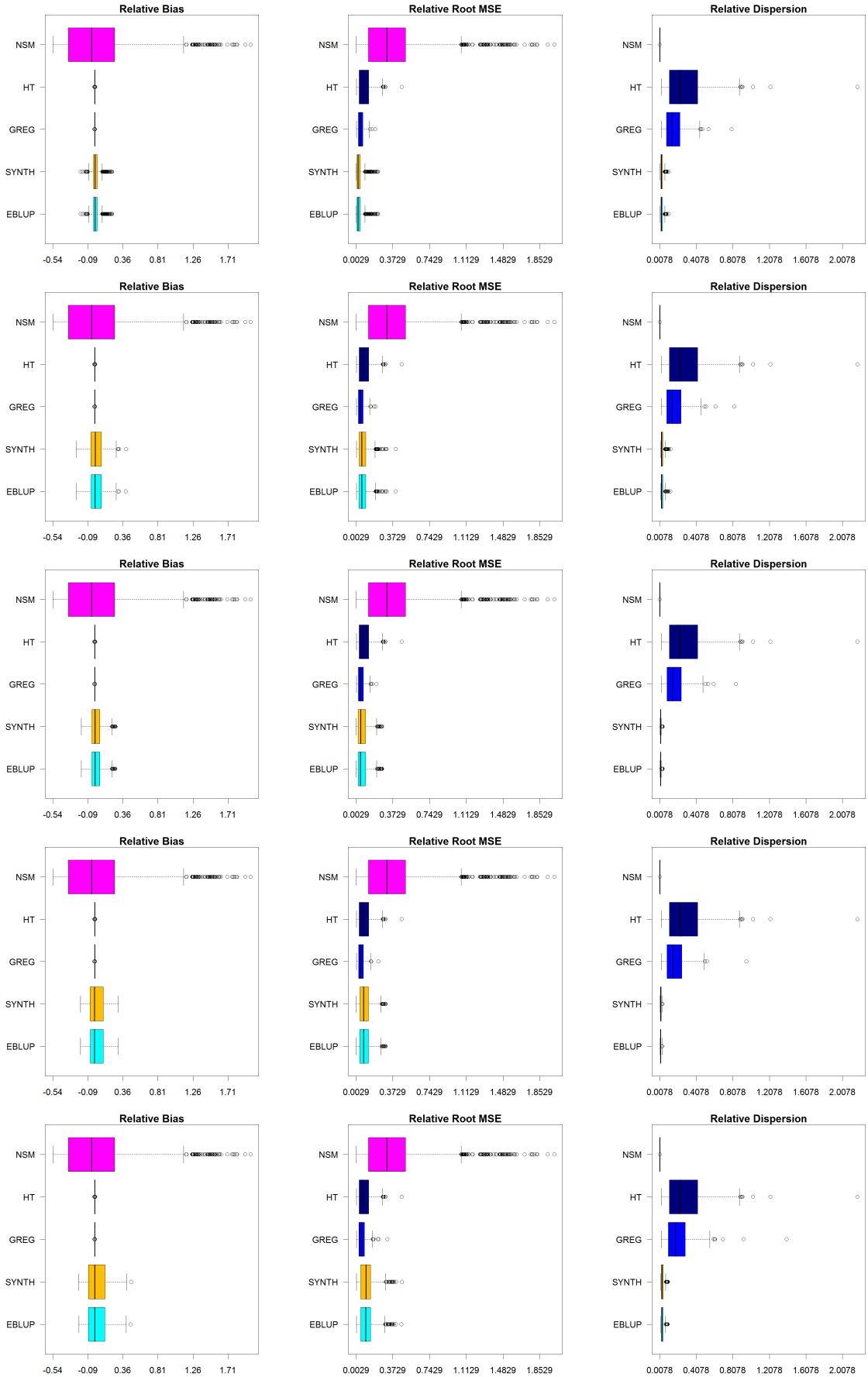

Figure 4.9: Measures of `SISoL`, sampled under `StratRS1` using auxiliary variable sets `UT1` (top row), `UT2`, `R1`, `R2`, and `EC` (bottom row)

First and foremost, the question of the data availability has to be addressed, regarding both the definition of the composite indicator and the auxiliary variables for model assisted (e.g. `GREG`) or model based (e.g. `SYNTH`) estimation. This includes investigation of data quality (cf. KEI deliverable 3.3 as well as investigations carried out in Appendices A and B of this report). Under predetermined assumption, the SAE approach turned out to be reasonably suitable for estimation of (composite) indicator, thus providing a valuable tool for policy making in small (geographic or socio-demographic) areas. Nevertheless, one has to point out that data availability and accessibility to microdata is vital to generate reliable composite indicators. The KBE is by far more sophisticated due to the use of data from many different sources. The end-user may be restricted to applying area-based models due to disclosure limitations.

Chapter 5

Summary and outlook

The preceding Sections have shown that improving the accuracy of single and composite indicators may become an important task in future applications, especially in policy making. Disclosure limitations may circumvent that end-users perform these estimation tasks on their own and especially the task of accuracy measurement. Despite the dispute over boon or bane of composite indicators, it seems sensible that National Statistical Institutes provide the necessary information on composite indicators which shall include improved accuracy of input values (single indicators) and adequate accuracy measures. This should help to gain better information for single and composite indicators for policy making.

The end-user will surely benefit from ongoing developments of meta data and quality report provision. As a major task within the wishlist (from the methodologist's view at least), more detailed country specific and variable specific information could be an important task for improving the outcome of research as input for policy making. However, it seems also necessary to teach the end-user in using this information adequately.

Within the simulation study some possible drawbacks become obvious such as data availability and the need of reliable auxiliary information. This was already the case for the ZUMA indicators which are derived from one single source of information. The case of KBE which was the main focus within the KEI project is still much more sophisticated since many data sources from many countries are to be used when building composite indicators on a larger scale.

Additionally, the present indicators were all linear. Especially the nonlinear Laeken indicators need further research also in the case of single indicators. Some of this new research tasks will be continued within the AMELI project, which forms part of the European Commission's 7th framework programme (cf. `http://ameli.surveystatistics.net`). This will bring both continuity and enhanced informativeness to the task, and challenge, of evaluating advanced indicators for policy making in Europe.

Appendix A

The semi-synthetic dataset

A.1 List of variables of interest

A full list of variables covered by the GMC is available in German on GESIS homepage under `www.gesis.org/en/services/data/official-microdata/microcensus/microcensus-grundfile`.

The list of EU-SILC target variables is covered by Commission regulation N^o 1983/2003 of 7th November 2003. The document is available via `EUR-Lex` providing a free access to European Union law:

`http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:298:0034:0085:EN:PDF`.

A complete list of ESS variables is downloadable from the ESS data homepage under `http://ess.nsd.uib.no/` with linkage to ESS rounds 1, 2, and 3. The latest list of variables is available under

`http://ess.nsd.uib.no/index.jsp?year=2007&country=&module=documentation`.

variable	
code	description
EF1	Federal state
EF3	Sample units
EF4	Consecutive number of household in sample unit
EF5	Consecutive number of person in household
EF28	Consecutive number of family in household
EF30	Age
EF32	Gender
EF35	Marital status
EF52	Citizenship
EF71	Attendance at school at present
EF74	Kind of school
EF110	Employment in reference week
EF127	Professional status
EF130	Public sector employment
EF134	Type of employment contract
EF138	Work part-time or full time
EF141	Contracted hours
EF214	Notified at the Federal Employment Office
EF215	Collecting unemployment benefits
EF258	Graduation yes/no
EF259	Highest level of education
EF260	Training qualification yes/no
EF261	Highest level of training qualification
EF338	What do you do for a living? / predominant livelihood
EF358	Accomodation allowance yes/no?
EF359	Benefit payments yes/no?
EF372	Net income per member of household
EF504	Employed (EU definition) yes/no?
EF505	Population in principal residence
EF506	Population in private households
EF507	Position in household (-> head of household)
EF509	Position in familiy (-> head of household)
EF521	Number of people in private household
EF539	Household net income
EF553	Marital status and cohabitation
EF627	Marital status and type of cohabitation
EF637	Number of children in cohabition aged 27 or over
EF640	Number of children in cohabition aged under 18
EF712	Size of building
EF750	Household expansion factor
EF751	Person expansion factor

Table A.1: List of GMC variables of interest

variable	
code	description
D-File	
DB030	Household ID
DB040	Region
DB075	Rotational group
DB080	Household design weight
DB090	Household cross-sectional weight
H-File	
HB030	Household ID
HB070	Person responding to the household questionnaire
HH030	Number of rooms available to the hosuehold
HH040	Leaking roof, damp walls/floors/foundation, or rot in window frames or floor
HH050	Ability to keep home adequately warm
HH080	Bath or shower in dwelling
HH090	Indoor flushing toilet for sole use of household
HS010	Arrears on mortgage or rent payments
HS020	Arrears on utility bills
HS040	Capacity to afford paying for one week annual holiday away from home
HS050	Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day
HS060	Capacity to face unexpected financial expenses
HS070	Do you have a telephone (including mobile phone)?
HS080	Do you have a colour TV?
HS090	Do you have a computer?
HS100	Do you have a washing machine?
HS110	Do you have a car?
HS120	Ability to make ends meet
HS140	Financial burden of the total housing cost
HS160	Problems with the dwelling: too dark, not enough light
HS170	Noise from neighbours or from the street
HS180	Pollution, crime or other environmental problems
HS190	Crime, violence or vadalism in the area
HY020	Total disponible household income
HY025	Within-household non-response inflation factor
HY050N	Familiy/children related allowences
HY060N	Social exclusion not elsewhere classified
HY070N	Housing allowences
HY120G	Regular taxes on wealth
HY130G	Regular inter-household cash transfer paid
R-File	
RB040	Household ID
RB030	Person ID

Table A.2: List of EU-SILC variables of interest (table 1)

variable	
code	description
P-File	
PB030	Person ID
PB070	Personel design weight for selected respondent
PB140	Year of birth
PB150	Sex
PB190	Marital status
PB220A	Citizenship 1
PE010	Current education activity
PE020	ISCED level currently attended
PE040	Highest ISCED level attained
PL040	Status in employment
PH010	General health
PH020	Suffer from any chronic (long-standing) illness or condition
PH030	Limitation in activities because of health problems
PH040	Unmet need for medical examination or treatment
PH060	Unmet need for dental examination or treatment
PY070N	Value of goods produced by own-consumption
PY080N	Pension from individual private plans

Table A.3: List of EU-SILC variables of interest (table 2)

variable		
code1	code2	description
A7	NETUSE	Personal use of internet/e-mail/www
B14	WRKPRTY	Worked in political party or action group last 12 months
B15	WRKORG	Worked in another organisation or association last 12 months
C2	SCLMEET	How often socially meet with friends, relatives or colleagues
C6	AESFDRK	Feeling of safety of walking alone in local area after dark
C8	BRGHMEF	Worry about home burgled has effect on quality of life
C26	CTZCNTR	Citizen of country
D53	WRINCO	Worried that income in old age will not be adequate to cover later years
E1	WKVLORG	Involved in work for voluntary or charitable organisations; at least once every six months
E12	FLTLNL	Felt lonely
E32	STFSDLV	Satisfied with standard of living
E45	FLCLPLA	Feel close to the people in local area
E48	STFJB	How satisfied with job
E49	STFJBOT	Satisfied with balance between time on job and time on other aspects
E51	JBSTRS	Find job stressful
E52	UEMPNYR	Become unemployed in the next 12 months, how likely
F1	HHMMB	Number of people living regularly as member of household
F2	GNDR	Gender
F3	YRBRN	Year of birth
F6	EDLVADE	Highest level of education, Germany
F8	UEMPLA	Doing last 7 days: Unemployed, actively looking for job
F8	UEMPLI	Doing last 7 days: Unemployed, not actively looking for job
F12	EMPLREL	Employment relation
F21	WKHTOT	Total hours normally worked per week in main job overtime included
F32	HINCTNT	Household's total net income, all sources
F62	MARITALA	Legal marital status
F72	MBLTPH	Personally have mobile telephone
Admin	regionde	Region, Germany

Table A.4: List of ESS variables of interest

A.2 Auxiliary variable sets

regressor			frequency		auxiliary varible set					
code	description	label	persons	households	UT1	UT2	R1	R2	NSI	EC
szh	social affiliation	self-employed	5.38%	6.61%	x	x			x	
		civil servant, employee, and apprentice	46.00%	46.72%	x	x			x	
		pensioner	26.88%	35.13%						
		other, or stay-at-home person	21.74%	11.54%	x	x		x	x	
hhtyp	type of household	single-person household	20.65%	36.78%	x	x			x	
		couple-household without children	33.07%	29.50%	x	x			x	
		single parent	2.96%	3.94%	x	x			x	
		couple-household with children	24.69%	18.48%	x	x			x	
		other households	18.62%	11.29%						
inc	household income	< 900 Euro	9.29%	14.75%					x	
		900 Euro - 1300 Euro	12.81%	17.21%					x	
		1300 Euro - 2600 Euro	43.60%	42.42%					x	
		2600 Euro - 3600 Euro	18.65%	14.54%					x	
		> 3600 Euro	15.66%	11.08%						
EF1.1	Federal state	Schleswig-Holstein, Mecklenburg-Western Pomerania	5.67%	5.67%						
EF1.2		Bremen, Hamburg, Berlin	7.38%	8.36%	x	x			x	x
EF1.3		Lower Saxony	9.72%	9.77%	x	x			x	x
EF1.4		North Rhine-Westphalia	19.53%	19.66%	x	x			x	x
EF1.5		Hesse	7.42%	7.30%	x	x			x	x
EF1.6		Rhineland-Palatinate, Saarland	6.04%	5.99%	x	x			x	x
EF1.7		Baden-Wuerttemberg	12.61%	12.29%	x	x			x	x
EF1.8		Bavaria	16.05%	15.71%	x	x			x	x
EF1.9		Brandenburg	3.32%	3.22%	x	x			x	x
EF1.10		Saxony, Saxony-Anhalt and Thuringia	12.24%	12.03%	x	x			x	x
EF30.1	Age	aged < 25 years	12.04%	4.07%	x	x		x	x	
EF30.2		aged 25-35 years	13.81%	13.22%						
EF30.3		aged 35-45 years	19.52%	20.31%	x	x		x	x	
EF30.4		aged 45-55 years	16.85%	17.60%	x	x		x	x	
EF30.5		aged 55-65 years	15.09%	16.24%	x	x		x	x	
EF30.6		aged > 65 years	22.70%	28.57%	x	x		x	x	
EF32	Gender	male	47.77%	67.59%	x	x		x	x	
EF35.1	Marital status	single	27.07%	23.08%	x	x	x	x	x	
EF35.2		married	57.27%	52.13%	x	x	x	x	x	
EF35.3		widowed	8.85%	14.37%	x	x	x	x	x	
EF35.4		divorced or separated	6.81%	10.42%						
EF52	Citizenship	German	93.60%	94.29%	x	x	x	x	x	
EF110.1	Employment	employed (working part- or full time)	51.38%	53.33%	x	x				
EF110.2		economically inactive	48.62%	46.67%						
EF130.1	Public sector employment	yes	9.94%	10.03%	x	x		x		
EF130.2		no	41.44%	43.30%						
EF130.0		non-response/ economically inactive	48.63%	46.67%						
EF138.1	Work part- or full time	full time	39.71%	46.69%	x	x	x	x		
EF138.8		part time	11.67%	6.64%	x	x	x			
EF138.0		not applicable (economically inactive)	48.62%	46.67%						
EF215.1	Collecting unemployment benefits	yes, Arbeitslosengeld	2.94%	3.25%	x		x			
EF215.2		yes, Arbeitslosenhilfe	2.73%	3.25%						
EF215.8		no	2.06%	1.79%						
EF215.9		non-response	0.01%	0.01%						
EF215.0		not applicable (e.g. children, not searching for work, not notified at the FEO)	92.26%	91.70%						
EF259.1	Highest level of education	CSE	41.74%	44.72%	x	x		x		
EF259.2		secondary scholl (GDR)	7.16%	7.17%	x	x		x		
EF259.3		secondary modern school certificate	18.70%	16.66%	x	x				
EF259.4		higher education entrance qualification	4.75%	5.60%						
EF259.5		general qualification for university entrance	16.05%	17.24%						
EF259.9		non-response	0.93%	0.94%						
EF259.0		scholar	10.68%	7.67%						
EF261.1	Highest level of training qualification	internship/ practical training	1.14%	1.20%						
EF261.2		vocational preparatory class	0.16%	0.15%						
EF261.3		apprentice	43.80%	45.87%						
EF261.4		vocational school	3.04%	2.95%	x	x				
EF261.5		foreman/ university of cooperative education	5.94%	7.72%	x	x	x	x		
EF261.6		professional school (GDR)	1.48%	1.45%	x	x				
EF261.7		advanced technical collage	0.65%	0.83%	x	x				
EF261.8/9		university	9.46%	11.62%	x	x	x	x		
EF261.10		PhD	1.18%	1.59%						
EF261.99		non-response	1.36%	1.50%						
EF261.0		not applicable (students)	31.78%	25.13%						

Table A.5: Auxiliary variable sets applied in simulation study (table 1)

regressor			frequency		auxiliary varible set					
code	description	label	persons	households	UT1	UT2	R1	R2	NSI	EC
EF338.1	What do you do for a living?	employed (working part- or full time)	46.83%	50.88%						
EF338.2		to be on dole	5.53%	6.44%						
EF338.3		to draw a pension	27.61%	36.13%						
EF338.4		support payments from parents/ dependants	16.72%	2.77%						
EF338.5		independatnt means, rental income, interest	0.48%	0.65%						
EF338.6		to be on welfare	1.49%	1.84%	x	x				
EF338.7		compulsory long term care insurance	0.08%	0.04%	x	x				
EF338.8		other support, e.g. student loan and grants	1.26%	1.25%	x	x				
EF358.1	Housing allowances	yes	2.89%	4.88%	x	x				
EF358.0		no	97.11%	95.12%						
EF359.1	Benefit payments	yes	2.03%	2.61%	x	x				
EF359.0		no	97.97%	97.39%						
EF521.1	Number of people in private household	1 person	20.65%	36.78%	x	x	x	x	x	x
EF521.2		2 persons	37.82%	34.74%	x	x	x	x	x	x
EF521.3		3 persons	18.83%	13.84%	x	x	x	x	x	x
EF521.4		4 persons	16.09%	10.79%	x	x	x	x	x	x
EF521.5		5 or more persons	6.61%	3.85%						
EF539.1	Household monthly net income	0-500 Euro	1.85%	3.10%	x		x			
EF539.2		500-900 Euro	7.43%	11.65%	x		x			
EF539.3		900-1300 Euro	12.81%	17.21%	x		x			
EF539.4		1300-2000 Euro	25.05%	26.29%	x					
EF539.5		2000-2600 Euro	18.55%	16.13%	x		x			
EF539.6		2600-3600 Euro	19.40%	14.54%	x		x			
EF539.7		3600-5000 Euro	10.18%	7.22%	x		x			
EF539.8		> 5000 Euro	5.47%	3.87%			x			
EF30 & EF32	Age cross sex (male)	aged < 16 years	0.00%	0.00%						x
		aged 16-20 years	2.73%	0.00%						x
		aged 20-25 years	3.34%	2.00%						x
		aged 25-30 years	3.12%	3.51%						x
		aged 30-35 years	3.73%	5.29%						x
		aged 35-40 years	4.81%	7.32%						x
		aged 40-45 years	4.96%	7.77%						x
		aged 45-50 years	4.32%	6.96%						x
		aged 50-55 years	3.97%	6.46%						x
		aged 55-60 years	3.31%	5.47%						x
		aged 60-65 years	4.05%	6.80%						x
		aged 65-70 years	3.79%	6.43%						x
		aged 70-75 years	2.46%	4.20%						x
		aged > 75	3.18%	5.38%						
	Age cross sex (female)	aged < 16 years	0.00%	0.00%						x
		aged 16-20 years	0.00%	0.00%						x
		aged 20-25 years	5.96%	2.06%						x
		aged 25-30 years	3.21%	2.19%						x
		aged 30-35 years	3.75%	2.23%						x
		aged 35-40 years	4.81%	2.59%						x
		aged 40-45 years	4.94%	2.63%						x
		aged 45-50 years	4.39%	2.20%						x
		aged 50-55 years	4.17%	1.99%						x
		aged 55-60 years	3.40%	1.69%						x
		aged 60-65 years	4.34%	2.28%						x
		aged 65-70 years	4.23%	2.69%						x
		aged 70-75 years	3.04%	2.55%						x
		aged > 75	5.99%	7.31%						
		Number of regressors:			56	48	21	21	34	39

Table A.6: Auxiliary variable sets applied in simulation study (table 2)

regressor			frequency		auxiliary varible set									
code	description	label	persons	households	SI1 R1	SI1 R2	SI2 R1	SI2 R2	SI3 R1	SI4 R1	SI5 R1	SI6 R1	SI6 R2	SI7 R1
szh	social affiliation	self-employed	5.38%	6.61%		x								
		civil servant, employee, and apprentice	46.00%	46.72%										
		pensioner	26.88%	35.13%										
		other, or stay-at-home person	21.74%	11.54%										x
hhtyp	type of household	single-person household	20.65%	36.78%										
		couple-household without children	33.07%	29.50%										
		single parent	2.96%	3.94%		x								
		couple-household with children	24.69%	18.48%										
		other households	18.62%	11.29%										
EF1.1	Federal state	Schleswig-Holstein, Mecklenburg-Western Pomerania	5.67%	5.67%										
EF1.2		Bremen, Hamburg, Berlin	7.38%	8.36%										
EF1.3		Lower Saxony	9.72%	9.77%										
EF1.4		North Rhine-Westphalia	19.53%	19.66%										
EF1.5		Hesse	7.42%	7.30%										
EF1.6		Rhineland-Palatinate, Saarland	6.04%	5.99%										
EF1.7		Baden-Wuerttemberg	12.61%	12.29%										
EF1.8		Bavaria	16.05%	15.71%										
EF1.9		Brandenburg	3.32%	3.22%					x					
EF1.10		Saxony, Saxony-Anhalt and Thuringia	12.24%	12.03%										
EF30.1	Age	aged < 25 years	12.04%	4.07%							x			x
EF30.2		aged 25-35 years	13.81%	13.22%										
EF30.3		aged 35-45 years	19.52%	20.31%										
EF30.4		aged 45-55 years	16.85%	17.60%										
EF30.5		aged 55-65 years	15.09%	16.24%		x								
EF30.6		aged > 65 years	22.70%	28.57%		x								
EF32	Gender	male	47.77%	67.59%					x		x		x	
EF35.1	Marital status	single	27.07%	23.08%			x	x	x		x	x	x	x
EF35.2		married	57.27%	52.13%			x	x	x		x	x	x	x
EF35.3		widowed	8.85%	14.37%			x	x	x		x	x	x	x
EF35.4		divorced or separated	6.81%	10.42%										
EF52	Citizenship	German	93.60%	94.29%		x		x		x	x	x	x	
EF110.1	Employment	employed (working part- or full time)	51.38%	53.33%				x	x					
EF110.2		economically inactive	48.62%	46.67%										
EF130.1	Public sector employment	yes	9.94%	10.03%										
EF130.2		no	41.44%	43.30%										
EF130.0		non-response/ economically inactive	48.63%	46.67%										
EF138.1	Work part- or full time	full time	39.71%	46.69%		x								
EF138.8		part time	11.67%	6.64%										
EF138.0		not applicable (economically inactive)	48.62%	46.67%										
EF215.1	Collecting unemployment benefits	yes, Arbeitslosengeld	2.94%	3.25%										
EF215.2		yes, Arbeitslosenhilfe	2.73%	3.25%										x
EF215.8		no	2.06%	1.79%										
EF215.9		non-response	0.01%	0.01%										
EF215.0		not applicable (e.g. children, not searching for work, not notified at the FEO)	92.26%	91.70%										
EF259.1	Highest level of education	CSE	41.74%	44.72%		x				x				
EF259.2		secondary scholl (GDR)	7.16%	7.17%										
EF259.3		secondary modern school certificate	18.70%	16.66%		x								
EF259.4		higher education entrance qualification	4.75%	5.60%										
EF259.5		general qualification for university entrance	16.05%	17.24%										
EF259.9		non-response	0.93%	0.94%										
EF259.0		scholar	10.68%	7.67%										
EF261.1	Highest level of training qualification	internship/ practical training	1.14%	1.20%										
EF261.2		vocational preparatory class	0.16%	0.15%										
EF261.3		apprentice	43.80%	45.87%										
EF261.4		vocational school	3.04%	2.95%										
EF261.5		foreman/ university of cooperative education	5.94%	7.72%						x				
EF261.6		professional school (GDR)	1.48%	1.45%						x				
EF261.7		advanced technical collage	0.65%	0.83%						x				
EF261.8/9		university	9.46%	11.62%				x		x		x	x	
EF261.10		PhD	1.18%	1.59%										
EF261.99		non-response	1.36%	1.50%										
EF261.0		not applicable (students)	31.78%	25.13%										

Table A.7: Auxiliary variable sets specified using R routine regsubsets (table 1)

regressor			frequency		auxiliary varible set									
code	description	label	persons	households	SI1 R1	SI1 R2	SI2 R1	SI2 R2	SI3 R1	SI4 R1	SI5 R1	SI6 R1	SI6 R2	SI7 R1
EF338.1	What do you do for a living?	employed (working part- or full time)	46.83%	50.88%										
EF338.2		to be on dole	5.53%	6.44%										
EF338.3		to draw a pension	27.61%	36.13%										
EF338.4		support payments from parents/ dependants	16.72%	2.77%										
EF338.5		independatnt means, rental income, interest	0.48%	0.65%										
EF338.6		to be on welfare	1.49%	1.84%		x								
EF338.7		compulsory long term care insurance	0.08%	0.04%										
EF338.8		other support, e.g. student loan and grants	1.26%	1.25%										
EF358.1	Housing allowances	yes	2.89%	4.88%										
EF358.0		no	97.11%	95.12%										
EF359.1	Benefit payments	yes	2.03%	2.61%										
EF359.0		no	97.97%	97.39%										
EF521.1	Number of people in private household	1 person	20.65%	36.78%	x		x	x	x	x	x	x	x	x
EF521.2		2 persons	37.82%	34.74%	x		x	x	x	x	x	x	x	x
EF521.3		3 persons	18.83%	13.84%	x		x	x	x	x	x	x	x	x
EF521.4		4 persons	16.09%	10.79%		x	x	x	x	x	x	x	x	x
EF521.5		5 or more persons	6.61%	3.85%										
EF539.1	Household monthly net income	0-500 Euro	1.85%	3.10%	x									
EF539.2		500-900 Euro	7.43%	11.65%	x									
EF539.3		900-1300 Euro	12.81%	17.21%	x									
EF539.4		1300-2000 Euro	25.05%	26.29%	x									
EF539.5		2000-2600 Euro	18.55%	16.13%										
EF539.6		2600-3600 Euro	19.40%	14.54%	x		x							
EF539.7		3600-5000 Euro	10.18%	7.22%	x		x					x		
EF539.8		> 5000 Euro	5.47%	3.87%	x		x							
		Number of regressors:			10	2	7	4	4	4	4	5	4	4

Table A.8: Auxiliary variable sets specified using R routine regsubsets (table 2)

A.3 Frequency distribution of variables

regressor			EU-SILC frequency		GMC frequency		difference persons		difference households	
code	description	label	Persons	Households	Persons	Households	absolute	relative	absolute	relative
EF1.1	Federal state	Schleswig-Holstein, Mecklenburg-Western Pom.	5.49%	5.67%	5.67%	5.67%	-0.18%	-3.13%	0.00%	-0.03%
EF1.2		Bremen, Hamburg, Berlin	5.99%	6.45%	7.38%	8.36%	-1.39%	-18.85%	-1.91%	-22.83%
EF1.3		Lower Saxony	9.03%	9.07%	9.72%	9.77%	-0.69%	-7.13%	-0.70%	-7.13%
EF1.4		North Rhine-Westphalia	21.31%	21.01%	19.53%	19.66%	1.77%	9.07%	1.34%	6.83%
EF1.5		Hesse	7.56%	7.52%	7.42%	7.30%	0.14%	1.95%	0.22%	2.99%
EF1.6		Rhineland-Palatinate, Saarland	6.26%	6.16%	6.04%	5.99%	0.21%	3.50%	0.18%	2.98%
EF1.7		Baden-Wuerttemberg	12.85%	12.76%	12.61%	12.29%	0.23%	1.84%	0.47%	3.78%
EF1.8		Bavaria	15.96%	15.89%	16.05%	15.71%	-0.10%	-0.60%	0.18%	1.16%
EF1.9		Brandenburg	3.31%	3.29%	3.32%	3.22%	-0.01%	-0.42%	0.07%	2.13%
EF1.10		Saxony, Saxony-Anhalt and Thuringia	12.26%	12.18%	12.24%	12.03%	0.01%	0.10%	0.15%	1.26%
EF30.1	Age	aged < 25 years	6.97%	1.19%	12.04%	4.07%	-5.07%	-42.10%	-2.87%	-70.62%
EF30.2		aged 25-35 years	9.48%	8.71%	13.81%	13.22%	-4.32%	-31.30%	-4.51%	-34.09%
EF30.3		aged 35-45 years	24.11%	26.10%	19.52%	20.31%	4.59%	23.51%	5.79%	28.51%
EF30.4		aged 45-55 years	21.87%	22.74%	16.85%	17.60%	5.02%	29.79%	5.14%	29.21%
EF30.5		aged 55-65 years	17.55%	18.19%	15.09%	16.24%	2.46%	16.32%	1.95%	11.97%
EF30.6		aged > 65 years	20.02%	23.07%	22.70%	28.57%	-2.68%	-11.82%	-5.50%	-19.25%
EF32	Gender	male	46.53%	42.57%	47.77%	67.59%	-1.24%	-2.59%	-25.02%	-37.02%
EF35.1	Marital status	single	18.91%	16.76%	27.07%	23.08%	-8.15%	-30.13%	-6.32%	-27.39%
EF35.2		married	65.27%	58.20%	57.27%	52.13%	8.00%	13.96%	6.07%	11.64%
EF35.3		widowed	4.88%	7.38%	8.85%	14.37%	-3.97%	-44.85%	-6.99%	-48.63%
EF35.4		divorced or separated	10.94%	17.66%	6.81%	10.42%	4.13%	60.60%	7.24%	69.54%
EF52	Citizenship	German	98.14%	98.58%	93.60%	94.29%	4.53%	4.84%	4.28%	4.54%
EF110	Employment	employed (working part- or full time)	51.93%	49.36%	51.38%	53.33%	0.55%	1.07%	-3.97%	-7.45%
EF127.1	Professional status	self-employed with employees	2.89%	2.62%	4.52%	5.97%	-1.63%	-36.11%	-3.36%	-56.22%
EF127.2		self-employed without employees	5.11%	5.21%	5.15%	6.08%	-0.04%	-0.79%	-0.87%	-14.39%
EF127.3		employee	91.30%	91.49%	89.53%	87.60%	1.77%	1.98%	3.89%	4.44%
EF127.4		familiy workers	0.70%	0.69%	0.79%	0.35%	-0.10%	-12.00%	0.34%	98.74%
EF358	Housing allowances	yes	5.53%	7.20%	2.89%	4.88%	2.64%	91.47%	2.32%	47.58%
EF359	Benefit payments	yes	2.40%	3.04%	2.03%	2.61%	0.37%	18.30%	0.44%	16.70%
EF521.1	Number of people in private household	1 person	14.13%	25.03%	20.65%	36.78%	-6.53%	-31.61%	-11.75%	-31.95%
EF521.2		2 persons	35.47%	35.03%	37.82%	34.74%	-2.35%	-6.22%	0.29%	0.85%
EF521.3		3 persons	22.61%	18.59%	18.83%	13.84%	3.79%	20.11%	4.75%	34.29%
EF521.4		4 persons	20.47%	16.07%	16.09%	10.79%	4.38%	27.24%	5.28%	48.92%
EF521.5		5 or more persons	7.32%	5.28%	6.61%	3.85%	0.71%	10.75%	1.43%	37.09%
EF539.1	Household monthly net income	0-500 Euro	0.68%	1.08%	1.85%	3.10%	-1.18%	-63.60%	-2.02%	-65.03%
EF539.2		500-900 Euro	4.09%	6.75%	7.43%	11.65%	-3.34%	-44.94%	-4.90%	-42.04%
EF539.3		900-1300 Euro	7.29%	10.56%	12.81%	17.21%	-5.51%	-43.05%	-6.65%	-38.63%
EF539.4		1300-2000 Euro	17.04%	20.75%	25.05%	26.29%	-8.01%	-31.96%	-5.54%	-21.08%
EF539.5		2000-2600 Euro	17.59%	17.20%	18.55%	16.13%	-0.96%	-5.20%	1.07%	6.63%
EF539.6		2600-3600 Euro	23.20%	20.45%	18.65%	14.54%	4.55%	24.40%	5.91%	40.67%
EF539.7		3600-5000 Euro	18.28%	14.36%	10.18%	7.22%	8.10%	79.58%	7.15%	99.06%
EF539.8		> 5000 Euro	11.76%	8.75%	5.47%	3.87%	6.29%	114.83%	4.89%	126.49%

Table A.9: Relative frequency distribution of EU-SILC regressors in datasets (table 1)

regressor			frequency [%]		difference	
code	description	label	ESS	GMC	absolute	relative
EF1.1	Federal state	Schleswig-Holstein, Mecklenburg-Western Pom.	7.97%	5.67%	2.30%	40.63%
EF1.2		Bremen, Hamburg, Berlin	10.04%	7.38%	2.66%	36.01%
EF1.3		Lower Saxony	6.83%	9.72%	-2.89%	-29.71%
EF1.4		North Rhine-Westphalia	16.25%	19.53%	-3.28%	-16.80%
EF1.5		Hesse	5.38%	7.42%	-2.04%	-27.44%
EF1.6		Rhineland-Palatinate, Saarland	4.76%	6.04%	-1.28%	-21.21%
EF1.7		Baden-Wuerttemberg	8.85%	12.61%	-3.76%	-29.83%
EF1.8		Bavaria	11.75%	16.05%	-4.30%	-26.81%
EF1.9		Brandenburg	5.12%	3.32%	1.80%	54.18%
EF1.10		Saxony, Saxony-Anhalt and Thuringia	23.03%	12.24%	10.79%	88.13%
EF30.1	Age	aged < 25 years	5.43%	12.04%	-6.60%	-54.84%
EF30.2		aged 25-35 years	14.49%	13.81%	0.69%	4.98%
EF30.3		aged 35-45 years	22.20%	19.52%	2.68%	13.75%
EF30.4		aged 45-55 years	20.08%	16.85%	3.24%	19.20%
EF30.5		aged 55-65 years	16.25%	15.09%	1.16%	7.70%
EF30.6		aged > 65 years	21.53%	22.70%	-1.17%	-5.14%
EF32	Gender	male	50.93%	47.77%	3.17%	6.63%
EF35.1	Marital status	single	22.36%	27.07%	-4.71%	-17.39%
EF35.2		married	59.73%	57.27%	2.46%	4.30%
EF35.3		widowed	7.97%	8.85%	-0.88%	-9.95%
EF35.4		divorced	9.94%	6.81%	3.13%	45.91%
EF52	Citizenship	German	96.84%	93.60%	3.24%	3.46%
EF258.8	Graduation	left school without school leaving qualification	1.86%	2.32%	-0.46%	-19.83%
EF259.1	Highest level of education	CSE	32.25%	41.74%	-9.49%	-22.74%
EF259.3		secondary modern school certificate	34.11%	25.86%	8.25%	31.92%
EF259.4		higher education entrance qualification	13.35%	20.80%	-7.44%	-35.79%
EF261.8		Technical college	7.19%	3.80%	3.39%	89.20%
EF261.9		university degree	11.13%	5.66%	5.47%	96.56%
EF504	Employed	unemployed	9.32%	7.07%	2.24%	31.73%
EF521.1	Number of people in private household	1 person	23.65%	20.65%	3.00%	14.53%
EF521.2		2 persons	39.49%	37.82%	1.67%	4.42%
EF521.3		3 persons	18.27%	18.83%	-0.56%	-2.95%
EF521.4		4 persons	13.61%	16.09%	-2.48%	-15.41%
EF521.5		5 or more persons	4.97%	6.61%	-1.64%	-24.79%
EF539.1	Household monthly net income	< 150€	0.41%	0.10%	0.31%	315.01%
EF539.2		150-300€	0.88%	0.33%	0.55%	168.19%
EF539.3		300-500€	1.81%	1.43%	0.38%	26.96%
EF539.4		ESS: 500-1000€; GMC: 500-1100€	11.59%	13.25%	-1.66%	-12.52%
EF539.5		ESS: 1000-1500€; GMC: 1100-1500€	16.56%	14.56%	2.00%	13.75%
EF539.6		1500-2000€	18.43%	17.48%	0.95%	5.44%
EF539.7		ESS: 2000-2500€; GMC: 2000-2600€	16.67%	18.55%	-1.88%	-10.15%
EF539.8		ESS: 2500-3000€; GMC: 2600-3200€	11.96%	12.75%	-0.80%	-6.25%
EF539.9		ESS: 3000-5000€; GMC: 3200-5000€	14.70%	16.08%	-1.38%	-8.56%
EF539.10		5000-7500€	5.12%	4.09%	1.03%	25.23%
EF539.11		7500-10000€	1.35%	0.77%	0.58%	74.84%
EF539.12		> 10000€	0.52%	0.61%	-0.10%	-15.51%

Table A.10: Relative frequency distribution of ESS regressors in datasets (table 2)

variable				frequency		difference	
code	description	value	labels	original	generated	absolute	relative
HS040	Capacity to afford paying for one week annual holiday	1	yes	78.66%	76.84%	-1.82%	-2.31%
HS050	Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day	1	yes	89.11%	87.63%	-1.48%	-1.66%
HS060	Capacity to face unexpected financial expenses	1	yes	77.54%	74.12%	-3.42%	-4.41%
HS090	Do you have a computer?	1	yes	72.81%	66.47%	-6.34%	-8.70%
HS100	Do you have a washing machine?	1	yes	96.70%	94.66%	-2.05%	-2.12%
HS110	Do you have a car?	1	yes	82.90%	77.10%	-5.80%	-6.99%
HS120	Ability to make ends meet	1	with great difficulty	3.02%	3.48%	0.46%	15.40%
		2	with difficulty	7.28%	8.42%	1.14%	15.63%
		3	with some difficulty	34.40%	38.40%	4.00%	11.63%
		4	fairly easily	27.53%	26.57%	-0.96%	-3.47%
		5	easily, very easily	27.77%	23.12%	-4.65%	-16.73%
HS140	Financial burden of the total housing cost	1	a heavy burden	22.18%	22.21%	0.03%	0.15%
		2	somewhat a burden	59.27%	59.21%	-0.06%	-0.11%
		3	not burden at all	18.55%	18.58%	0.03%	0.16%
HH050	Ability to keep home adequately warm	1	yes	99.71%	99.68%	-0.04%	-0.04%
HH030	Number of rooms available to the hosuehold	1	1 room	2.28%	4.59%	2.31%	101.29%
		2	2 rooms	14.99%	21.31%	6.32%	42.17%
		3	3 rooms	27.14%	29.98%	2.85%	10.49%
		4	4 rooms	24.02%	21.48%	-2.54%	-10.56%
		5	5 rooms	16.64%	12.47%	-4.17%	-25.05%
		6	6 rooms	8.40%	5.83%	-2.57%	-30.58%
		7	7 or more rooms	6.53%	4.34%	-2.20%	-33.65%
PH010	General health	1	very good	11.36%	10.93%	-0.44%	-3.84%
		2	good	48.32%	47.02%	-1.30%	-2.68%
		3	fair	31.64%	32.72%	1.08%	3.41%
		4	bad or very bad	8.67%	9.33%	0.65%	7.51%
PH020	Suffer from any chronic (long-standing) illness or	1	yes	38.81%	40.32%	1.51%	3.90%
PH030	Limitation in activities because of health problems	1	yes, strongly limited	7.56%	8.03%	0.48%	6.31%
		2	yes, limited	28.92%	30.01%	1.09%	3.79%
		3	no, not limited	63.53%	61.96%	-1.57%	-2.47%

Table A.11: Frequency distribution of original and generated EU-SILC variables

variable					frequency		difference	
code 1	code 2	description	value	labels	original	generated	absolute	relative
A7	NETUSE	Personal use of internet/ e-mail/ www	1	yes	56.83%	54.24%	-2.59%	-4.56%
B14 & B15	*membership*	Worked in political party or action group and/ or worked in another organsiation last 12 months	1	yes	23.24%	23.70%	0.46%	1.97%
C2	SCLMEET	How often socially meet with friends, relatives or colleagues	1	less than once a month or never	8.49%	8.44%	-0.05%	-0.54%
			2	once a month or several times a month	57.45%	56.54%	-0.91%	-1.59%
			3	once a week or several times a week	34.06%	35.02%	0.96%	2.82%
C6	AESFDRK	Feeling of safety of walking alone in local area after dark	1	yes	77.02%	75.86%	-1.16%	-1.51%
C8	BRGHMEF	Worry about home burgled has effect on quality of life	1	yes	67.14%	67.73%	0.59%	0.88%
D53	WRINCO	Worried that income in old age will not be adequate to cover later years	1	not worried, not worried at all	17.34%	17.72%	0.38%	2.21%
			2	neutral (3-7 points of 10)	45.24%	46.03%	0.79%	1.74%
			3	worried or extremely worried	37.42%	36.25%	-1.17%	-3.13%
E1	WKVLORG	Involved in work for voluntary or charitable organisations; at least once every six months	1	yes	36.96%	34.84%	-2.12%	-5.73%
E12	FLTLNL	Felt lonely	1	yes	71.64%	71.43%	-0.20%	-0.28%
E32	STFSDLV	Satisfied with standard of living	1	dissatisfied or extremely dissatisfied	9.83%	9.53%	-0.31%	-3.12%
			2	neutral (4-6 points of 10)	27.43%	28.51%	1.08%	3.94%
			3	satisfied or extremely satisfied	62.73%	61.96%	-0.77%	-1.23%
E45	FLCLPLA	Feel close to the people in local area	1	yes	61.75%	59.80%	-1.95%	-3.16%
F72	MBLTPH	Personally have mobile telephone	1	yes	78.57%	76.30%	-2.27%	-2.89%

Table A.12: Frequency distribution of original and generated ESS variables

Appendix B

The semi-synthetic dataset

B.1 Data Quality in generated population

B.1.1 Comparison EU-SILC and GMC

Deviations in the frequency distributions are found on the personal as well as on the household level. These discrepancies are due, in part, to the fact that EU-SILC 2005 comprises only 25% of GMC households. Some 75% of the households were taken from the current economic calculations (`Laufende Wirtschaftrechungen`) and also the `German sample survey on income and expenditure` (EVS) as a quota sample. With a growing number of GMC households, increased correspondence in the dataset structures is only to be expected.

Deviations in household structure may also be due to what is generated by the household head, who had to be identified - in the absence of having been designated in the database. Another point to note is that data lines deleted as a result of the non-response item may cause the frequency distributions to change somewhat.

- German Microcensus: 42,733 (8.55%) data lines
- EU-SILC: 2,029 (8.21%) data lines
- ESS: 984 (33.75%) data lines.

According to table A.9 persons under the age of 35 (EF30.1 and EF30.2) are under-represented in EU-SILC, especially those under 25. On the whole, more persons between the ages of 35 and 65 are represented. The over-65 component (EF30.6) is smaller in EU-SILC than in GMC. These differences in the age spread are due, in part, to the fact that EU-SILC characteristics of interest were only available for those persons and households that had sent in a questionnaire. The targetted population only takes in persons over the age of 15.

Singles (EF35.1) and widowed persons (EF35.3) are under-represented over and against married couples and spouses living apart. The number of one-person households (EF521.1)

is too low, while households of three or more persons (EF521.3-.5) are more strongly represented. Households with incomes under Euro 2,000 (EF539.1-.4) are under-represented. A high relative deviation from GMC is posted for households and persons with incomes under Euro 500 as well as those with incomes between Euro 500 and 1300, while a preponderance of high-income households is represented in EU-SILC, especially households with incomes under Euro 3,600 (EF539.7 and EF539.8). The number of the self-employed and corporate households (EF127.1 and EF127.2) is lower in EU-SILC than in GMC.

Deviations found in the frequency distributions are reflected in the correlation tables, which underscore the differences between EU-SILC and GMC. Tables B.1 und B.2 show the correlation structures pertaining between regressors for the household and personal dataset. Tabular fields that are not colour-coded designate correlation coefficients below +- 0.1, i.e. they point to a very weak correlative linkage between the variables. The resulting differences for the correlation coefficients are set out in table B.3.

Equivalent correlation structures are available for the Federal state characteristic. More or less large deviations are found for the relationships between age, gender, marital status, household size, employment, wage earners, and income. Differences are posted above all for the following regressors: Persons under 25 and above 65 (EF30.1 and EF30.6), gender (EF32), marital status: single and married (EF35.1 and EF35.2), employment (EF110), wage earners (EF 127.3), single-person households and families with children (EF521.1 and EF521.3-5) as well as all income classes.

On the whole, the correlation structures show more deviations on the person level than in the household datasets.

Result:

Measured in terms of frequency distributions, the EU-SILC variables can be implemented with differential degrees of quality. The maximal absolute deviation is 6.3% and the maximal relative deviation 101%, posted for categories 1 and 2 respectively of the variable HH030 (number of rooms per household). HH030 can-in contrast to the original frequency distribution-be implemented only with fairly large deviations (cf. table A.9). Especially the peripheral classes show sizeable deviations. There are more households with 1 and 3 rooms represented. This is at the expense of the characteristic classes of 5, 6, or 7 and more rooms.

In GMC, more households are represented that can handle month's end only with (a certain amount of) financial difficulty (HS120.1-.3). The number of households able to shoulder unexpected financial burdens (HS060) is likewise lower. Fewer households own a computer (HS090) or a car (HS110). The number of these is, respectively, 8.7% and 6.9% lower (in relative terms) than in the case of the original dataset (absolute deviation: 6.3% and 5.8% respectively).

The number of persons who (subjectively judged) are in poor health (PH010.4), suffer from chronic diseases (PH020) or have restricting disabilities (PH030.1) is somewhat larger in each of these cases.

Based on univariate frequencies, household availability of goods is lower and average health slightly worse than in the original EU-SILC dataset. Apart from HH030, HS090, HS110 and HS120, the frequencies are largely conformant with the original dataset material.

To be noted in this context is that both the correlation coefficients and the regressors entered into the Logit models are statistically significant. To the extent that deviations obtain between original and generated frequency distributions for EU-SILC and ESS characteristics, this is primarily due to differential patterning of datasets, as can be derived indeed from the correlation tables or restricted frequency distributions.

Table B.6 shows for the EU-SILC characteristics of interest the deviations found in the correlation coefficients. The selected regressors generate, on the whole, only a small number of differences (see colour-coded fields). Deviations are mainly apparent for the following regressors:

- Persons under 25 (EF30.1): correlates with HH030.1 and the person-related characteristics PH010.1, PH020, and PH030.
- Sex (EF32): correlates with HS090, HS100, HS110 and HH030.2.
- Wage earners (EF127.1): correlates with HS040, HS090, HS110, HH30.7; also with the heath-related variables PH010, PH020 and PH030.
- Households with incomes between Euro 1,300 and 2,000 (EF539.4): correlates with HS040, HS050, HS090, HS110 and HH030.2.

What cannot be represented are the relationships between individual EU-SILC characteristics. This holds especially for the relationships between:

- HS120 and HS140
- H010, PH020, and PH030
- each, HS040, HS050, and HS060 with HS120 and HS140.

Tables B.4 and B.5 contrasts the correlations per variable as found between the original and the semi-synthetic dataset (columns original and gen.). For the regressors (listed line-by-line) we mostly get identical colour patterns per variable. Deviations in the generated EU-SILC variables are evident in the bottom third of the table.

B.1.2 Comparison ESS and GMC:

The frequency distributions of the regressors diverge from those for GMC (cf. table A.10). The dataset includes only 2,916 lines, with 984 having to be excluded on grounds of item non-response. Also, the ESS is not designed as a GMC sub-sample. Given all this, the discrepancies are unsurprising.

A maximal absolute difference of 10.79% (in relative terms: 88.13%) is found for the variable EF1.10 (Federal states of Saxony, Saxony-Anhalt and Thuringia). Generally speaking, the western Federal states are under-represented, while persons in eastern Germany are over-represented. As with EU-SILC, the percentage of the population under 25 years of age is too low compared with GMC. Single households and spouses living apart

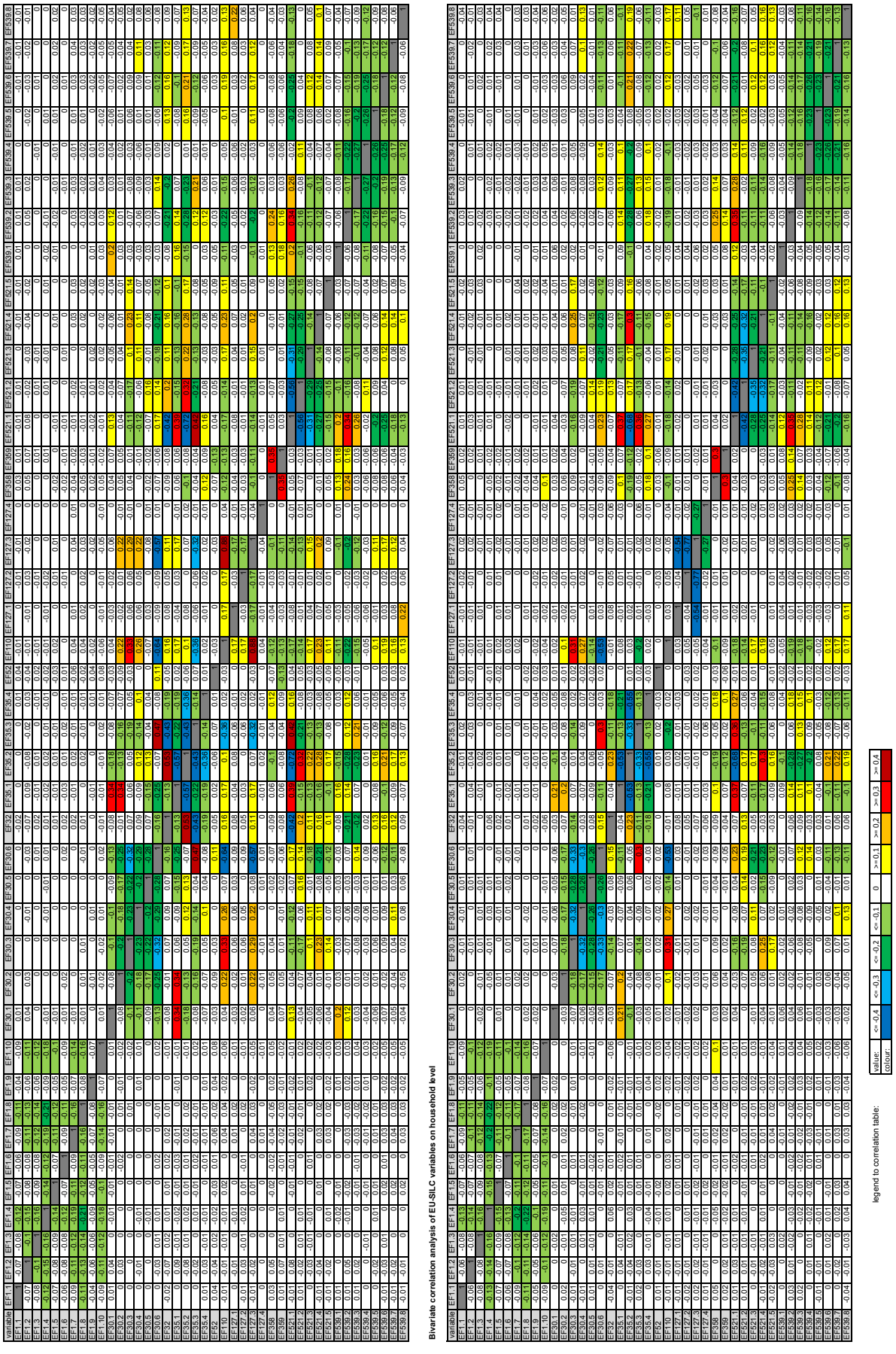

Table B.1: Bivariate correlation coefficients of EU-SILC and GMC household data

Table B.2: Bivariate correlation coefficients of EU-SILC and GMC person data

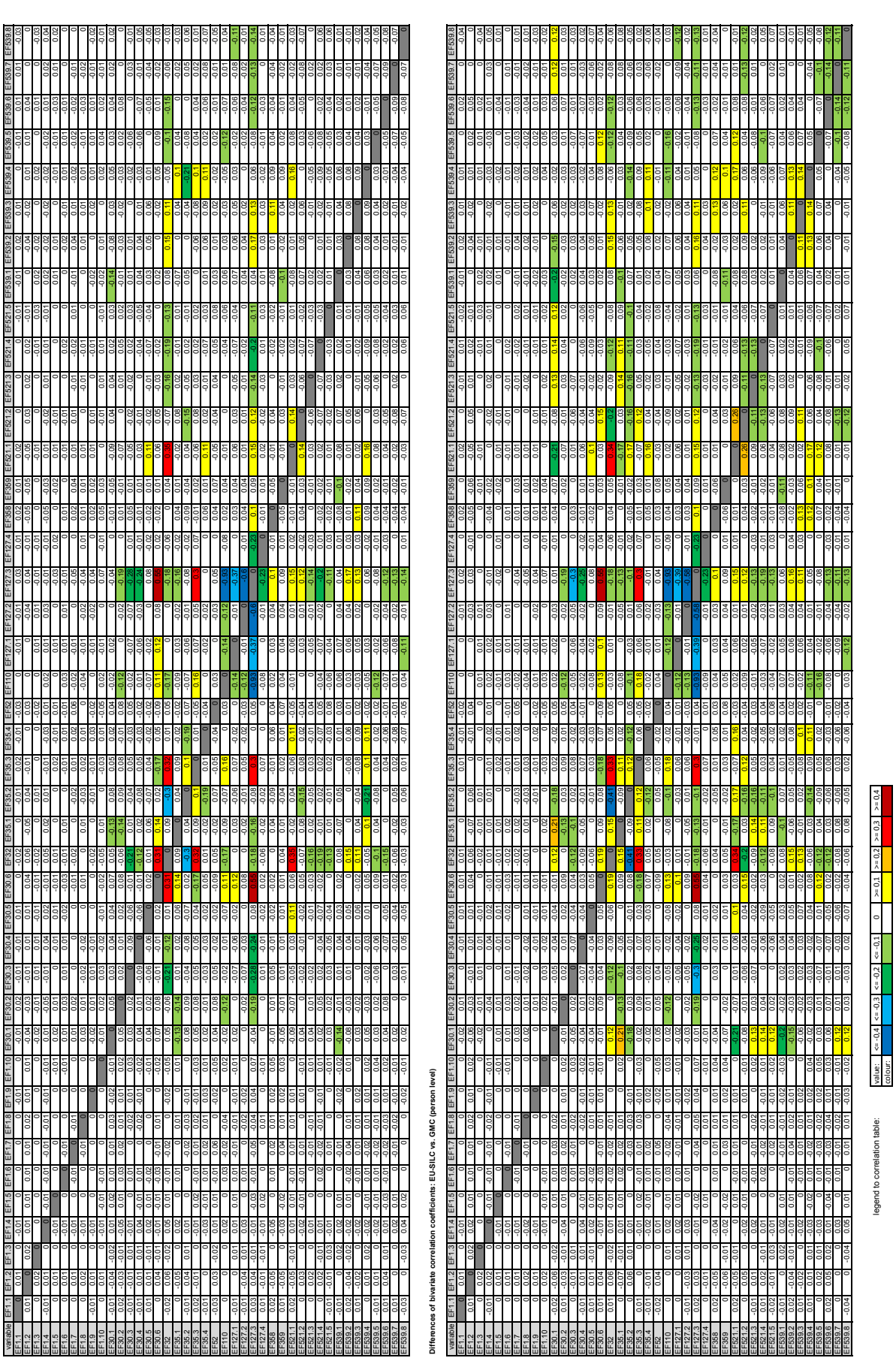

Table B.3: Differences of bivariate correlation coefficients EU-SILC and GMC data

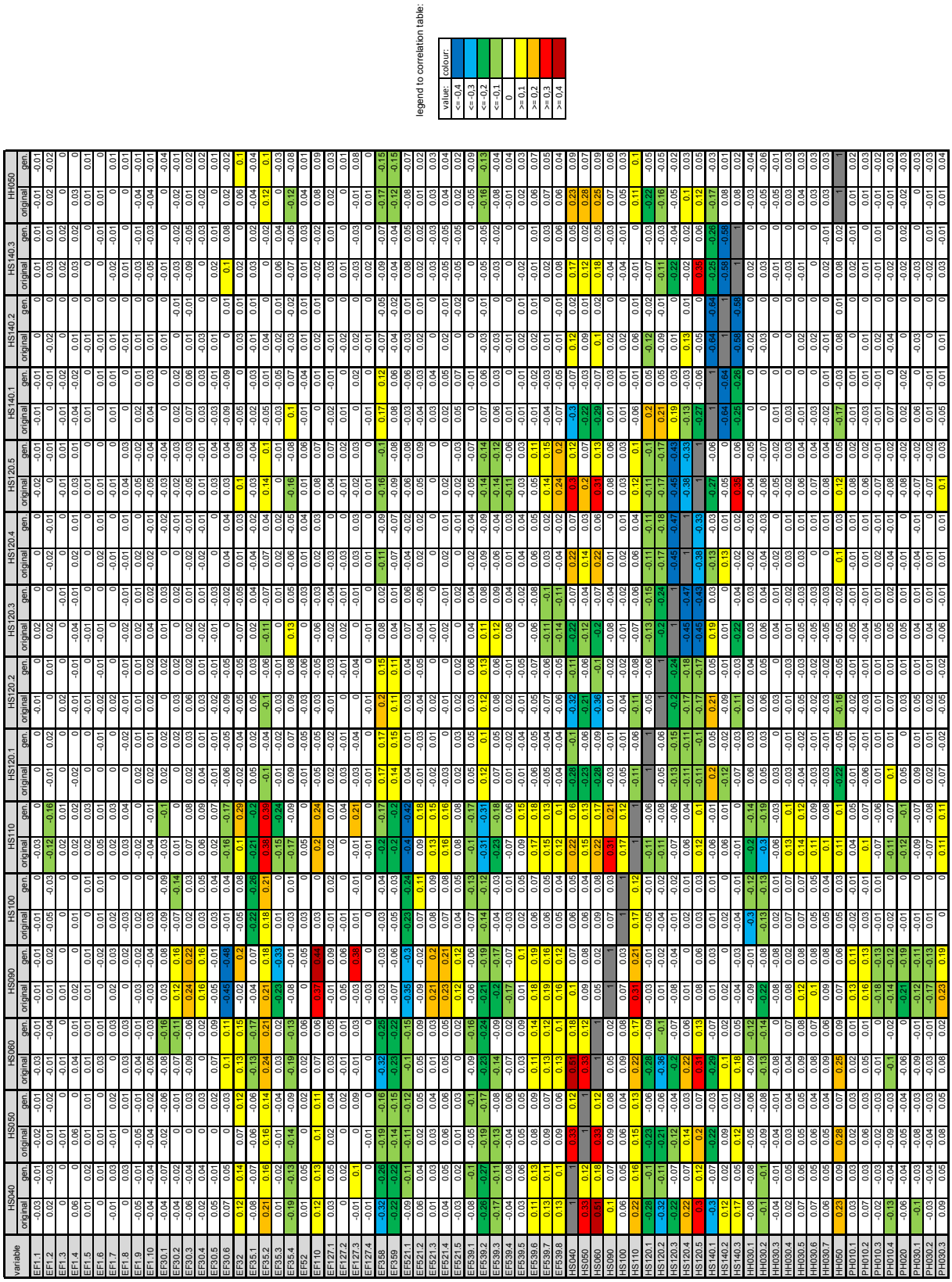

Table B.4: Pairwise comparison of bivariate correlation coefficients of original and generated EU-SILC variables (table 1)

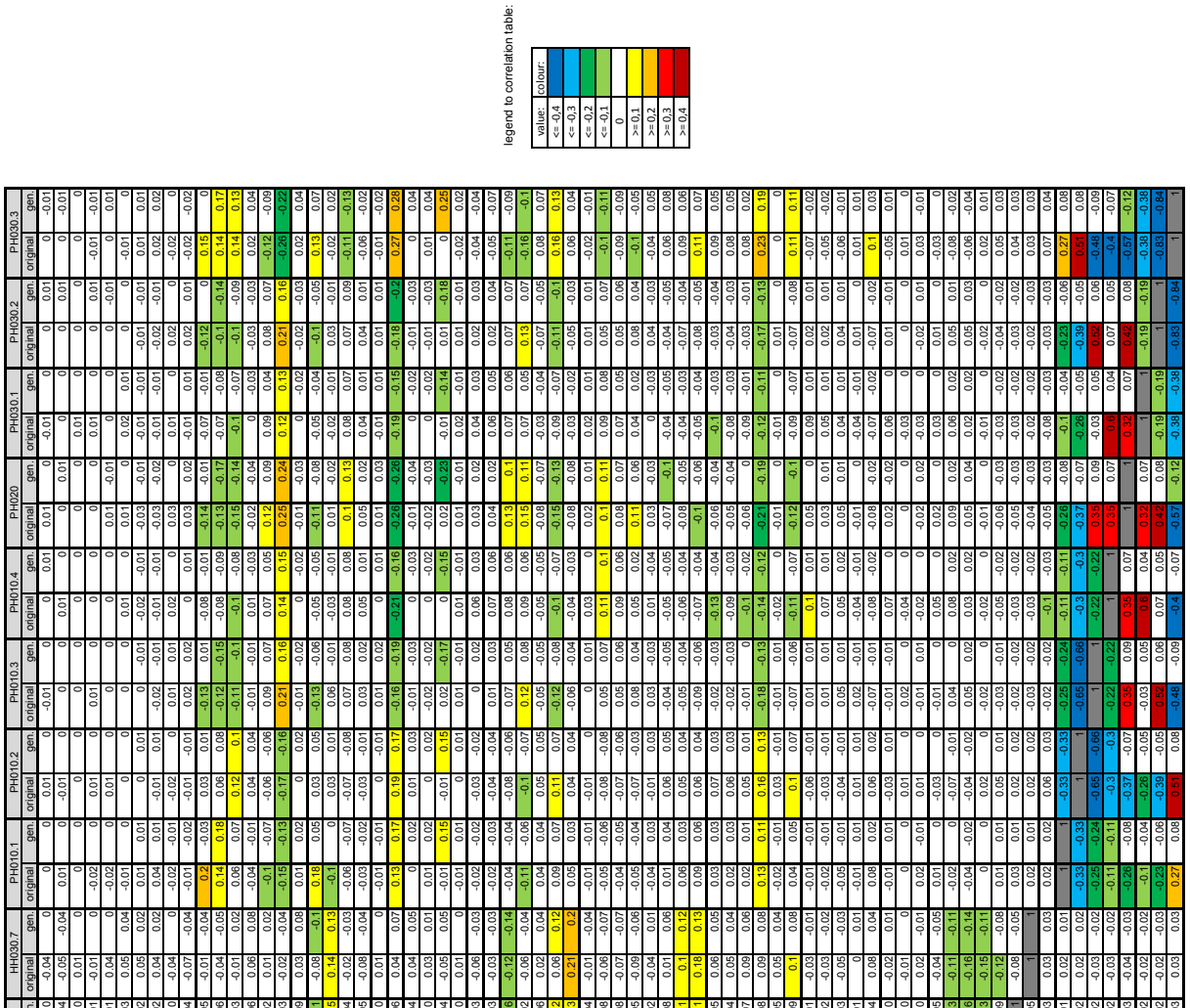

Table B.5: Pairwise comparison of bivariate correlation coefficients of original and generated EU-SILC variables (table 2)

Table B.6: Differences in bivariate correlation coefficients of original and generated EU-SILC variables

are insufficiently represented. Differences are chiefly found in the frequency distribution for educational qualifications. There are more persons in the ESS who have graduated from comprehensive schools, technical colleges or universities. The income distribution differs in the case of the lower two income classes. Persons with incomes below Euro 500 are more strongly represented in the ESS, but still only represent approx. 1% of all persons in the dataset. The absolute deviation lies between 0.31 and 0.55 percentage points, while the relative deviation lies - owing to the small share of the population involved - between 26.96% and 315%.

Comparison of the correlation matrices (regressors) of GMC and ESS reveals a large correlative similarity between the datasets. Major deviations in the correlation coefficients are set out in table B.8, chiefly for the following regressors:

- Correlation between EF30.1 (persons under 25), EF35.1 and EF35.2 (marital status single or married), EF259.1 and EF259.4 (CSE and Abitur [German school leaver's certificate conferring right to tertiary study]), and EF521.2 and EF521.4 (two- and four-person households)
- Correlation between EF258.8 (school leaver's certificate), EF30.6 (persons over 65), EF35.3 (widowed persons), EF52 (citizenship), EF521.1 and EF521.5 (one-person households or households of more than 4 persons)
- Correlation between EF259.4, Ef261.8 and EF261.9 (persons with Abitur and university graduates).

The ESS variables correlate (correlation coefficient greater 0.1) with a widely fluctuating number of regressors. Table B.9 for instance, shows for the NETUSE variable a correlative nexus of up to 22 regressors; on the other hand, the `membership` characteristic is only correlated with 3 regressors. Dummy variables-such as e.g. EF1.1 to EF1.10 (Federal state) or EF52 (citizenship)-have poor explanatory value for an ESS characteristic, and are entered into the Logit models as explanatory variables only in a few cases.

Result:

Generally speaking, the ESS variables could be synthesised very well, notwithstanding the small datasets (only 1,932 non-response-free lines) and the relatively high discrepancies noted in the frequencies of regressors. Original and generated ESS variables differ in their frequencies only marginally: The maximal absolute deviation is only 2.59% (in relative terms: -4.56%), as noted for the NETUSE variable. The maximal relative deviation is -5.73 percentage points (in absolute terms: -2.12 %), as found in the WKVLORG variable.

The generated correlation structure closely resembles that of the original dataset (as measured on the empirical correlation coefficient), in particular the correlation between ESS variables and regressors. Deviations occur in how the generated variables relate to each other. The correlation between the following variables could not be represented:

- WKVLORG und Membership
- WRINCO und STFSDLV
- FLCLPLA und Membership, SCLMEET.3 und WKVLORG.

Table B.7: Bivariate correlation coefficients of ESS and GMC data

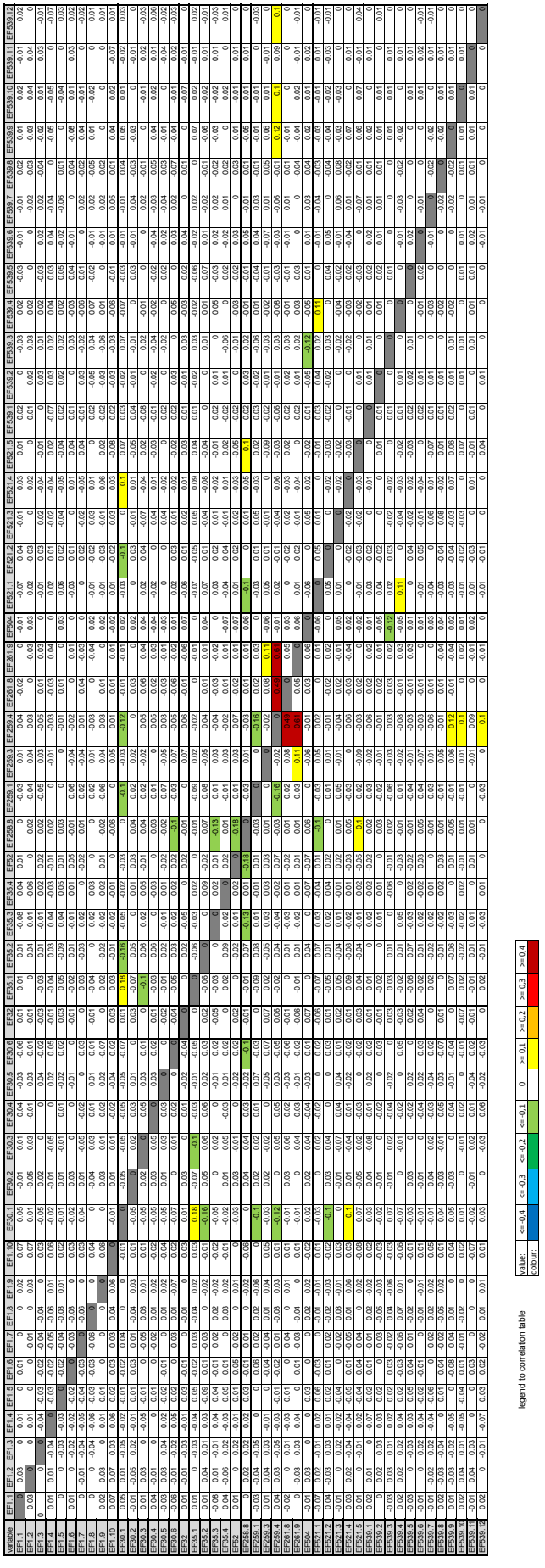

Table B.8: Differences of bivariate correlation coefficients ESS and GMC data

Table B.9: Bivariate correlation coefficients of ESS and GMC data

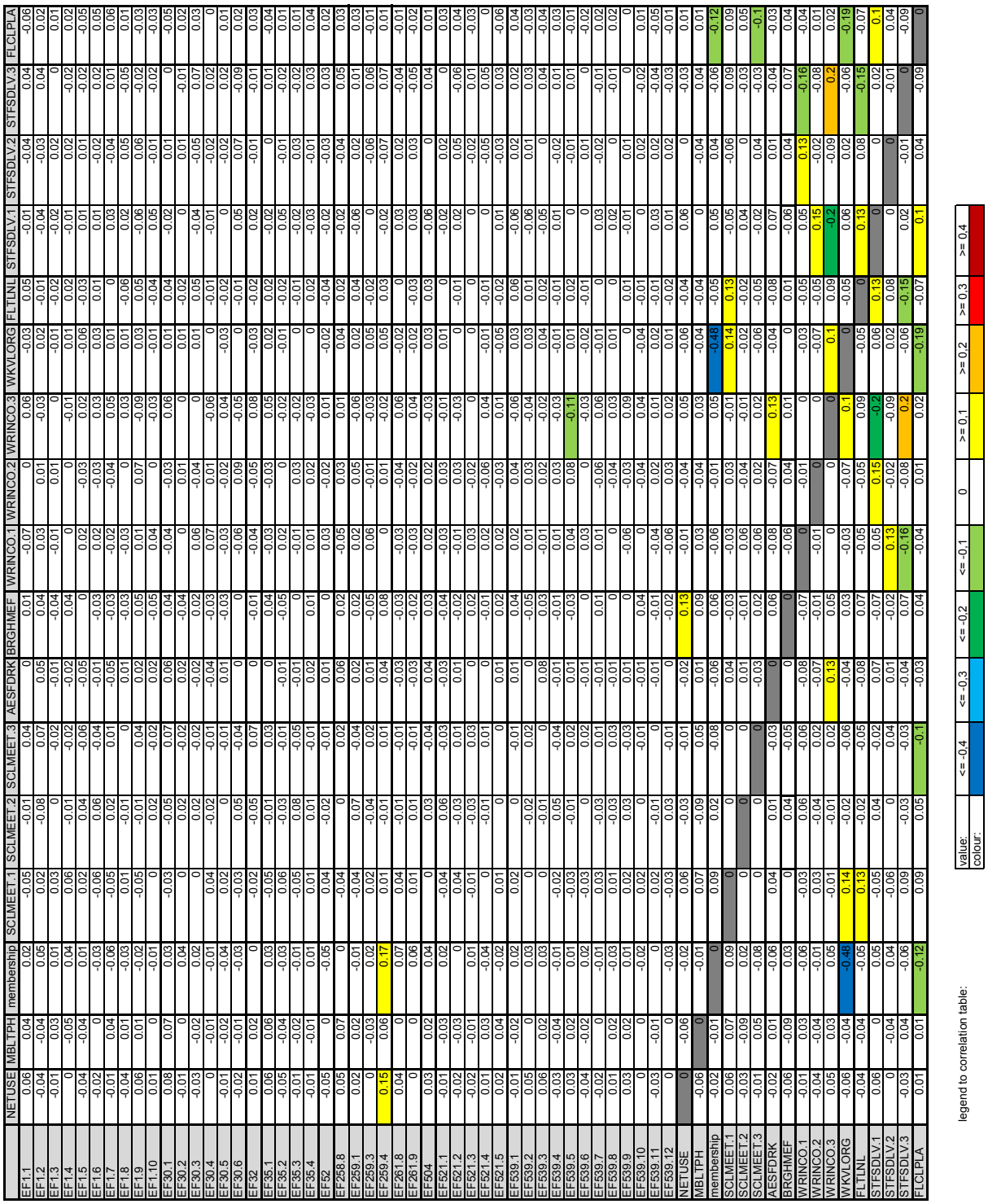

Table B.10: Bivariate correlation coefficients of ESS and GMC data

Table B.9 contrasts the correlations between the original and semi-synthetic datasets for each variable (original and generated values are compared column by column). Deviations in the correlation coefficients (see also table B.10) between original- and semi-synthetic datasets are largely confined to the bottom quarter of the table, where the relationships between the ESS characteristics are set out.

The colour pattern is largely identical for each variable of interest in respect of the regressors (listed line-by-line). Deviations are evident in the generated ESS variables (bottom quarter of table).

B.2 Correlation structures of single indicators

Table B.11 shows how the single indicators correlate both with each other and with the regressors available in GMC. Multiple regressors are correlated with both the single indicators and the composite indicator. The strength of the linear relationship is great. The (in part) high correlation coefficients result, inter al., from data aggregation, i.e. indicators are formed from several discrete variables.

The precision of regression-based estimation methods is highly dependent on the auxiliary information entered and how this correlates with the characteristic to be estimated. The individual auxiliary variable sets using various dummy variables (which also vary in number) generated different results in the simulation study. The variance of point estimates is heavily dependent on the auxiliary variable set used. For an overview of the composition of the auxiliary variable set, readers are referred to Tables A.5 bis A.8. The relationships between the variables are plausible (in terms of content) and permit e.g. the following conclusions:

The living standard (single indicator 2 and sub-indicators 2.1 to 2.5) strongly depends on income. The higher the household income, the more financial options are available to a household. Childless married couples have higher incomes than childless couples, single parents or divorced persons. Tertiary qualifications and a full-time job are important factors in a high household income. Households drawing state benefits (unemployment support, housing allowance, welfare aid) have a lower standard of living. Income increases with advancing age until retirement. Retired person households (persons over 65) and widowed persons are, on average, less well provided with goods and have a smaller financial cushion. Households whose chief income provider is aged between 25 and 55 are better furnished with goods, especially in the case of families. Persons in this age cohort are fearful, however, in comparison with today's retirees, that they will not have an adequate income in old age. Among married couples with children, fear of poverty in old age is smaller than in other household types.

The health of persons (single indicator 5) displays, contrary to all other indicators, hardly any significant linear relationships with other variables. Aggregation of the variables PH010, PH020 and PH030 results in altered correlation coefficients: The linear relationships with the characteristics of age, marital status, job status, household size and income that were used in generating these EU-SILC variables are no longer evident. Estimation of this indicator might - if the correlation coefficients are anything to go by - involve accepting less precision.

code	description	label	S1	S2	S21	S22	S23	S24	S25	S3	S4	S5	S6	S7	CI
szh	social affiliation	self-employed	0.15	0.1	0.06	0.14	0.07	0.06	-0.01	0.05	0.17	0.02	0.06	0.07	0.18
		civil servant, employee, and apprentice	0.23	0.18	0.08	0.47	0.06	0.11	-0.12	0.08	0.15	0.07	0.08	0.2	0.3
		pensioner	-0.11	-0.01	0.1	-0.54	0.06	0.01	0.24	-0.07	-0.19	-0.07	-0.05	0.21	-0.1
		other, or stay-at-home person	-0.32	-0.35	-0.31	-0.04	-0.23	-0.23	-0.17	-0.06	-0.08	-0.01	-0.1	-0.69	-0.45
hhtyp	type of household	single-person household	-0.14	-0.31	-0.1	-0.44	-0.08	-0.19	-0.06	-0.09	-0.12	-0.04	-0.62	-0.02	-0.3
		couple-household without children	0.18	0.21	0.15	0.07	0.15	0.09	0.13	0.09	0.06	-0.01	0.34	0.07	0.21
		single parent	-0.15	-0.15	-0.22	0.05	-0.15	-0.07	-0.07	-0.09	-0.03	0.02	0.03	-0.09	-0.12
		couple-household with children	-0.03	0.12	0.01	0.35	-0.03	0.11	-0.06	0.02	0.09	0.04	0.23	-0.01	0.11
		other households	0.08	0.11	0.05	0.12	0.05	0.07	0.03	0.02	0	0.02	0.16	0	0.09
inc	household income	< 900 Euro	-0.5	-0.46	-0.36	-0.32	-0.25	-0.26	-0.13	-0.13	-0.17	-0.02	-0.3	-0.24	-0.49
		900 Euro - 1300 Euro	-0.3	-0.22	-0.12	-0.26	-0.15	-0.13	0	-0.1	-0.16	-0.01	-0.2	-0.03	-0.27
		1300 Euro - 2600 Euro	0.01	0.1	0.12	0.06	0.03	0.05	0.05	0.02	-0.03	-0.01	0.09	0.08	0.06
		2600 Euro - 3600 Euro	0.34	0.25	0.16	0.27	0.15	0.16	0.01	0.1	0.13	0.03	0.19	0.09	0.3
		> 3600 Euro	0.52	0.33	0.2	0.28	0.24	0.19	0.06	0.12	0.28	0.02	0.22	0.09	0.44
EF1.1	Federal state	Schleswig-Holstein, Mecklenburg-Vorp.	-0.03	-0.01	-0.01	0	-0.01	-0.01	0	-0.01	0	0	0	-0.02	-0.02
EF1.2		Bremen, Hamburg, Berlin	-0.01	-0.05	-0.04	-0.04	-0.04	-0.02	-0.02	-0.03	0.03	0	-0.07	-0.04	-0.03
EF1.3		Niedersachsen	0	0.02	0.01	0.01	0.01	0.01	0.01	0.01	-0.02	0	0	0.02	0
EF1.4		Nordrhein-Westfalen	0.02	0.03	0.03	0	0.02	0.01	0.02	0	-0.04	0	-0.01	0.01	0
EF1.5		Hessen	0.04	0.03	0.01	0.03	0.01	0.02	0.01	0.02	0.02	0	0.03	0.02	0.05
EF1.6		Rheinland-Pfalz, Saarland	0.01	0.02	0	0.01	0.01	0.01	0.02	0.01	-0.02	0	0	0.02	0.01
EF1.7		Baden-Württemberg	0.06	0.04	0.02	0.04	0.01	0.02	0.01	0.02	0.01	0	0.01	0.04	0.05
EF1.8		Bayern	0.05	0.05	0.03	0.03	0.03	0.02	0.01	0.09	-0.01	0.01	0.01	0.04	0.06
EF1.9		Brandenburg	-0.05	-0.04	-0.03	-0.01	-0.02	-0.05	0	-0.12	0.02	0	0	-0.05	-0.06
EF1.10		Sachsen, Sachsen-Anhalt, Thüringen	-0.13	-0.1	-0.06	-0.07	-0.04	-0.05	-0.06	-0.06	0.04	0	0.02	-0.08	-0.08
EF30.1	Age	aged < 25 years	-0.15	-0.13	-0.12	0.05	-0.05	-0.09	-0.13	-0.05	-0.05	0	-0.17	-0.04	-0.14
EF30.2		aged 25-35 years	-0.03	-0.06	-0.07	0.19	-0.06	-0.03	-0.14	-0.02	0.05	0.01	-0.11	-0.06	-0.03
EF30.3		aged 35-45 years	0.03	0	-0.07	0.26	-0.05	0.02	-0.11	0.02	0.09	0.03	0.07	-0.06	0.06
EF30.4		aged 45-55 years	0.09	0.06	0.01	0.19	-0.02	0.03	-0.02	0.06	0.07	0.03	0.1	-0.08	0.09
EF30.5		aged 55-65 years	0.08	0.04	0.04	0.01	0.05	0.02	0	0.08	0.03	-0.01	0.09	-0.02	0.06
EF30.6		aged > 65 years	-0.09	0.01	0.13	-0.56	0.09	0.01	0.27	-0.09	-0.17	-0.06	-0.06	0.19	-0.09
EF32	Gender	male	0.16	0.24	0.17	0.33	0.1	0.11	0.01	0.19	0.16	0.02	0.33	-0.01	0.26
EF35.1	Marital status	single	-0.05	-0.2	-0.12	0.04	-0.07	-0.15	-0.2	-0.02	0.05	0	-0.31	-0.09	-0.11
EF35.2		married	0.14	0.34	0.2	0.33	0.13	0.21	0.1	0.13	0.12	0.03	0.55	0.06	0.31
EF35.3		widowed	-0.07	-0.04	0.04	-0.47	0.03	0.01	0.18	-0.1	-0.22	-0.04	-0.29	0.1	-0.18
EF35.4		divorced or separated	-0.08	-0.23	-0.21	-0.05	-0.16	-0.15	-0.09	-0.06	-0.03	0	-0.13	-0.08	-0.15
EF52	Citizenship	German	0.12	0.08	0.08	-0.03	0.11	0.03	0.04	0	0.08	-0.01	-0.01	0.07	0.1
EF110.1		Employed	0.31	0.24	0.11	0.54	0.09	0.14	-0.12	0.11	0.23	0.08	0.11	0.23	0.39
EF130.1		Employed in public sector	0.18	0.12	0.06	0.19	0.06	0.07	-0.02	0.04	0.19	0.02	0.04	0.09	0.2
EF138.1		Working full-time	0.35	0.26	0.15	0.52	0.12	0.15	-0.1	0.12	0.22	0.07	0.12	0.25	0.41
EF138.8		Working part time	-0.09	-0.05	-0.08	0.05	-0.06	-0.02	-0.04	-0.04	0.03	0.02	-0.02	-0.02	-0.04
EF215.1/2		Collecting unemployment benefits	-0.24	-0.28	-0.22	-0.03	-0.17	-0.22	-0.14	-0.04	-0.07	-0.01	-0.07	-0.88	-0.45
EF259.1	Highest education of level	CSE	-0.21	-0.11	0	-0.32	-0.04	-0.09	0.1	-0.07	-0.44	-0.02	-0.03	0.03	-0.28
EF259.2		secondary scholl (GDR)	-0.09	-0.09	-0.07	0.04	-0.06	-0.06	-0.07	-0.06	-0.02	0.01	0.02	-0.14	-0.11
EF259.3		secondary modern school certificate	0.07	0.02	0	0.11	0.01	0.03	-0.09	0.03	-0.1	0.01	-0.05	0.01	0
EF259.4		higher education entrance qualification	0.12	0.08	0.04	0.13	0.04	0.05	-0.01	0.05	0.28	0.01	0.05	0.02	0.19
EF259.5		general qualification for university entrance	0.25	0.17	0.06	0.25	0.09	0.11	0	0.09	0.56	0.01	0.05	0.05	0.38
EF261.4	Highest level of training qualification	vocational school	0.02	0	0	0.02	0	0	-0.01	0	-0.04	0	-0.02	0	-0.01
EF261.5		foreman/ university of cooperative education	0.1	0.06	0.05	0.07	0.04	0.04	-0.01	0.03	0.45	0.01	0.06	0.03	0.23
EF261.6		professional school (GDR)	0.01	-0.01	0	-0.02	0	0	0	-0.02	0.19	0	0.01	0	0.06
EF261.7		advanced technical collage	0.09	0.05	0.03	0.05	0.04	0.03	0.01	0.03	0.14	0	0.02	0.02	0.11
EF261.8/9/10		university degree, PhD	0.33	0.24	0.11	0.25	0.13	0.15	0.05	0.12	0.62	0.01	0.15	0.05	0.46
EF338>5		on welfare, care insurance, student loan	-0.19	-0.21	-0.24	-0.05	-0.15	-0.1	-0.08	-0.06	-0.06	0	-0.06	-0.11	-0.18
EF358.1		Receipt of housing allowances	-0.26	-0.31	-0.37	-0.1	-0.24	-0.14	-0.09	-0.08	-0.11	0	-0.08	-0.29	-0.31
EF359.1		Receipt of benefit payments	-0.2	-0.25	-0.29	-0.1	-0.18	-0.11	-0.07	-0.07	-0.1	0	-0.06	-0.17	-0.23
EF521.1	Number of people in private household	1 person	-0.14	-0.31	-0.1	-0.44	-0.08	-0.19	-0.06	-0.09	-0.12	-0.04	-0.62	-0.02	-0.3
EF521.2		2 persons	0.14	0.16	0.08	0.07	0.11	0.07	0.12	0.06	0.04	-0.01	0.34	0.04	0.17
EF521.3		3 persons	0.04	0.08	0.01	0.22	-0.01	0.07	-0.03	-0.01	0.03	0.02	0.17	-0.02	0.08
EF521.4		4 persons	0	0.11	0.03	0.25	0	0.1	-0.03	0.02	0.07	0.04	0.19	0.01	0.11
EF539.1	Household monthly net income	0-500 Euro	-0.24	-0.23	-0.15	-0.08	-0.09	-0.17	-0.12	-0.03	-0.05	-0.01	-0.15	-0.14	-0.23
EF539.2		500-900 Euro	-0.42	-0.38	-0.31	-0.31	-0.22	-0.2	-0.08	-0.13	-0.16	-0.02	-0.25	-0.19	-0.42
EF539.3		900-1300 Euro	-0.3	-0.22	-0.12	-0.26	-0.15	-0.13	0	-0.1	-0.16	-0.01	-0.2	-0.03	-0.27
EF539.4		1300-2000 Euro	-0.1	0.01	0.06	-0.04	-0.04	0.01	0.04	-0.02	-0.07	-0.01	0	0.03	-0.06
EF539.5		2000-2600 Euro	0.14	0.13	0.08	0.14	0.08	0.06	0.01	0.05	0.04	0.01	0.12	0.07	0.15
EF539.6		2600-3600 Euro	0.34	0.25	0.16	0.27	0.15	0.16	0.01	0.1	0.13	0.03	0.19	0.09	0.3
EF539.7		3600-5000 Euro	0.38	0.23	0.14	0.21	0.17	0.15	0.01	0.09	0.19	0.02	0.17	0.07	0.32
EF539.8		> 5000 Euro	0.34	0.23	0.13	0.17	0.17	0.1	0.08	0.07	0.2	0.01	0.12	0.05	0.29
EF30 & EF32	Age cross sex (male)	aged 20-25 years	-0.09	-0.07	-0.06	0.05	-0.03	-0.06	-0.08	-0.01	-0.03	0	-0.1	-0.04	-0.09
		aged 25-30 years	-0.03	-0.04	-0.03	0.09	-0.03	-0.03	-0.08	0.01	0	0	-0.06	-0.04	-0.03
		aged 30-35 years	0.02	0.02	0	0.15	0	0.01	-0.07	0.01	0.04	0.01	-0.02	-0.03	0.03
		aged 35-40 years	0.04	0.03	0	0.16	-0.01	0.02	-0.05	0.04	0.06	0.02	0.05	-0.02	0.06
		aged 40-45 years	0.04	0.04	0	0.17	-0.01	0.03	-0.05	0.03	0.05	0.02	0.07	-0.02	0.07
		aged 45-50 years	0.06	0.07	0.03	0.14	0	0.04	-0.01	0.06	0.05	0.02	0.08	-0.04	0.08
		aged 50-55 years	0.09	0.07	0.04	0.14	0.01	0.04	0	0.06	0.05	0.02	0.09	-0.04	0.09
		aged 55-60 years	0.09	0.07	0.06	0.06	0.06	0.03	0	0.06	0.04	0.01	0.1	-0.05	0.08
		aged 60-65 years	0.05	0.06	0.04	0.04	0.04	0.03	0.01	0.07	0.04	-0.01	0.11	0.02	0.08
		aged 65-70 years	0	0.06	0.08	-0.13	0.06	0.03	0.11	0.01	0.01	-0.02	0.11	0.08	0.05
		aged 70-75 years	-0.01	0.05	0.07	-0.11	0.04	0.02	0.1	0.01	-0.01	-0.01	0.08	0.06	0.03
		aged > 75	0	0.06	0.08	-0.14	0.06	0.03	0.11	0.02	-0.01	-0.02	0.06	0.07	0.03
	Age cross sex (female)	aged 20-25 years	-0.1	-0.09	-0.09	0.03	-0.04	-0.05	-0.09	-0.06	-0.02	0	-0.11	-0.01	-0.09
		aged 25-30 years	-0.04	-0.07	-0.07	0.04	-0.05	-0.04	-0.06	-0.04	0.02	0	-0.08	-0.02	-0.04
		aged 30-35 years	-0.01	-0.05	-0.06	0.05	-0.04	-0.02	-0.05	-0.04	0.03	0.01	-0.06	-0.03	-0.02
		aged 35-40 years	-0.02	-0.06	-0.09	0.05	-0.05	-0.03	-0.05	-0.03	0.02	0.01	-0.01	-0.04	-0.03
		aged 40-45 years	-0.03	-0.06	-0.08	0.05	-0.05	-0.03	-0.05	-0.03	0.02	0.01	-0.01	-0.04	-0.03
		aged 45-50 years	-0.01	-0.04	-0.05	0.02	-0.05	-0.02	-0.02	-0.02	0.01	0.01	-0.02	-0.04	-0.02
		aged 50-55 years	0	-0.04	-0.05	0	-0.04	-0.02	-0.02	-0.02	0	0.01	-0.04	-0.03	-0.03
		aged 55-60 years	-0.01	-0.05	-0.02	-0.06	-0.01	-0.03	-0.02	-0.01	-0.02	0	-0.05	-0.03	-0.04
		aged 60-65 years	-0.03	-0.06	-0.04	-0.09	-0.02	-0.02	-0.01	-0.01	-0.04	-0.01	-0.08	0.02	-0.05
		aged 65-70 years	-0.04	-0.05	0	-0.21	0	-0.02	0.06	-0.07	-0.08	-0.02	-0.1	0.05	-0.08
		aged 70-75 years	-0.05	-0.04	0.01	-0.22	0	-0.02	0.06	-0.06	-0.1	-0.02	-0.11	0.05	-0.09
		aged > 75	-0.09	-0.07	0.02	-0.39	0.01	-0.02	0.11	-0.11	-0.18	-0.03	-0.21	0.08	-0.16
Sub-Indicator 1		Income	1	0.55	0.41	0.35	0.42	0.31	0.12	0.21	0.37	0.02	0.18	0.27	0.75
Sub-Indicator 2		Standard of Living	0.55	1	0.52	0.41	0.56	0.65	0.54	0.16	0.26	0.02	0.27	0.32	0.65
Sub-Indicator 21		Affordarbility of goods	0.41	0.52	1	0.11	0.27	0.2	0.13	0.12	0.14	0	0.13	0.25	0.41
Sub-Indicator 22		Possession of Durables	0.35	0.41	0.11	1	0.11	0.16	-0.1	0.15	0.31	0.06	0.29	0.03	0.43
Sub-Indicator 23		Ability to make ends meet	0.42	0.56	0.27	0.11	1	0.17	0.1	0.1	0.16	0	0.11	0.2	0.41
Sub-Indicator 24		Satisfied with standard of living	0.31	0.65	0.2	0.16	0.17	1	0.09	0.09	0.15	0.01	0.17	0.23	0.4
Sub-Indicator 25		Worried that income in old age not be adequate to cover last years	0.12	0.54	0.13	-0.1	0.1	0.09	1	0.02	0.01	-0.01	0.08	0.15	0.2
Sub-Indicator 3		Housing	0.21	0.16	0.12	0.15	0.1	0.09	0.02	1	0.13	0.01	0.09	0.05	0.42
Sub-Indicator 4		Education	0.37	0.26	0.14	0.31	0.16	0.15	0.01	0.13	1	0.02	0.17	0.08	0.62
Sub-Indicator 5		Health	0.02	0.02	0	0.06	0	0.01	-0.01	0.01	0.02	1	0.03	0	0.23
Sub-Indicator 6		Social Relations	0.18	0.27	0.13	0.29	0.11	0.17	0.08	0.09	0.17	0.03	1	0.08	0.43
Sub-Indicator 7		Work	0.27	0.32	0.25	0.03	0.2	0.23	0.15	0.05	0.08	0	0.08	1	0.51
Composite Index		Composite Indicator	0.75	0.65	0.41	0.43	0.41	0.4	0.2	0.42	0.62	0.23	0.43	0.51	1

legend to correlation table:

value:	<= -0,4	<= -0,3	<= -0,2	<= -0,1	0	>= 0,1	>= 0,2	>= 0,3	>= 0,4
colour:									

Table B.11: Bivariate correlation coefficients of single indicators and CIoILC

The variable EF1 (Federal state) correlates with the indicators only to a (very) limited degree. The dummy variables that result from cross-combining the characteristics of age (in 5-year stages) and gender correlate only in individual cases with the indicators. As for the cross-combination of age and gender, the suspicion is that the class width chosen for the age variables has been too narrow. The extent to which these concerns are justified is analysed within the simulation study.

Single indicators 1 (`income`), 2 (`living standard`) and 4 (`education`) are highly correlated with each other. The correlative relationships of the sub-indicators entered into single indicator 2 are less pronounced. This is encouraging from the perspective of dataset generation, as from both substantive and statistical properties different sub-indicators are to be generated. Single indicators 3 (`housing`) and 6 (`social relations`) correlate with income and living standard, albeit far less strongly. Especially the social relations indicator is dependent on other characteristics like marital status or household type. The health indicator is not correlated to any meaningful extent with any other single indicator.

Furthermore, the plausibility of the generated variables and indicators is assessed using conditional distributions. In the following, this assessment is illustrated for sub-indicator 2.1 (`affordability of goods`) and single indicator 5 (`health`).

Figure B.1 shows the distribution of sub-indicator 2.1 conditioned on the variables of `household net income` (hne), `age` (EF30), `marital status` (EF35), `social affiliation` (szh) and `type of household` (hhtyp). The figures confirms previous results. The greater a household's income, the more financial options it has at its disposal and the larger the point score for the sub-indicator. The number of households whose household-adjusted incomes lie below Euro 900 decreases steadily as the score increases. Households scoring between 1 and 5 points stem, in 75% of cases, from income class 1. Only 5% of households scoring the highest point of 5 come from income class 1. Those households that are less financially well off (1 point) can be further characterised. The number of households whose primary carer is under the age of 25 (EF30.1) is the largest posted (33%). Earning capacity grows with increasing age, giving households more financial options. Widowed persons are particularly badly off (EF35.4). Households scoring a single (1) point make up 68% of the total. Childless married couples (EF35.2 and hhtyp.2) are better able to shoulder financial burdens. Some 32% of households scoring 5 points are married couples (EF35.2), mostly without children (cf. hhtyp.2 and hhtyp.4). The number of single households (EF35.1) remains, across the point distribution spread, relatively constant on 20%. Single-person households are represented in all income classes. The number of single parents (hhtyp.3) is relatively high (83%) among households scoring low points. Furthermore, the number of other households where unemployment is high is especially large (szh.4). The point scale (1 to 5) is broad enough to include all social classes, with the number of self-employed persons lowest among the lowest-scoring households.

A person's subjective estimation of her/his state of health (single indicator 5) is, as figure B.2 shows, only weakly dependent on income. The numbers of income classes vary only slightly for alternative point scores with this indicator, with 1 point reflecting a very poor health status and 5 a very good one. Persons over 55, unlike younger persons, tend to suffer from health ailments or disabilities. Among those in poor health there are disproportionally many pensioners and widowed persons. Compared with sub-indicator 2.1, univariate frequency distributions show no distinctive structures for individual socio-

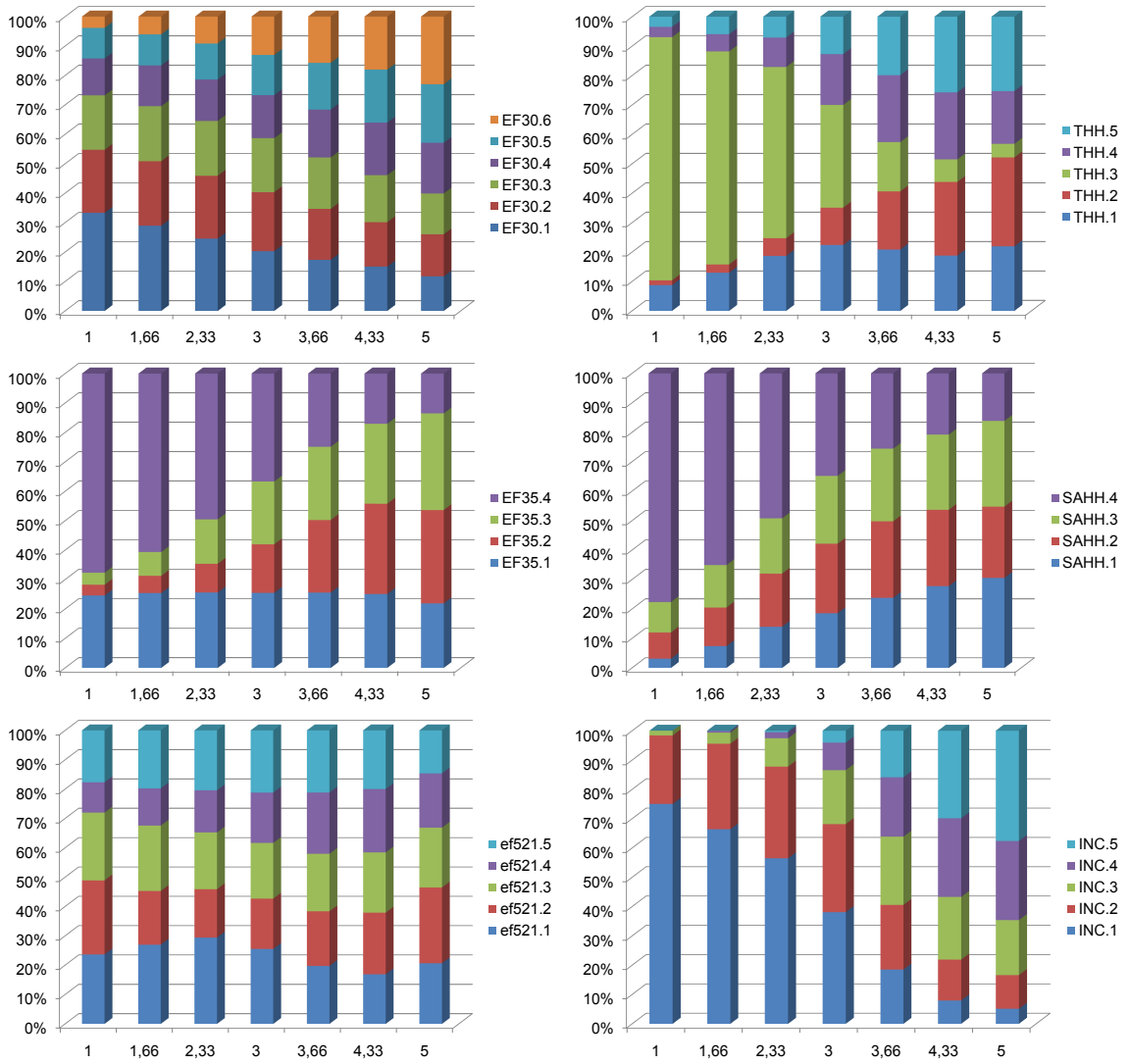

Figure B.1: Sub-indicator 2.1 (`affordability of goods`): conditioned frequency distributions

demographic groups. A person's health status is, on the whole, a very individual thing and cannot be specified by one variable alone.

The figures lend themselves to further analysis, especially with regard to the socio-demographic groups not comprehensively mentioned. For now, we will leave this task to our readers. Our interpretations are chiefly designed to help in interpreting the figures.

Based on the comparison of frequency distributions and correlation structures existing between the original and generated variables, as well as the conditional distributions evaluated for the individual variables and indicators, the dataset can, in the main, be rated as reliable, especially bearing in mind the EU-SILC and ESS database.

With an eye to the simulation study and its results, it is worth pointing out that the latter cannot be more than approximations of reality. The semi-synthetic population can, however, shed light on the risks and benefits of calculating composite indicators. By modifying the composite indicator of individual living conditions, we have addressed a thematic field of considerable relevance, where reliable advice is much needed by politicians.

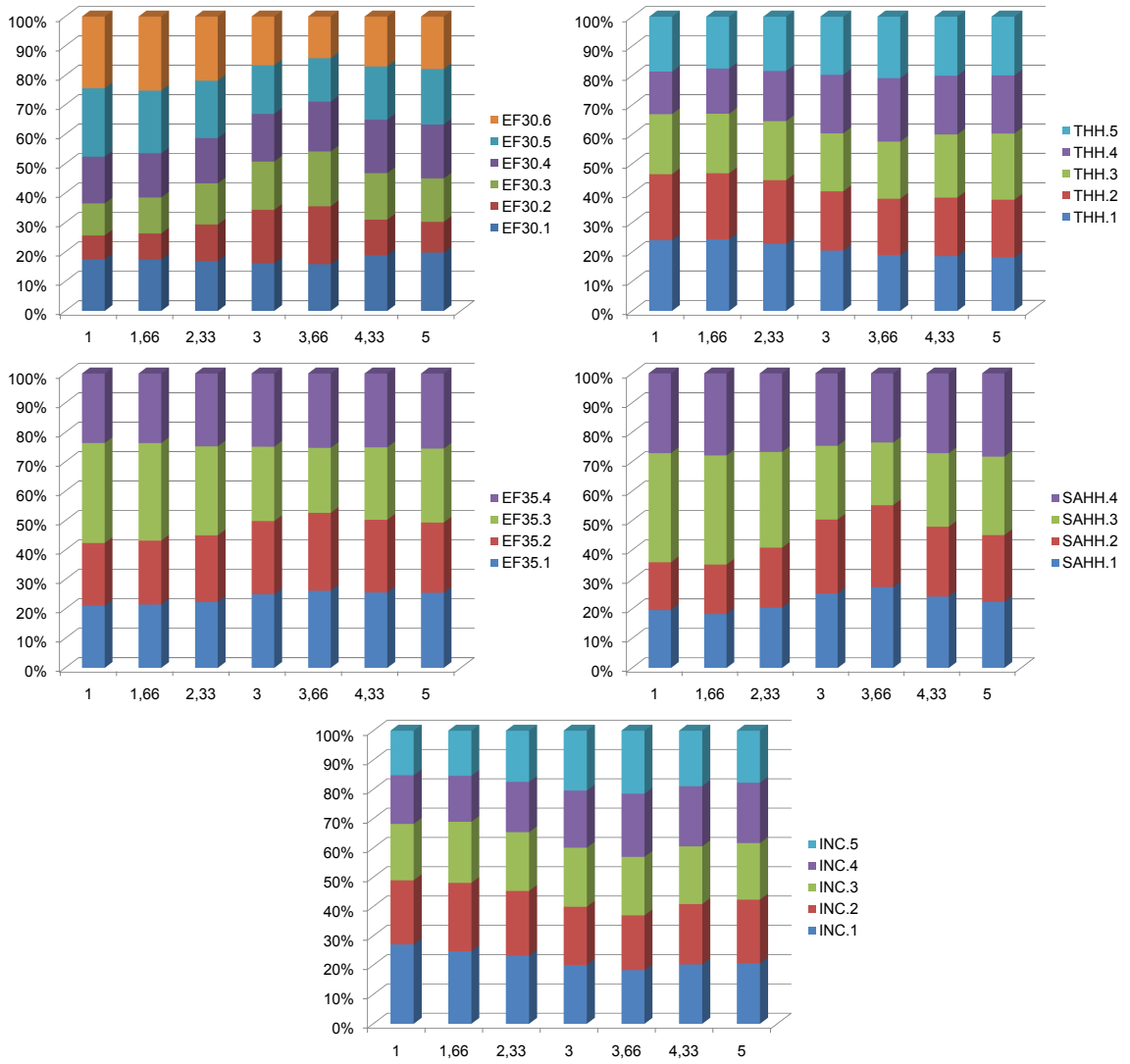

Figure B.2: Single indicator 5 (`health`): conditioned frequency distributions

Appendix C

Stratification plan

According to Destatis (2004), EU-SILC stratification variables are as follows:

1. Federal state, 10 classes
2. type of household (hhtyp), 5 classes
3. social affiliation of head of household (szh), 4 classes
4. household net income (hne), 5 classes.

The original stratification proposed by DESATATIS has been modified in the simulation study. A few cells (stratification classes) were collapsed in order to guarantee, that every cell covers at least two households.

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 5	172	172	3	2	8	4
2	1 2 1	341	341	6	4	15	9
3	1 2 2	492	492	8	6	22	13
4	1 2 3 -1 2 5	665	665	10	8	30	17
5	1 3 1	692	692	25	8	31	18
6	1 3 2	758	758	23	9	34	20
7	1 3 3 -1 3 5	502	502	12	6	23	13
8	1 4 1 -1 4 5	662	662	8	8	30	17
9	2 1 1 -2 1 5	214	428	3	2	10	6
10	2 2 1 -2 2 3	755	1507	12	9	34	20
11	2 2 4	388	776	4	4	17	10
12	2 2 5	223	445	3	3	10	6
13	2 3 1	51	102	3	2	2	2
14	2 3 2	270	540	10	3	12	7
15	2 3 3	1334	2668	40	15	60	35
16	2 3 4 -2 3 5	317	633	7	4	14	8
17	2 4 1 -2 4 5	365	730	4	4	16	9
18	3 1 1 -3 1 5	19	45	3	2	2	2
19	3 2 1	46	103	5	2	2	2
20	3 2 2	101	226	9	2	5	3
21	3 2 3 -3 2 5	184	482	12	2	8	5
22	3 4 1*	103	247	7	2	5	3
23	3 4 2*	75	197	4	2	3	2
24	3 4 3 -3 4 5*	41	124	2	2	2	2
25	4 1 1 -4 1 3	80	295	3	2	4	2
26	4 1 4	52	208	3	2	2	2
27	4 1 5	78	308	3	2	4	2
28	4 2 1 -4 2 2	41	152	4	2	2	2
29	4 2 3	801	2968	42	9	36	21
30	4 2 4	471	1796	19	5	21	12
31	4 2 5	275	1080	11	3	12	7
32	4 4 1 -4 4 2*	123	473	4	2	6	3
33	4 4 3 -4 4 5*	273	1062	4	3	12	7
34	5 1 1 -5 1 4	74	221	3	2	3	2
35	5 1 5	44	149	3	2	2	2
36	5 2 1 -5 2 2	46	110	3	2	2	2
37	5 2 3	286	790	18	3	13	7
38	5 2 4	202	636	12	2	9	5
39	5 2 5	174	599	10	2	8	5
40	5 4 1*	25	57	2	2	2	2
41	5 4 2*	62	152	3	2	3	2
42	5 4 3*	263	739	8	3	12	7
43	5 4 4 -5 4 5*	147	461	4	2	7	4

Table C.1: Stratification plan implemented for Federal states SCH and MVP (EF1.1)

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 5	595	595	3	7	27	15
2	1 2 1	756	756	8	9	34	20
3	1 2 2	1013	1013	12	12	46	26
4	1 2 3 -1 2 5	1835	1835	16	21	83	47
5	1 3 1	787	787	35	9	35	20
6	1 3 2	1127	1127	31	13	51	29
7	1 3 3 -1 3 5	1066	1066	21	12	48	28
8	1 4 1 -1 4 5	1707	1707	11	20	77	44
9	2 1 1 -2 1 5	345	689	6	4	16	9
10	2 2 1 -2 2 3	842	1678	17	10	38	22
11	2 2 4	443	885	8	5	20	11
12	2 2 5	365	727	6	4	16	9
13	2 3 1	32	64	4	2	2	2
14	2 3 2	156	312	14	2	7	4
15	2 3 3	1391	2782	54	16	63	36
16	2 3 4 -2 3 5	475	949	12	5	21	12
17	2 4 1 -2 4 5	513	1025	5	6	23	13
18	3 1 1 -3 1 5	62	144	3	2	3	2
19	3 2 1	54	118	7	2	2	2
20	3 2 2	149	331	12	2	7	4
21	3 2 3 -3 2 5	350	854	21	4	16	9
22	3 4 1*	129	293	9	2	6	3
23	3 4 2*	168	432	6	2	8	4
24	3 4 3 -3 4 5*	64	219	3	2	3	2
25	4 1 1 -4 1 3	124	474	6	2	6	3
26	4 1 4	55	218	4	2	2	2
27	4 1 5	122	456	6	2	5	3
28	4 2 1 -4 2 2	73	249	6	2	3	2
29	4 2 3	650	2379	61	8	29	17
30	4 2 4	423	1588	29	5	19	11
31	4 2 5	298	1150	17	3	13	8
32	4 4 1 -4 4 2*	160	569	6	2	7	4
33	4 4 3 -4 4 5*	326	1354	6	4	15	8
34	5 1 1 -5 1 5	129	368	9	2	6	3
35	5 2 1 -5 2 3	340	881	27	4	15	9
36	5 2 4	196	597	19	2	9	5
37	5 2 5	200	661	17	2	9	5
38	5 4 1*	128	311	7	2	6	3
39	5 4 3*	308	870	11	4	14	8
40	5 4 4 -5 4 5*	157	497	6	2	7	4

Table C.2: Stratification plan implemented for Federal states HAM, BRE, BER (EF1.2)

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 2	118	118	2	2	5	3
2	1 1 3	120	120	2	2	5	3
3	1 1 4 -1 1 5	84	84	2	2	4	2
4	1 2 1	442	442	8	5	20	11
5	1 2 2	868	868	13	10	39	22
6	1 2 3	1346	1346	20	16	61	35
7	1 2 4 -1 2 5	150	150	3	2	7	4
8	1 3 1	1192	1192	37	14	54	31
9	1 3 2	1307	1307	35	15	59	34
10	1 3 3	938	938	26	11	42	24
11	1 3 4 -1 3 5	90	90	3	2	4	2
12	1 4 1	938	938	11	11	42	24
13	1 4 2	89	89	2	2	4	2
14	1 4 3 -1 4 5	46	46	2	2	2	2
15	2 1 1 -2 1 3	150	299	2	2	7	4
16	2 1 4	83	166	2	2	4	2
17	2 1 5	147	294	3	2	7	4
18	2 2 1 -2 2 2	98	196	3	2	4	3
19	2 2 3	991	1977	19	11	45	26
20	2 2 4	676	1349	10	8	30	17
21	2 2 5	478	953	7	6	22	12
22	2 3 1	132	264	6	2	6	3
23	2 3 2	560	1120	17	6	25	14
24	2 3 3	2092	4184	60	24	94	54
25	2 3 4	456	910	11	5	21	12
26	2 3 5	211	421	6	2	10	5
27	2 4 1	75	150	2	2	3	2
28	2 4 2	125	250	2	2	6	3
29	2 4 3 -2 4 5	219	436	2	3	10	6
30	3 1 1 -3 1 2	13	32	2	2	2	2
31	3 1 3	15	42	2	2	2	2
32	3 1 4 -3 1 5	12	31	2	2	2	2
33	3 2 1	61	139	7	2	3	2
34	3 2 2	174	401	13	2	8	5
35	3 2 3	228	590	23	3	10	6
36	3 2 4 -3 2 5	37	102	3	2	2	2
37	3 4 1*	113	277	10	2	5	3
38	3 4 2*	107	295	8	2	5	3
39	3 4 3 -3 4 5*	48	161	3	2	2	2
40	4 1 1 -4 1 2	27	98	2	2	2	2

Table C.3: Stratification plan implemented for Federal state NIE (EF1.3), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	4 1 3	122	465	6	2	5	3
42	4 1 4	108	436	5	2	5	3
43	4 1 5	167	682	8	2	8	4
44	4 2 1	9	32	2	2	2	2
45	4 2 2	78	274	4	2	4	2
46	4 2 3	1526	5745	71	18	69	40
47	4 2 4	933	3671	37	11	42	24
48	4 2 5	544	2193	23	6	25	14
49	4 4 1*	34	128	2	2	2	2
50	4 4 2*	99	368	3	2	4	3
51	4 4 3*	244	1022	6	3	11	6
52	4 4 4 -4 4 5*	64	288	2	2	3	2
53	5 1 1 -5 1 3	41	134	2	2	2	2
54	5 1 4	42	144	2	2	2	2
55	5 1 5	83	300	6	2	4	2
56	5 2 1	10	23	2	2	2	2
57	5 2 2	39	89	3	2	2	2
58	5 2 3	332	914	28	4	15	9
59	5 2 4	372	1207	22	4	17	10
60	5 2 5	386	1344	22	4	17	10
61	5 4 1*	37	84	3	2	2	2
62	5 4 2*	83	210	4	2	4	2
63	5 4 3*	404	1083	12	5	18	10
64	5 4 4*	210	683	5	2	9	5
65	5 4 5*	150	489	3	2	7	4

Table C.4: Stratification plan implemented for Federal state NIE (EF1.3), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1	97	97	3	2	4	3
2	1 1 2	85	85	2	2	4	2
3	1 1 3	220	220	5	3	10	6
4	1 1 4 -1 1 5	129	129	2	2	6	3
5	1 2 1	890	890	19	10	40	23
6	1 2 2	1633	1633	31	19	74	42
7	1 2 3	3003	3003	46	35	135	78
8	1 2 4	298	298	4	3	13	8
9	1 2 5	111	111	2	2	5	3
10	1 3 1	2099	2099	85	24	95	54
11	1 3 2	2759	2759	78	32	124	71
12	1 3 3	2046	2046	58	24	92	53
13	1 3 4	155	155	4	2	7	4
14	1 3 5	45	45	2	2	2	2
15	1 4 1	1700	1700	25	20	77	44
16	1 4 2	256	256	3	3	12	7
17	1 4 3 -1 4 5	120	120	2	2	5	3
18	2 1 1 -2 1 2	44	88	2	2	2	2
19	2 1 3	194	386	4	2	9	5
20	2 1 4	159	318	3	2	7	4
21	2 1 5	313	626	6	4	14	8
22	2 2 1	45	89	2	2	2	2
23	2 2 2	170	339	4	2	8	4
24	2 2 3	2071	4134	42	24	93	54
25	2 2 4	1441	2876	24	17	65	37
26	2 2 5	980	1959	16	11	44	25
27	2 3 1	184	368	12	2	8	5
28	2 3 2	955	1910	38	11	43	25
29	2 3 3	4270	8537	134	49	192	111
30	2 3 4	837	1674	26	10	38	22
31	2 3 5	426	851	12	5	19	11
32	2 4 1	188	374	3	2	8	5
33	2 4 2	249	497	3	3	11	6
34	2 4 3	420	836	3	5	19	11
35	2 4 4 -2 4 5	127	254	2	2	6	3
36	3 1 1 -3 4 5	1560	3990	162	18	70	40
37	4 1 1	5	21	2	2	2	2
38	4 1 2	16	61	2	2	2	2
39	4 1 3	219	847	14	3	10	6
40	4 1 4	187	724	10	2	8	5

Table C.5: Stratification plan implemented for Federal state NRW (EF1.4), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	4 1 5	325	1286	17	4	15	8
42	4 2 1	27	94	4	2	2	2
43	4 2 2	157	553	9	2	7	4
44	4 2 3	3129	11901	161	36	141	81
45	4 2 4	1814	7039	83	21	82	47
46	4 2 5	1168	4713	51	14	53	30
47	4 4 1*	93	334	5	2	4	2
48	4 4 2*	255	965	8	3	11	7
49	4 4 3*	589	2435	13	7	27	15
50	4 4 4*	102	428	2	2	5	3
51	4 4 5*	60	249	2	2	3	2
52	5 1 1 -5 1 2	7	19	2	2	2	2
53	5 1 3	56	152	5	2	3	2
54	5 1 4	73	225	5	2	3	2
55	5 1 5	148	506	12	2	7	4
56	5 2 1	20	45	3	2	2	2
57	5 2 2	52	121	6	2	2	2
58	5 2 3	753	2074	62	9	34	19
59	5 2 4	633	2044	50	7	29	16
60	5 2 5	735	2607	50	9	33	19
61	5 4 1*	96	221	6	2	4	2
62	5 4 2*	184	464	10	2	8	5
63	5 4 3*	804	2220	28	9	36	21
64	5 4 4*	373	1126	10	4	17	10
65	5 4 5*	255	849	7	3	11	7

Table C.6: Stratification plan implemented for Federal state NRW (EF1.4), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1	61	61	2	2	3	2
2	1 1 2	57	57	2	2	3	2
3	1 1 3	114	114	2	2	5	3
4	1 1 4 -1 1 5	62	62	2	2	3	2
5	1 2 1	265	265	6	3	12	7
6	1 2 2	570	570	10	7	26	15
7	1 2 3	1128	1128	15	13	51	29
8	1 2 4	138	138	2	2	6	4
9	1 2 5	71	71	2	2	3	2
10	1 3 1	722	722	28	8	33	19
11	1 3 2	856	856	26	10	39	22
12	1 3 3 -1 3 5	819	819	21	9	37	21
13	1 4 1	494	494	8	6	22	13
14	1 4 2 -1 4 5	112	112	2	2	5	3
15	2 1 1 -2 1 3	117	232	2	2	5	3
16	2 1 4	61	122	2	2	3	2
17	2 1 5	144	287	2	2	6	4
18	2 2 1	11	22	2	2	2	2
19	2 2 2	64	128	2	2	3	2
20	2 2 3	734	1462	15	8	33	19
21	2 2 4	529	1055	8	6	24	14
22	2 2 5	431	854	6	5	19	11
23	2 3 1	78	156	4	2	4	2
24	2 3 2	341	682	12	4	15	9
25	2 3 3	1473	2946	45	17	66	38
26	2 3 4	372	743	8	4	17	10
27	2 3 5	175	350	4	2	8	5
28	2 4 1	54	106	2	2	2	2
29	2 4 2	66	131	2	2	3	2
30	2 4 3 -2 4 5	190	377	2	2	9	5
31	3 1 1 -3 1 5	28	68	3	2	2	2
32	3 2 1	43	97	6	2	2	2
33	3 2 2	98	232	10	2	4	3
34	3 2 3	190	498	17	2	9	5
35	3 2 4	24	61	2	2	2	2
36	3 2 5	7	16	2	2	2	2
37	3 4 1*	72	173	8	2	3	2
38	3 4 2*	50	136	6	2	2	2
39	3 4 3 -3 4 5*	37	122	3	2	2	2
40	4 1 1	5	22	2	2	2	2

Table C.7: Stratification plan implemented for Federal state HES (EF1.5), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	4 1 2	82	327	4	2	4	2
42	4 1 3	81	311	3	2	4	2
43	4 1 4	154	591	6	2	7	4
44	4 1 5	12	42	2	2	2	2
45	4 2 1	49	180	3	2	2	2
46	4 2 2	1135	4250	54	13	51	29
47	4 2 3	739	2844	28	9	33	19
48	4 2 4	521	2046	17	6	23	13
49	4 2 5	19	73	2	2	2	2
50	4 4 1*	72	273	3	2	3	2
51	4 4 2*	171	698	5	2	8	4
52	4 4 3*	58	218	2	2	3	2
53	4 4 4 -4 4 5*	44	117	2	2	2	2
54	5 1 1 -5 1 3	34	104	2	2	2	2
55	5 1 4	82	277	4	2	4	2
56	5 1 5	9	17	2	2	2	2
57	5 2 1	36	79	2	2	2	2
58	5 2 2	291	791	20	3	13	8
59	5 2 3	298	958	17	3	13	8
60	5 2 4	369	1310	17	4	17	10
61	5 2 5	35	78	2	2	2	2
62	5 4 1*	59	154	3	2	3	2
63	5 4 2*	319	888	10	4	14	8
64	5 4 3*	162	498	4	2	7	4
65	5 4 4*	125	411	2	2	6	3

Table C.8: Stratification plan implemented for Federal state HES (EF1.5), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 5	159	159	3	2	7	4
2	1 2 1	225	225	6	3	10	6
3	1 2 2	453	453	9	5	20	12
4	1 2 3	880	880	12	10	40	23
5	1 2 4 -1 2 5	112	112	2	2	5	3
6	1 3 1	889	889	24	10	40	23
7	1 3 2	770	770	21	9	35	20
8	1 3 3	603	603	16	7	27	16
9	1 3 4 -1 3 5	48	48	2	2	2	2
10	1 4 1 -1 4 5	504	504	8	6	23	13

Table C.9: Stratification plan implemented for Federal states RLP and SAL (EF1.6), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
11	2 1 1 -2 1 3	72	144	2	2	3	2
12	2 1 4 -2 1 5	149	297	3	2	7	4
13	2 2 1 -2 2 3	617	1231	13	7	28	16
14	2 2 4	392	782	7	5	18	10
15	2 2 5	258	514	4	3	12	7
16	2 3 1	103	206	3	2	5	3
17	2 3 2	388	776	10	4	17	10
18	2 3 3	1269	2538	38	15	57	33
19	2 3 4	242	483	7	3	11	6
20	2 3 5	115	229	3	2	5	3
21	2 4 1 -2 4 2	100	198	2	2	5	3
22	2 4 3 -2 4 5	120	239	2	2	5	3
23	3 1 1 -3 1 5	36	88	3	2	2	2
24	3 2 1	38	85	4	2	2	2
25	3 2 2	76	179	9	2	3	2
26	3 2 3 -3 2 5	178	474	16	2	8	5
27	3 4 1*	62	142	6	2	3	2
28	3 4 2*	58	156	4	2	3	2
29	3 4 3 -3 4 5*	29	86	3	2	2	2
30	4 1 1 -4 1 3	93	354	4	2	4	2
31	4 1 4	60	235	3	2	3	2
32	4 1 5	94	377	6	2	4	2
33	4 2 1	9	33	2	2	2	2
34	4 2 2	39	152	3	2	2	2
35	4 2 3	926	3480	45	11	42	24
36	4 2 4	637	2470	23	7	29	16
37	4 2 5	405	1610	15	5	18	10
38	4 4 1 -4 4 2*	89	334	3	2	4	2
39	4 4 3 -4 4 5*	147	590	4	2	7	4
40	5 1 1 -5 1 5*	129	419	6	2	6	3
41	5 2 1 -5 2 2	31	77	2	2	2	2
42	5 2 3	237	673	17	3	11	6
43	5 2 4	224	702	15	3	10	6
44	5 2 5	281	1004	14	3	13	7
45	5 4 1*	31	71	2	2	2	2
46	5 4 2*	52	127	3	2	2	2
47	5 4 3*	277	749	8	3	12	7
48	5 4 4*	164	489	3	2	7	4
49	5 4 5*	102	318	2	2	5	3

Table C.10: Stratification plan implemented for Federal states RLP and SAL (EF1.6), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1	93	93	2	2	4	2
2	1 1 2	53	53	2	2	2	2
3	1 1 3	154	154	2	2	7	4
4	1 1 4	43	43	2	2	2	2
5	1 1 5	62	62	2	2	3	2
6	1 2 1	536	536	11	6	24	14
7	1 2 2	1042	1042	17	12	47	27
8	1 2 3	2072	2072	26	24	93	54
9	1 2 4	234	234	2	3	11	6
10	1 2 5	82	82	2	2	4	2
11	1 3 1	1323	1323	49	15	60	34
12	1 3 2	1456	1456	44	17	66	38
13	1 3 3	1232	1232	33	14	56	32
14	1 3 4 -1 3 5	120	120	3	2	5	3
15	1 4 1	704	704	15	8	32	18
16	1 4 2	129	129	2	2	6	3
17	1 4 3 -1 4 5	75	75	2	2	3	2
18	2 1 1 -2 1 2	26	52	2	2	2	2
19	2 1 3	157	313	3	2	7	4
20	2 1 4	112	224	2	2	5	3
21	2 1 5	225	446	3	3	10	6
22	2 2 1	20	40	2	2	2	2
23	2 2 2	78	154	3	2	4	2
24	2 2 3	1093	2178	24	13	49	28
25	2 2 4	918	1828	14	11	41	24
26	2 2 5	650	1293	9	8	29	17
27	2 3 1	137	274	7	2	6	4
28	2 3 2	447	894	21	5	20	12
29	2 3 3	2341	4678	76	27	105	61
30	2 3 4	531	1060	15	6	24	14
31	2 3 5	278	555	7	3	13	7
32	2 4 1	57	114	2	2	3	2
33	2 4 2	105	209	2	2	5	3
34	2 4 3 -2 4 5	287	572	2	3	13	7
35	3 1 1	4	9	2	2	2	2
36	3 1 2	12	33	2	2	2	2
37	3 1 3	22	54	3	2	2	2
38	3 1 4 -3 1 5	14	37	2	2	2	2
39	3 2 1	67	158	9	2	3	2
40	3 2 2	151	351	17	2	7	4

Table C.11: Stratification plan implemented for Federal state BW (EF1.7), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	3 2 3	368	952	29	4	17	10
42	3 2 4	36	100	3	2	2	2
43	3 2 5	18	54	2	2	2	2
44	3 4 1*	75	190	12	2	3	2
45	3 4 2*	76	197	10	2	3	2
46	3 4 3 -3 4 5*	55	188	4	2	2	2
47	4 1 1	7	27	2	2	2	2
48	4 1 2	15	63	2	2	2	2
49	4 1 3	147	588	8	2	7	4
50	4 1 4	138	541	6	2	6	4
51	4 1 5	299	1229	10	3	13	8
52	4 2 1	15	52	3	2	2	2
53	4 2 2	59	211	5	2	3	2
54	4 2 3	1976	7527	92	23	89	51
55	4 2 4	1480	5834	47	17	67	38
56	4 2 5	1069	4428	29	12	48	28
57	4 4 1*	38	139	3	2	2	2
58	4 4 2*	86	328	5	2	4	2
59	4 4 3*	225	901	8	3	10	6
60	4 4 4*	50	211	2	2	2	2
61	4 4 5*	30	125	2	2	2	2
62	5 1 1 -5 1 3	56	159	3	2	3	2
63	5 1 4	50	166	3	2	2	2
64	5 1 5	187	628	7	2	8	5
65	5 2 1	14	34	2	2	2	2
66	5 2 2	44	102	3	2	2	2
67	5 2 3	480	1290	35	6	22	12
68	5 2 4	512	1675	28	6	23	13
69	5 2 5	751	2700	28	9	34	19
70	5 4 1*	51	121	3	2	2	2
71	5 4 2*	95	227	6	2	4	2
72	5 4 3*	478	1259	16	6	22	12
73	5 4 4*	270	816	6	3	12	7
74	5 4 5*	246	792	4	3	11	6

Table C.12: Stratification plan implemented for Federal state BW (EF1.7), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1	139	139	2	2	6	4
2	1 1 2	122	122	2	2	5	3
3	1 1 3	262	262	3	3	12	7
4	1 1 4	87	87	2	2	4	2
5	1 1 5	109	109	2	2	5	3
6	1 2 1	597	597	12	7	27	15
7	1 2 2	1406	1406	20	16	63	36
8	1 2 3	2541	2541	31	29	115	66
9	1 2 4	297	297	3	3	13	8
10	1 2 5	139	139	2	2	6	4
11	1 3 1	1825	1825	56	21	82	47
12	1 3 2	1717	1717	51	20	77	44
13	1 3 3	1407	1407	38	16	63	36
14	1 3 4	134	134	3	2	6	3
15	1 3 5	47	47	2	2	2	2
16	1 4 1	966	966	17	11	44	25
17	1 4 2	159	159	2	2	7	4
18	1 4 3 -1 4 5	100	100	2	2	5	3
19	2 1 1 -2 1 2	60	119	2	2	3	2
20	2 1 3	212	422	3	2	10	5
21	2 1 4	137	274	2	2	6	4
22	2 1 5	270	538	4	3	12	7
23	2 2 1	27	50	2	2	2	2
24	2 2 2	97	191	3	2	4	3
25	2 2 3	1481	2947	28	17	67	38
26	2 2 4	1158	2308	15	13	52	30
27	2 2 5	790	1570	10	9	36	20
28	2 3 1	283	565	8	3	13	7
29	2 3 2	760	1520	25	9	34	20
30	2 3 3	2748	5495	89	32	124	71
31	2 3 4	580	1157	17	7	26	15
32	2 3 5	284	567	8	3	13	7
33	2 4 1	83	166	2	2	4	2
34	2 4 2	127	252	2	2	6	3
35	2 4 3	297	590	2	3	13	8
36	2 4 4	64	127	2	2	3	2
37	2 4 5	47	94	2	2	2	2
38	3 1 1	19	43	2	2	2	2
39	3 1 2	10	24	2	2	2	2
40	3 1 3	39	102	3	2	2	2

Table C.13: Stratification plan implemented for Federal state BAY (EF1.8), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	3 1 4	8	21	2	2	2	2
42	3 1 5	12	37	2	2	2	2
43	3 2 1	87	189	11	2	4	2
44	3 2 2	239	581	20	3	11	6
45	3 2 3	438	1130	34	5	20	11
46	3 2 4	40	115	3	2	2	2
47	3 2 5	12	37	2	2	2	2
48	3 4 1*	114	270	15	2	5	3
49	3 4 2*	109	294	11	2	5	3
50	3 4 3*	89	293	5	2	4	2
51	3 4 4 -3 4 5*	10	33	2	2	2	2
52	4 1 1	15	60	2	2	2	2
53	4 1 2	30	109	2	2	2	2
54	4 1 3	227	890	9	3	10	6
55	4 1 4	223	861	7	3	10	6
56	4 1 5	408	1627	11	5	18	11
57	4 2 1	21	73	3	2	2	2
58	4 2 2	100	357	6	2	5	3
59	4 2 3	2388	9002	106	28	108	62
60	4 2 4	1696	6702	55	20	76	44
61	4 2 5	1216	4954	34	14	55	31
62	4 4 1*	36	122	3	2	2	2
63	4 4 2*	111	404	6	2	5	3
64	4 4 3*	349	1326	9	4	16	9
65	4 4 4*	81	323	2	2	4	2
66	4 4 5*	44	173	2	2	2	2
67	5 1 1 -5 1 2	21	63	2	2	2	2
68	5 1 3	82	244	3	2	4	2
69	5 1 4	89	292	3	2	4	2
70	5 1 5	214	741	8	2	10	6
71	5 2 1	13	27	2	2	2	2
72	5 2 2	58	126	4	2	3	2
73	5 2 3	668	1827	41	8	30	17
74	5 2 4	675	2187	33	8	30	17
75	5 2 5	867	3069	33	10	39	22
76	5 4 1*	61	148	4	2	3	2
77	5 4 2*	110	253	6	2	5	3
78	5 4 3*	706	1934	18	8	32	18
79	5 4 4*	430	1291	7	5	19	11
80	5 4 5*	323	1067	4	4	15	8

Table C.14: Stratification plan implemented for Federal state BAY (EF1.8), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 5	81	81	2	2	4	2
2	1 2 1	198	198	4	2	9	5
3	1 2 2	259	259	3	3	12	7
4	1 2 3 -1 2 5	251	251	2	3	11	6
5	1 3 1	444	444	17	5	20	11
6	1 3 2	516	516	15	6	23	13
7	1 3 3 -1 3 5	160	160	4	2	7	4
8	1 4 1 -1 4 5	372	372	5	4	17	10
9	2 1 1 -2 1 5	105	209	2	2	5	3
10	2 2 1	11	21	2	2	2	2
11	2 2 2	41	82	2	2	2	2
12	2 2 3	374	745	6	4	17	10
13	2 2 4	156	310	2	2	7	4
14	2 2 5	94	187	2	2	4	2
15	2 3 1	20	40	2	2	2	2
16	2 3 2	153	306	6	2	7	4
17	2 3 3	925	1850	27	11	42	24
18	2 3 4 -2 3 5	76	152	2	2	3	2
19	2 4 1	38	76	2	2	2	2
20	2 4 2	76	152	2	2	3	2
21	2 4 3 -2 4 5	96	191	2	2	4	2
22	3 1 1 -3 1 5	24	60	2	2	2	2
23	3 2 1	29	63	3	2	2	2
24	3 2 2	71	161	6	2	3	2
25	3 2 3	90	234	6	2	4	2
26	3 2 4 -3 2 5	11	32	2	2	2	2
27	3 4 1*	68	145	6	2	3	2
28	3 4 2*	48	138	3	2	2	2
29	3 4 3 -3 4 5*	17	57	2	2	2	2
30	4 1 1 -4 1 2	11	38	2	2	2	2
31	4 1 3	56	200	2	2	3	2
32	4 1 4	39	142	2	2	2	2
33	4 1 5	32	122	2	2	2	2
34	4 2 1	22	75	3	2	2	2
35	4 2 2	450	1588	25	5	20	12
36	4 2 3	243	879	9	3	11	6
37	4 2 4	147	555	3	2	7	4
38	4 2 5	20	69	2	2	2	2
39	4 4 1*	47	166	2	2	2	2
40	4 4 2*	129	492	3	2	6	3

Table C.15: Stratification plan implemented for Federal state BRA (EF1.9), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
41	4 4 3*	23	97	2	2	2	2
42	4 4 4 -4 4 5*	29	81	2	2	2	2
43	5 1 1 -5 1 3	26	84	2	2	2	2
44	5 1 4 -5 1 4	22	71	2	2	2	2
45	5 1 5 -5 1 5	7	13	2	2	2	2
46	5 2 1	33	77	2	2	2	2
47	5 2 2	218	613	12	3	10	6
48	5 2 3	153	494	7	2	7	4
49	5 2 4	95	339	4	2	4	2
50	5 2 5	27	56	2	2	2	2
51	5 4 1*	51	128	2	2	2	2
52	5 4 2*	232	650	6	3	10	6
53	5 4 3*	48	153	2	2	2	2
54	5 4 4*	13	55	2	2	2	2

Table C.16: Stratification plan implemented for Federal state BRA (EF1.9), Table 2

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
1	1 1 1 -1 1 5	294	294	6	3	13	8
2	1 2 1	819	819	14	9	37	21
3	1 2 2	1058	1058	14	12	48	27
4	1 2 3 -1 2 5	791	791	8	9	36	20
5	1 3 1	1749	1749	61	20	79	45
6	1 3 2	2291	2291	56	27	103	59
7	1 3 3 -1 3 5	814	814	16	9	37	21
8	1 4 1 -1 4 5	1477	1477	18	17	67	38
9	2 1 1 -2 1 5	402	800	4	5	18	10
10	2 2 1 -2 2 2	236	471	6	3	11	6
11	2 2 3	1454	2904	24	17	66	38
12	2 2 4	445	884	6	5	20	12
13	2 2 5	199	395	3	2	9	5
14	2 3 1	63	126	4	2	3	2
15	2 3 2	575	1149	21	7	26	15
16	2 3 3	3738	7474	101	43	168	97
17	2 3 4 -2 3 5	179	358	3	2	8	5
18	2 4 1	108	211	3	2	5	3
19	2 4 2	297	592	3	3	13	8
20	2 4 3 -2 4 5	391	778	2	5	18	10

Table C.17: Stratification plan implemented for Federal states SAC, SAA and THÜ (EF1.10), Table 1

stratum number	strata DESTATIS hhtyp, szh, hne	dataset sim study N_h HH	N_h Pers	9.870 HH acc. NSI n_h QS	n_h RS	9.870 HH n_h RS	5.775 HH n_h RS
21	3 1 1 -3 1 5	40	99	5	2	2	2
22	3 2 1	101	223	12	2	5	3
23	3 2 2	232	533	20	3	10	6
24	3 2 3	307	748	21	4	14	8
25	3 2 4 -3 2 5	23	72	2	2	2	2
26	3 4 1*	234	525	20	3	11	6
27	3 4 2*	141	373	11	2	6	4
28	3 4 3 -3 4 5*	46	147	3	2	2	2
29	4 1 1 -4 1 2	36	123	2	2	2	2
30	4 1 3	201	743	8	2	9	5
31	4 1 4	122	473	3	2	5	3
32	4 1 5	96	369	5	2	4	2
33	4 2 1	18	59	3	2	2	2
34	4 2 2	101	335	8	2	5	3
35	4 2 3	1641	5825	90	19	74	42
36	4 2 4	774	2911	34	9	35	20
37	4 2 5	371	1428	14	4	17	10
38	4 4 1*	64	215	5	2	3	2
39	4 4 2*	201	717	8	2	9	5
40	4 4 3 -4 4 5*	511	1945	10	6	23	13
41	5 1 1 -5 1 3	114	344	4	2	5	3
42	5 1 4 -5 1 4	78	264	3	2	4	2
43	5 1 5 -5 1 5	84	284	3	2	4	2
44	5 2 1	21	45	2	2	2	2
45	5 2 2	72	170	6	2	3	2
46	5 2 3	801	2282	45	9	36	21
47	5 2 4	544	1783	26	6	25	14
48	5 2 5	359	1277	16	4	16	9
49	5 4 1*	88	194	5	2	4	2
50	5 4 2*	179	431	8	2	8	5
51	5 4 3*	771	2182	20	9	35	20
52	5 4 4 -5 4 5*	310	1029	3	4	14	8

Table C.18: Stratification plan implemented for Federal states SAC, SAA and THÜ (EF1.10), Table 2

Bibliography

Deming, W. E.; Stephan, F. F. (1940) : On a least squares adjustment of a sampled frequency table when expected marginal totals are known. In: Annals of Mathematical Statistics, Vol. 11, pages 427-444.

Destatis (2004) : Feinkonzept zur Bereitstellung von Daten für die Gemeinschaftsstatistik über Einkommen und Lebensbedingungen (EU-SILC). Stand Februar 2004, 36 pages.

Deville, J.-C.; Särndal, C.-E. (1992) : Calibration Estimators in Survey Sampling. In : Journal of the American Statistical Association, Vol. 87, pages 376 - 382.

Deville, J.-C.; Särndal, C.-E. and Sautory, O. (1993) : Generalized Raking Procedures in Survey Sampling. In : Journal of the American Statistical Association, Vol. 88, pages 1013 - 1021.

Estevao, V. M.; Särndal, C.-E. (2003) : A new perspective on calibration estimators. Proceedings of the 2003 Joint Statistical Meeting - Section on Survey Research Methods, pages 1346-1356.

EUROSTAT, European Commission (2004) : Description of Target Variables: Cross-sectional and Longitudinal. Document EU-SILC 065/04, 335 pages.

EUROSTAT, European Commission (2008) : Algorithms to compute Overarching Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). WORKING GROUP S̈TATISTICS ON LIVING CONDITIONS. Document LC-ILC/11/08/EN - REV. 61 pages.

Fox, R. (2002) : An R and S-Plus companion to applied regression. Thousand Oaks, CA: Sage Publ, XVI + 312 pages.

Horvitz, D. G.; Thompson D. J. (1952) : A Generalization of Sampling Without Replacement from a Finite Universe. In : Journal of the American Statistical Association, 47, pages 663 - 685.

Jiang, J.; Lahiri, P. (2006) : Mixed model prediction and small area estimation. In: Test 15 (1) 196, 2006.

Körner, T.; Mayer, I; Nimmergut, A. (2003) : Haushalte Heute - Pilotstudie zur Umsetzbarkeit einer Dauerstichprobe befragungsbereiter Haushalte (Access-Panel) in der amtlichen Statistik. Stand September 2003. Wiesbaden, Statistisches Bundesamt.

Körner, T.; Meyer, I.; Minkel, H.; Timm, U. (2005) : LEBEN IN EUROPA - Die neue Statistik über Einkommen und Lebensbedingungen. In: Wirtschaft und Statistik, 11/2005, pages 1137-1152. DESTATIS.

Körner, T.; Nimmergut, A.; Nökel, J.; Rohloff, S. (2006) : Die Dauerstichprobe befragungsbereiter Haushalte - Die neue Auswahlgrundlage für freiwillige Haushaltsbefragungen. In: Wirtschaft und Statistik, 05/2006, pages 451-467. DESTATIS.

Münnich, R. et. al. (2004) : Workpackage 1: Variance Estimation in Complex Surveys: Deliverables 1.1 and 1.2. `http://www.dacseis.de` - IST-2000-26057-DACSEIS Reports.

Münnich, R. et. al. (2004) : Workpackage 10: Variance Estimation for Small Area Estimates: Deliverables 10.1 and 10.2. `http://www.dacseis.de` - IST-2000-26057-DACSEIS Reports.

Münnich, R.; Knobelspies, M. (2008) : Variablenselektion bei gebundener Hochrechnung. In: Austrian Journal of Statistics, Vol. 37, pages 335 - 347.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E. (2005) : Handbook on constructing composite indicators: methodology and user guide. OECD Statistics Working Paper STD/DOC(2005)3, `http://www.olis.oecd.org/olis/2005doc.nsf/LinkTo/std-doc(2005)3`.

Office for National Statistics et al. (2003) ; Heady, P.; Lehtonen, R. et al. : Report on the performance of the SStandardËstimators. Version 1.0/300503. EURAREA working paper IST-2000-26290.

Quatember, A. et. al. (2004) : Workpackage 2: Analysis of National Surveys: Deliverables 2.1 and 2.2. `http://www.dacseis.de` - IST-2000-26057-DACSEIS Reports.

Rao, C.R. (1965) : Linear Statistical Inference and its Applications. Wiley, New York. Wiley series in probability and mathematical statistics. XVIII + 522 pages.

Rao, J.N.K. (2003) : Small Area Estimation. Hoboken : John Wiley & Sons. Wiley series in survey methodology. XXIII + 313 pages.

Särndal, C.-E.; Swensson, B.; Wretman, J. (1992) : Model Assisted Survey Sampling. New York et al. : Springer (Springer Series in Statistics).

Sautory, O. (1991) : Redressement d'échantillons d'enquêtes auprès des ménages par calage sur marges. Internal working paper No. F 9103, National Institute for Statistics and Economic Studies (INSEE), Paris.

Sautory, O. (2003) : CALMAR 2: A new version of the CALMAR calibration adjustement program. In: Proceedings of Statistics Canada's Symposium 2003: Challenges in survey taking for the next decade.

Schroedter, J.; Yvonne, L.; Lüttinger, P. (2006) : Die Umsetzung der Bildungsskala ISCED-1997 für die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 und die Mikrozensen 1976-2004. ZUMA Methodenbericht 2006.

Stoop, I.; Jowell, R.; Mohler, P. (2002) : The European Social Survey: One Survey in Two Dozen Countries. In: Paper presented at the International Conference on Improving Surveys, Copenhagen, August 2002.

Vanderhoeft, C. (2003) : Generalised Calibration at Statistics Belgium; SPSS Module g-CALIB and Current Practices. Statistics Belgium Working Paper.