EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

Deliverable 7.1

Report on the Simulation Results

Version: 2011

Beat Hulliger, Andreas Alfons, Christian Bruch, Peter Filzmoser, Monique Graf, Jan-Philipp Kolb, Risto Lehtonen, Daniela Lussmann, Angelika Meraner, Ralf Münnich, Mikko Myrskylä, Desislava Nedyalkova, Jan Seger, Tobias Schoch, Matthias Templ, Maria Valaste, Ari Veijanen, and Stefan Zins

The project **FP7–SSH–2007–217322 AMELI** is supported by European Commission funding from the Seventh Framework Programme for Research.

http://ameli.surveystatistics.net/

Contributors to deliverable 7.1

Chapter 1: B. Hulliger, D. Lussman and T. Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 2: J.-P. Kolb, R. Münnich and S. Zins, University of Trier

Chapter 3: B. Hulliger, D. Lussman and T. Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 4: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 5: A. Alfons and M. Templ, Vienna Technical University

Chapter 6: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland.

Chapter 7: M. Graf and D. Nedyalkova, Swiss Federals Statistical Office ; R. Münnich, J. Seger and S. Zins, University of Trier

Section 7.1: M. Graf and D. Nedyalkova, Swiss Federals Statistical Office

Section 7.2: R. Münnich, J. Seger and S. Zins, University of Trier

Chapter 8: R. Lehtonen, A. Veijanen, M. Myrskylä and M. Valaste, University of Helsinki

Chapter 9: C. Bruch, R. Münnich, T. Zimmermann and S. Zins

Chapter 10: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland; A. Alfons, J. Holzer, P. Filzmoser, A. Meraner and M. Templ, Vienna Technical University

Section 10.2: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland

Section 10.3: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland

Section 10.4: A. Alfons, M. Templ, P. Filzmoser and J. Holzer, Vienna Technical University

Section 10.5: B. Hulliger and T. Schoch, University of Applied Sciences Northwestern Switzerland

Main responsibility

Beat Hulliger, University of Applied Sciences Northwestern Switzerland.

Evaluators

Internal experts:

Oliver Bode, *Federal Statistical Office of Germany*
Thomas Burg, *Statistics Austria*
Rudi Seljak, *Statistical Office of the Republic of Slovenia*

© http://ameli.surveystatistics.net/ - 2011

Aim and Objectives of Deliverable 7.1

Deliverable 7.1 of the AMELI project brings together the results of the various simulations under Workpackage 7 into one reference document. All research partners of the AMELI project have contributed with methods, simulations and a corresponding chapter. NSI partners of the AMELI project have contributed through evaluation of the results and of the deliverable. Workpackage 7 and Deliverable D7.1 were coordinated by FHNW.

The Deliverable was split into the main deliverable D7.1 and the Appendix, called D7.1-Appendix. The main deliverable gives an overview over the simulation setup in Part I. The reports on the simulations and recommendations can be found in Part II. The main deliverable is well suited for printing. It is a reference source for the results in the form of tables and graphs and contains R-code used for specific tasks in the simulations.

The chapters and sometimes the sections of the deliverable mention the names of the responsible authors. This reflects the heterogeneity of the simulations and methods well. Chapters also contain chapter bibliographies specific to the chapter.

Contents

© http://ameli.surveystatistics.net/ - 2011

© http://ameli.surveystatistics.net/ - 2011

Part I

Simulation Setup

© http://ameli.surveystatistics.net/ - 2011

Chapter 1

Introduction

1.1 Overview

The methods to be evaluated with the simulations have been developed under workpackages 2, 3 and 4 of the AMELI project. The corresponding deliverables (Graf et al., 2011), (Lehtonen et al., 2011), (Bruch et al., 2011) and (Hulliger et al., 2011b) describe the methods and preliminary simulations in detail.

The construction of the universes AMELIA and AAT-SILC for the simulations are described in (Alfons et al., 2011). These universes themselves use the real data sets EU-SILC, which is a collection of the European Statistics on Income and Living Conditions Databases maintained by Eurostat, and AT-SILC, which is the Austrian version of the SILC data.

Four simulation environments have been used for the simulations of the AMELI project: The simulation environment of the University of Trier (UT) based on the system Condor, the simulation environment of the University of Helsinki (UH), a new R-based simulation environment from University of Vienna (TUW), and a database-simulation environment built on top of simFrame by University of Applied Sciences Northwestern Switzerland (FHNW).

In this deliverable neither the simulation universes AMELIA nor the simulation environments are described. The specification of the simulation bed, sometimes called the scenarios, are described in the first part of the deliverable. In particular Chapter 2 explains the sample designs, Chapter 3 gives a proposal for domains which might be useful for the simulations, Chapter 4 discusses the contamination mechanisms and introduces the corresponding nomenclature, Chapter 5 introduces the missingness mechanisms and Chapter 6 defines the criteria for the analysis of results for later reference. The methods are shortly summarized in the corresponding chapters in Part II. These chapters deal with parametric estimation in Chapter 7, small area estimation in Chapter 8, variance estimation in Chapter 9 and univariate as well as multivariate robustness in Chapter 10.

1.2 Workflow and Metadata

The simulations of the AMELI project have used several simulation environments, several hardware platform and a range of methods. Nevertheless the coordination of the simulation set up and an overview of the simulation results was ensured. To understand the coordination process and the gathering of the information about the analysis carried out by the different partners the workflow of the information and the format of the resulting metadata is explained in this chapter.

Figure 1.1 describes the analysis process for the AMELI Results. Each simulation partner is responsible for maintaining the data of its simulations and for delivering a report. Simulation partners are UT, UH, TUW, Swiss Federal Statistical Office (SFSO) and FHNW. Each simulation which is used in the analysis must be accompanied by the Form AMELI Simulation Metadata. This form will give an overview over the simulations and permits to identify where the simulation data actually can be obtained. Each partner holds his own data and he is capable to provide simulation data as a simFrame R-object or in an object which inherits from it, such that the extractor functions of simFrame work.

For his simulations each partner delivers one metadata simulation form per simulation run. Metadata for a simulation run contains compulsory and voluntary metadata and voluntary results. The AMELI project has set up an SVN server for data management and exchange (http://svn.uni-trier.de/AMELI). Each metadata form is uploaded into the partner directory on the AMELI SVN server. A metalist is maintained which lists all existing metadata forms. Figure 1.2 shows where the different elements of the simulation results will be stored. Note that some of the data is stored both centrally on the SVN-Server and locally by each partner.

FHNW maintains a list with the metadata (Metalist, see Figure 1.3) of the uploaded files and a link to each metadata simulation form. If a partner wants to obtain data from another partner he must require it using the metadata identification (id) and it will be sent in simFrame format. Thus partners are responsible for archiving their data and for ensuring access.

The metalist is a compilation of all compulsory metadata (Figure 1.4). A metadata simulation form contains compulsory and any voluntary metadata from one simulation run, voluntary simulation results and any text a partner deems necessary to explain the simulation run. The part with the voluntary metadata, the results and text has no specification and can be filled out arbitrary. The compulsory metadata-part has to respect specifications in order to allow automatic transfer into the metalist. The content of the compulsory metadata is explained in detail in Table 1.1 and shown in Figure 1.4. Only the compulsory metadata have to be delivered obligatory, every other data is optional. Finally there will be one metalist linked to numerous metadata simulation forms. The metalist is reproduced in (Hulliger et al., 2011a, chap. 1).

The metadata simulation form can be delivered in a text file format (.csv) which could afterwards be imported in an Excel-sheet. Figure 1.5 shows how the text file could look like.

The reports on the simulation runs of a simulation partner are adapted to the needs of the methods and the simulations run by that partner. However, every report should contain

© http://ameli.surveystatistics.net/ - 2011

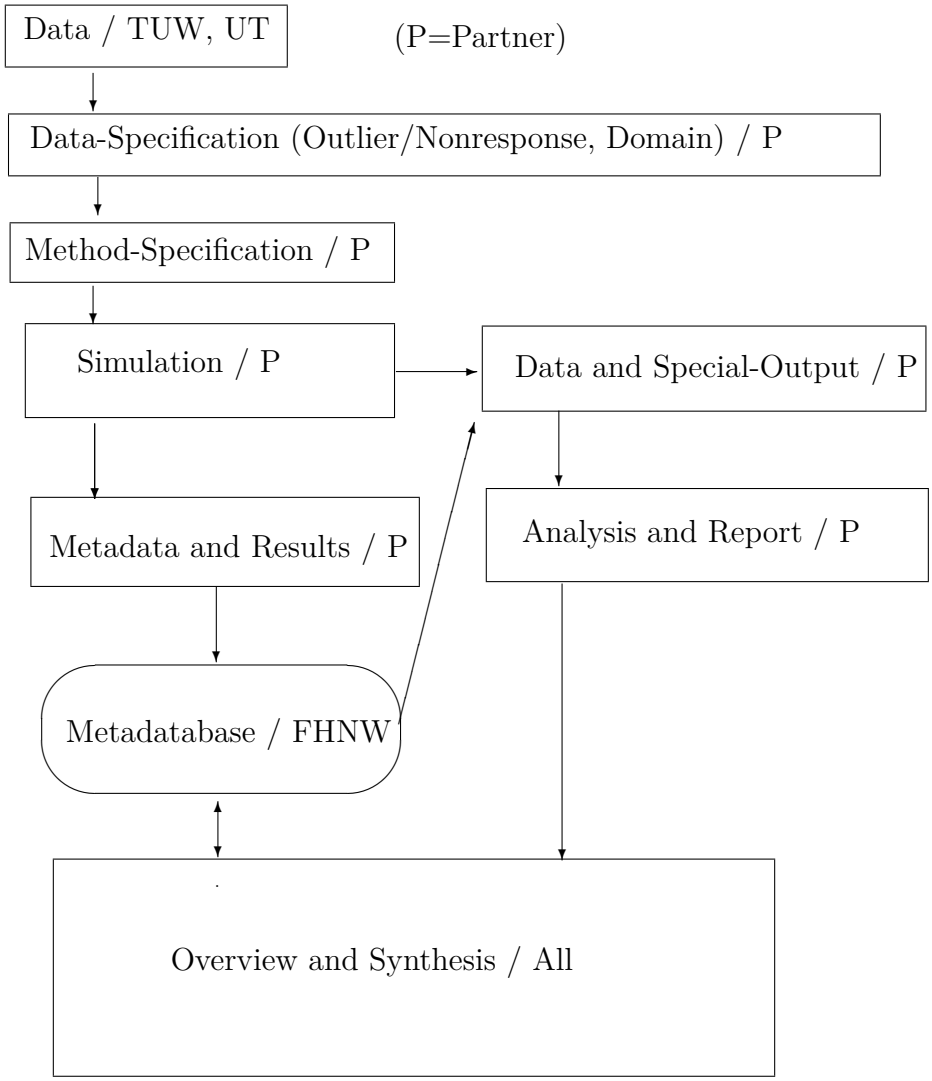


Figure 1.1: Analysis process

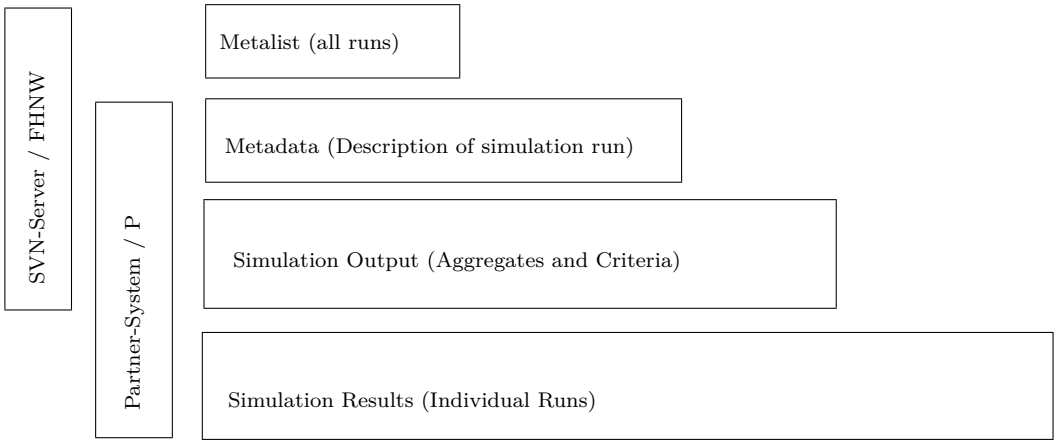


Figure 1.2: Metadata storage

recommendations for the applications of the methods. The reports are collected in Part II of this deliverable.

Metalist.xls

	A	B	C	D	E	F	G	H	I	J	
1	mdid	id	author	data	mean	arpr	rmpg	qsr	gini	universe	dom
2	1	N0045	FHNW	20101003	1	0	0	0	0	AMELIA	
3	2	N0046	FHNW	20101003	0	1	0	0	0	AMELIA	
4	3	N0047	FHNW	20101013	0	0	0	1	0	AAT-Silc	
5	4	N0048	FHNW	20101015	0	0	1	0	0	AMELIA	
6	5	N0049	FHNW	20101015	0	0	1	1	0	AMELIA	
7	6	N0050	FHNW	20101015	0	0	1	0	0	AMELIA	
8	7	N0051	FHNW	20101022	0	1	0	0	0	AAT-Silc	

Figure 1.3: Metalist

Compulsory metadata			Version 30.11.2010
Metadata-Label	Values	Name	Remarks
Identification		id	author can find the data in his database with identifier (e.g. N0001)
Author/Institution		author	
Date		date	Finishing date of simulation (yyyymmdd)
Mean		mean	1=used, 0=not used
ARPR		arpr	1=used, 0=not used
RMPG		rmpg	1=used, 0=not used
QSR		qsr	1=used, 0=not used
Gini		gini	1=used, 0=not used
Other		other	1=used, 0=not used; to be documented in voluntary part
Universe		universe	AMELIA, AAT-SILC, OTHER
Domain(s)		domain	laf, 22, n4f, OTHER
Sample design		design	1.2, 1.4a, 1.5a, 2.6, 2.7, OTHER
Sample size (nbr households)		size	
Outlier and cont. mechanism		oc	OCAR-CCAR, OCAR-NCAR, OAR-CCAR, OAR-NCAR, OTHER
Parameters outliers/contam.		ocpar	Free text
Nonresponse mechanism		nr	MCAR, MAR, OTHER
Parameters NR-mechanism		nrpar	Free text
Number of replicates		nrep	
Simulation environment		sw	simFrame, Dbsim, condor, OTHER

Figure 1.4: Compulsory metadata

Bibliography

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**): *Synthetic Data Generation of SILC Data*. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Bruch, C., **Münnich, R.** and **Zins, S.** (**2011**): *Variance Estimation for Complex Surveys*. Research Project Report WP3 – D3.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Graf, M., **Nedyalkova, D.**, **Münnich, R.**, **Seger, J.** and **Zins, S.** (**2011**): *Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion*. Research Project Report WP2 – D2.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

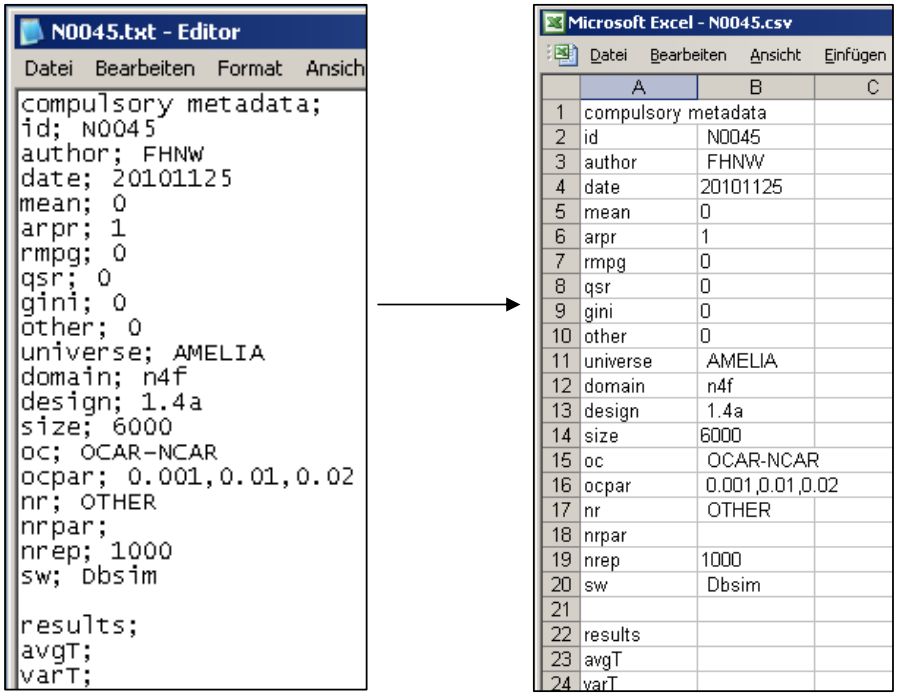


Figure 1.5: Textfile for metadata (left) and Excel-sheet (right)

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**a): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Hulliger, B., **Alfons, A.**, **Filzmoser, P.**, **Meraner, A.**, **Schoch, T.** and **Templ, M.** (**2011**b): *Robust Methodology for Laeken Indicators*. Research Project Report WP4 – D4.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Lehtonen, R., **Veijanen, A.**, **Myrskylä, M.** and **Valaste, M.** (**2011**): *Small Area Estimation of Indicators on Poverty and Social Exclusion*. Research Project Report WP2 – D2.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Table 1.1: Description of compulsory metadata

Name	Description
id	the identification must contain a capital letter for each partner (see below) followed by a sequence number (4 digits). The sequence number has to be assigned and administered by the partner. Capital letters for the id: H = University of Helsinki, N = University of Applied Sciences Northwestern Switzerland, S = Swiss Federal Statistical Office, T = University of Trier, W = Vienna University of Technology
author	the author/institution of a simulation run can be UT, UH, TUW, SFSO or FHNW
date	means the finishing date of simulation run. Format: yyyymmdd
mean, arpr, rmpg, gsr, gini	the used indicators must be binary coded (1=used, 0=not used)
other	if there were other indicators used than the listed ones, it is possible to add other indicators by using the label öthersänd defining it in the voluntary part of the metatdata (1=used, 0=not used)
universe	the universe is either AMELIA, AAT-SILC or OTHER. If there was used another universe than AMELIA or AAT-SILC it can be defined in the voluntary part of the metadata
domain	possible analysis domains: laf, 22, n4f, OTHER (see specifications in chapter 3). If there was used another domain than laf, 22 or n4f it can be defined in the voluntary part of the metadata
design	possible sample designs: 1.2, 1.4a, 1.5a, 2.6, 2.7, OTHER (see table 2.1). Another used sample design can be documented in the voluntary metadata part.
size	the sample size is counted by households and is absolute (not counted as sampling fraction)
oc	possible outlier and contamination mechanisms: OCAR-CCAR, OCAR-NCAR, CAR-CCAR, OAR-NCAR, OTHER; the first part of the reference (e.g. OCAR) defines the outlier model (see 4.2.1) and the second part (e.g. CCAR) indicates the contamination model (see 4.2.4).
ocpar	the parameters for outliers and contamination can be described by using free text
nr	the references for the missing data mechanisms are: MCAR, MAR, OTHER (see specifications in chapter 5.2).
nrpar	parameters for nonresponse-mechanism can be described by using free text
nrep	number of replicates
sw	the simulation environment can be simFrame, Dbsim, Condor or OTHER (other simulation environments can be described in the voluntary metadata part)

© http://ameli.surveystatistics.net/ - 2011

Chapter 2

Sampling Designs

In the following the five core sampling designs are shown in Table 2.1. Table 2.2 indicates the possible sampling fractions for a fix sample size of $n = 6000$. All sampling fractions refer to the AMELIA data set. The variable sample sizes are reported here for coordination purposes and as a proposal. The basic design for the AMELI simulations will use the fixed sample size sampling fractions with 6000 households.

Table 2.1: Sample designs

	$p_1(\cdot)$					$p_2(\cdot)$				
ID	PSU	Strata	π_{iI}	SM	Alloc.	SSU	Strata	π_{iII}	SM	Alloc.
1.2	HID	–	srs	1	–	–	–	–	–	–
1.4a	HID	NUTS2	srs	1	prop	–	–	–	–	–
1.5a	HID	NUTS2	pps	HHG	prop	–	–	–	–	–
2.6	CIT	NUTS2*DOU	srs	1	prop	HID	–	pps	HHG	–
2.7	CIT	NUTS2*DOU	srs	1	prop	HID	–	srs	–	–

Notes: SM: measure of size; Alloc: allocation; prop.: proportional; srs: simple random sampling without replacement; pps: sampling proportional to size (Midzuno); $p_k(\cdot)$: sampling design at the kth stage; π_{iI} and π_{iII}: sample inclusion probability at the first and second stage; *Variables*: HID: household identifier; HHG: household size; CIT: municipality identifier (LAU1); DOU: degree of urbanization.

Table 2.2: Sampling fractions

	variable sample size			fixed sample size, $n = 6000$		
ID(s)	f_I	f_{II}	f	f_I	f_{II}	f
1.2 / 1.4a / 1.5a	1%; 5%	–	1%; 5%	0.16%	–	0.16%
2.7 / 2.6	5%; 1.25%	20%; 80%	1%	16%	0.1%	0.16%
2.7 / 2.6	25%; 6.25%	20%; 80%	5%	–	–	–

Chapter 3

Analysis Domains

In order to evaluate simulation results not only on the level of the whole population but also on the level of domains for both universes two similar domains are proposed.

The two proposed domains of analysis with AAT-SILC are

1. dom.laf<-db040 == ”Lower Austria“ AND rb090== ”female“
2. dom.22<-hsize==4 AND eqSS==2.1

The first domain consists of the women in Lower Austria. It has a size of $N_D = 77\,0367$ or 9.4% of the total population in the universe. It should be rather homogenous in terms of income distributions and could be an aggregate of domains used in Small Area Estimation (SAE)

The second domain consists of persons in households with 2 adults and 2 children. It has a size of $N_D = 78\,4944$ or 9.6% of the total population. This domain may cut across the SAE domains. The size of both domains is large enough that a direct estimation is reasonable.

Two domains of analysis for AMELIA are

1. dom.n4f<-NUTS2==4 AND SEX=2
2. dom.22<-HHG==4 AND children==2

The first domain, women in the fourth simulated NUTS2 region, has size $N_D = 694\,837$ or 6.9% of the total AMELIA population. The second domain has size $N_D = 459\,504$ or 4.6%.

© http://ameli.surveystatistics.net/ - 2011

Chapter 4

Outlier and Contamination Settings

4.1 Introduction

We attempt to describe the contamination set-up for simulation bed derived from our needs. The notation and set-up described here follows the Milestone M4.2 description. The ingredients of the mechanisms we investigate are the following:

- U is the set of elements in the population of size N. We usually use the index i to indicate the elements of U.
- Y_i^* the true, complete data. For the description of the mechanisms, we think of the true Y_i^* as of random variables, which follow a superpopulation model. In the survey context we will fix them as one realisation from a superpopulation model. Any finite population characteristics would be a function of Y^*.
- Y_i is the observable data for unit i.
- R_{ij} denotes a response indicator for the i th observation of variable j. Given the vector $R_{i\cdot}$ we can split the observable data Y_i into an observed part Y_{io} and a missing part Y_{im}, which in turn may be composed as $Y_i = (Y_{io}, Y_{im})$.
- Y_{ic} is the contaminated data for unit i.
- O_i is an outlier indicator. For those observations with $O_i = 1$ the observable data Y_i is replaced by the contaminated data Y_{ic}. Thus the observed data Y_{io} actually consists of Y_{ico}. An important difference to R is that O is not directly observable. The outlier indicator thus is a latent variable.
- S_i is the sample indicator.
- X_i denotes covariables which are fully observed.
- Z_i denotes unobserved covariables.
- $\hat{Y}_i$ is imputed data, possibly after detection of outliers and imputation for outliers, i.e. $\hat{Y} = I(Y_o, X|D)$.

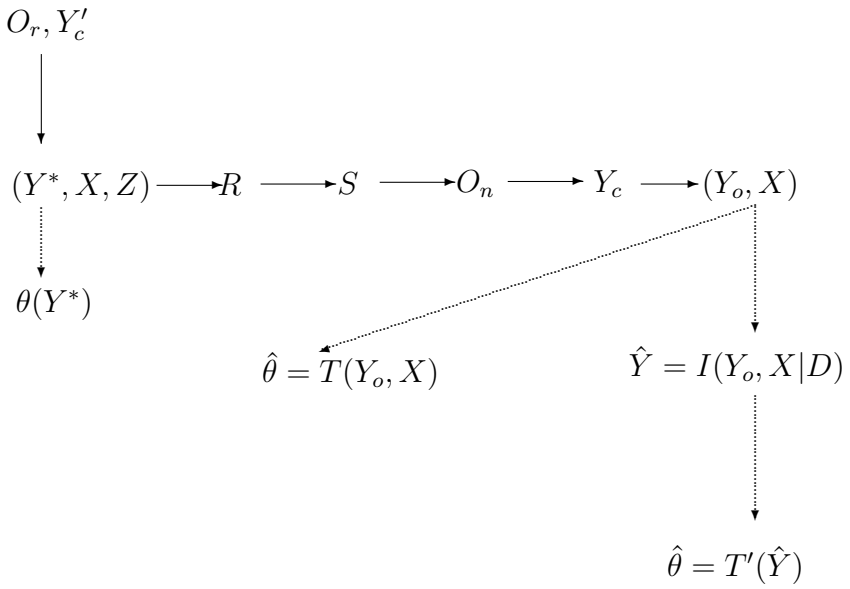


Figure 4.1: A model for the simulations

4.1.1 Setup

We will not consider here a process for representative outliers. In other words we consider Y_i^* as our starting point and assume therefore that representative outliers have been created already before and integrated into Y_i^*. Thus the outlier indicator O_i refers to non-representative outliers only. In Figure 4.1 the representative outliers are created by the process O_{ir} while the non-representative outliers are created by O_{in}. More so, we assume that the sampling mechanism is ignorable and we suppress it in the notation in this section. However, the models we are referring to, always pertain to the population and the sample design should be taken into account when estimating and testing the model parameters.

Furthermore, we assume that the observable data is not subject to other error mechanisms than the outlier process. In Figure 4.1, we show the relationship among the mechanisms and the process of estimation. Note that the exact sequence between the processes may be different. For instance, the response may depend on the sample and then in the process R would follow S. A scheme (scenario) consists of a complete description of all the elements in Figure 4.1. We consider both populations AMELIA and AAT-SILC. The units of the population are persons, which are grouped in households. For outlier schemes, the following variables are needed for each person.

1. Equivalized disposable income
2. Total personal income (i.e., aggregated persons income components).
3. Components of personal income:
   - income from employment (employed and self-employed)
   - income from transfers

© http://ameli.surveystatistics.net/ - 2011

4. Household income per capita (i.e., aggregated household-income components, divided by the size of the household; not equivalization size).

5. Components of household income per capita (i.e., aggregated and uniformly distributed over all household members):

   - income from employment (of children or not attributed)
   - income from capital
   - income from transfers

The outlier/contamination mechanism can be split up into two parts: in the first part, an observation is declared as an outlier. The second part is concerned with the contamination mechanism that creates the contamination. The outlier mechanism should be OCAR (outlying completely at random) and OAR (outlying at random) and non-ignorable, for example depending on the level of income.

The contamination may be created at the population level, i.e. $Y_{ci}, i \in U$ if we assume that there is no dependence of the contamination on the response, sampling and outlyingness mechanisms R, S and O. Further, the number of outliers in a sample is fixed. Therefore the outlier mechanism depends on the sampling mechanism and it can be simulated only after the sample has been drawn.

When it comes to detection, imputation or estimation two schemes should be considered for robust estimators:

1. The number of outliers is assumed known and therefore can be used for the determination of tuning constants. This may be useful in order to determine the relative properties of procedures conditional on the choice of the tuning constant.

2. The number of outliers is assumed not to be known and can thus not be used for the determination of tuning constants. This may be the more realistic scheme in practice but may entail problems when comparing the relative merits of procedures due to the additional effect of having to choose the tuning constant.

It may be argued that the second proposal is not realistic in a situation where a survey is repeated frequently and very good past knowledge about the amount of contamination is available.

4.1.2 Outlyingness- and Contamination Typology/Nomenclatura

As to outlyingness mechanism, we confine ourselves to study the OCAR- and the OAR-mechanism (see Table 4.1.2). Concerning contamination mechanisms, we consider the mechanisms: CCAR, CAR, and NCAR (see Table 4.1.2).

Table 4.1: Outlyingness Mechanisms

	Outlying completely at random	Outlying at random
Description	distribution of outliers is determined completely at random	distribution of outliers depends on covariates
`OType`	`OCAR`	`OAR`
`OTypeSpec`	`[NULL]`	outlier-determining variable and a corresponding outlier-probability vector
`OTypeEpsilon`	`0.001, 0.02`	`0.001, 0.02`

`OTypeEpsilon` denotes the number of outliers. These numbers correspond to 0.01% and 0.1% outliers in relative terms. `OTypeSpec` defines the mechanism.

Table 4.2: Contamination Mechanisms

	Contaminated completely at random	Contaminated at random	Not contaminated at random
Description	all observations are identically contaminated	the contamination distribution depends on observable covariates	the contamination distribution depends on the variable to be contaminated (and/or unobserved covariates)
`CType`	`CCAR`	`CAR`	`NCAR`
`CTypeMech`	$f(x)$	$f(x\|\theta)$	contamination distribution depends on x
`CTypeMechSpec`	specification of f	specification of f and θ	[e.g., `scale`$\cdot x$]

4.2 Univariate Outlyingness/Contamination

For all univariate methods (whether direct-, model-assisted-, or model-based estimation) we consider outliers exclusively in the (univariate) equivalized disposable income. The treatment of genuinely multivariate outliers, that is, jointly outlying observations in income components are to be destined only for the multivariate outlier-detection and imputation methods.

Moreover, both the outlyingness- and contamination mechanims operate on variables with data at individual level. Upon having declared observations as outliers and having contaminated them conformably, these contaminated observations are then re-distributed among the household members.[1] In particular, each household member with at least one

[1]Since outlyingness is defined on individual level, it may occur that several individuals of the same households have been declared as outliers. Moreso, the probability of multiple outlying persons per household depends on the household size. We confine ourselves to consider each household only once (even if more than one household member have been declared as outlier). As a result, the number of outlying households is not constant. This in turn may introduce some variability due to the variable number of outlying household. On the other hand, this outylingess schemes assures consisteny at the

© http://ameli.surveystatistics.net/ - 2011

öutlying individualïs assigned the outlying income divided by the household equivalization scale (i.e., modified OECD scale). Therefore, we can assure that all household members have the same equivalized disposable income; and thus the data are in line with the EUROSTAT definition. On the other, we note that the equivalization operation leads to treating outliers differently for different household sizes. Most notably, outliers that may be considered as far from the bulk of data in the case of one-person households may be rendered ßmallfor huge households.

4.2.1 Outlier Models

We consider only the outlier mechanisms OCAR (outlying completely at random) and OAR (outlying at random). The inverse-probability weights associated with the contaminated observations are obtained from the original, uncontaminated observations.

4.2.2 OCAR: Outlying Completely at Random

The OCAR outlyingness mechanism randomly assigns $O_i = 1$ for outlying observations and $O_i = 0$ otherwise, according to amount of outliers ε. The simplest `OCAR` function takes the observation vector `x` and ε (e.g., 1% or 10%) as arguments and returns a vector of logicals (`TRUE`, `FALSE`). In fact only the size of the sample is derived from the observation vector.

4.2.3 OAR: Outlying at Random

As to the OAR mechanisms, we consider the strategy: suppose the variables `main activity status` is categorized into four groups. Each attribute has been allocated a different outlyingness-probabilty based on subject matter knowledge. Next, one declares observations as outliers conditional on the assigned probabilities.

In particular, the outlyingness probabilites for the OAR-mechanism are defined according to the variable `main activity status` (i.e., variable `RB210` in AMELIA and `RB170` in AATSILC) (see Table 4.3). Thus, by means of the tabulated values, the outlying indicators O_i are obtained using an OCAR mechanism for each cell separately; note that the OAR-mechanism can be seen as an analogon to stratified sampling. Moreover, we normed the OAR-determining probabilities such that they sum up to one. Thus, the group-specific relative number of outliers follows from `OTypeEpsilon` $\cdot\,(0.7, 0.1, 0.15, 0.05)^T$. Due to the discretization (i.e, distributing the number of outliers to a small number of levels), the outlyingness probability may not be exactly proportional (rounding errors). Thus, the remaining number of outliers that has not been assigned to a certrain level due to the discretization are assigned to the level with the largest number of outliers (Hulliger et al., 2011, Chapter G). Alternatively, one may use the functions (due to Stefan Zins) that assign the level-specific outliers using a pps-type algorithm (Hulliger et al., 2011, Chapter G).

individual level. Alternatively, one may draw the outyling households w.r.t the household size (i.e. pps); this methods ensures a constant number of outlying households. An implementation of this scheme is displayed in (Hulliger et al., 2011).

Table 4.3: Probability distribution of OAR mechanism: Main activity status during the income reference period (EU-SILC User Data Base, variable: RB210; and RB170 in AAT-SILC)

At work	Unemployed	In retirement	Other inactive person
0.7	0.1	0.15	0.05

© http://ameli.surveystatistics.net/ - 2011

4.2.4 Contamination Models

Upon having declared some observations as outliers, these observations Y_i^* are replaced by the contaminated data Y_{ci}. Note that Y_{ci} may be simulated for each sample or it may be fixed at the population level, if it does not depend on other mechanisms.

For the univariate methods acting on the equivalized disposable income (i.e., no income components), we consider all three contamination set-ups.

4.2.5 CCAR: Contamination Completely at Random

Suppose the following CCAR contamination process for the equivalized disposable income: let the contamination distribution be defined as $Y_c \sim N(\mu, \sigma^2)$. Thus, the overall data, consisting of a fraction of $(1 - \texttt{OTypeEpsilon})$ good data (i.e., uncontaminated data) and a fraction `OTypeEpsilon` drawn according to the law of Y_c.

4.2.6 CAR: Contamination at Random

For the contamination at random mechanism we consider the following set-up: let the contamination distribution be defined as $Y_c \sim N(\mu_a, \sigma^2)$ for $a = \{1, \ldots, 4\}$, where a denotes the categories of a covariate, namely main activity status. Contrary to the CCAR setting, the location μ_a of the contaminated data in the CAR set-up is not the same for all $O_i = 1$, but it is defined conditional on the auxiliary variable `main activity status` (i.e., variable `RB2010` in AMELIA and variable `RB170` in AAT-SILC). The location parameters are shown in Tables 4.5 and 4.4 for AMELIA and AAT-SILC, respectively. By means of the tabulated location parameters and the contamination distribution, the bad data are obtained as in the case of the CCAR mechanism. (Similarly, we could also define a group-specific scale σ_a)

Table 4.4: Setup for the AMELIA data set: Location parameters μ_a ($a = \{1, \ldots, 4\}$) for the CAR mechanism conditional on the main activity status during the income reference period (variable RB210 in AMELIA)

At work	Unemployed	In retirement	Other inactive person
$3 \cdot 10^5$	$1 \cdot 10^5$	$2 \cdot 10^5$	$1 \cdot 10^5$

4.2.7 NCAR: Non-Ignorable Contamination

As a set up of practical interest which is non-ignorable (NCAR= not contaminated at random), we consider that the bad data follows by multiplying the true observation, Y_i^*, by a factor of 12. We may think of this type of contamination as of individuals reporting (approximately) their annual income when in actual fact they were asked to report the monthly income. Clearly, this operation acts differently on the observations, depending on the true value (thus the non-randomness of the contamination mechanism).

Table 4.5: Setup for the AAT-SILC data set: Location parameters μ_a ($a = \{1, \ldots, 4\}$) for the CAR mechanism conditional on the main activity status during the income reference period for variable: RB170 in AAT-SILC.

At work	Unemployed	In retirement	Other inactive person
$5 \cdot 10^5$	$2 \cdot 10^5$	$3 \cdot 10^5$	$2 \cdot 10^5$

4.2.8 Proposed Setup

In particular, we recommend the following combinations of outlyingness/contamination settings. The following snippets give a specification for the corresponding objects used by the simulation environment `simFrame`. An example call using these specifications follows in Subsection 4.2.14.

4.2.9 OCAR-CCAR-0.01

```
OCAR.CCAR <- list(OTypeEpsilon=0.01,
   OType="OCAR",
   OTypeSpec=NULL,
   CType="CCAR",
   CTypeMech="rnorm",
   CTypeMechSpec=list(mean=5e05, sd=2e04))
```

Note that for the AAT-SILC data set, you may set `mean=3e05` in `CTypeMechSpec` because the equivalized disposable income features less representative outliers (i.e., extreme observations that are already in the data).

4.2.10 OCAR-CCAR-0.001

```
OCAR.CCAR <- list(OTypeEpsilon=0.001,
   OType="OCAR",
   OTypeSpec=NULL,
   CType="CCAR",
   CTypeMech="rnorm",
   CTypeMechSpec=list(mean=5e05, sd=2e04))
```

Note that for the AAT-SILC data set, you may set `mean=3e05` in `CTypeMechSpec` because the equivalized disposable income features less representative outliers.

4.2.11 OCAR-NCAR

```
OCAR.NCAR <- list(OTypeEpsilon=0.001,
   OType="OCAR",
   OTypeSpec=NULL,
   CType="NCAR",
   CTypeMech=NULL,
   CTypeMechSpec=list(scale=12))
```

© http://ameli.surveystatistics.net/ - 2011

4.2.12 OAR-CCAR

```
OAR.CCAR <- list(OTypeEpsilon=0.01,
   OType="OAR",
   OTypeSpec=list(var="RB210", prob=c(0.7, 0.1, 0.15, 0.05)),
   CType="CCAR",
   CTypeMech="rnorm",
   CTypeMechSpec=list(mean=5e05, sd=2e04))
```

Again, note that for the AAT-SILC data set, you may set `mean=3e05` in `CTypeMechSpec` because the equivalized disposable income features less representative outliers.

Moreover, you must change the element `var` from `RB210` to `rb170` in the `OTypeSpec` object for the AAT-SILC data because of different naming.

4.2.13 OAR-NCAR

```
OAR.NCAR <- list(OTypeEpsilon=0.01,
   OType="OAR",
   OTypeSpec=list(var="RB210", prob=c(0.7, 0.1, 0.15, 0.05)),
   CType="NCAR",
   CTypeMech=NULL,
   CTypeMechSpec=list(scale=12))
```

Again, note that you must change the element `var` from `RB210` to `rb170` in the `OTypeSpec` object for the AAT-SILC data because of different naming.

4.2.14 An Example

You may find the specifications in the R-code snippets as list-objects that serves as argument for the `makeoutliers`-function (see Appendix). Here is an example,

Define the following OCAR-NCAR setup

```
setupOCAR.CCAR <- list(OTypeEpsilon=0.01,
   OType="OCAR",
   OTypeSpec=NULL,
   CType="CCAR",
   CTypeMech="rnorm",
   CTypeMechSpec=list(mean=1e05, sd=2e04))
```

On grounds of the defined OCAR-NCAR setup, one may generate outliers with the function-call of `makeoutliers`

```
dataOCAR.CCAR = makeoutliers(setupOCAR.CCAR, data, hid="db030", eqSS="eqSS", eqIncome="
    eqIncome",flag="TRUE")
```

where `data` denotes a `data.frame` that should be contaminated; `hid` specifies the variable name in `data` with the household-identification number; `eqSS` defines the name of the variable containing the equivalized household size; `eqIncome` denotes the variable name of the equivalized household income in `data`; by default `flag=TRUE`, which generates a new variable in `data` as flag of the outliers (1=outlier; 0=non-outlier).

4.3 Multivariate Outlyingness/Contamination

We start with some notation. Let Σ_0 denote a positive definite symmetric $(p \times p)$ covariance matrix of *good data* (i.e., majority of observations from a well-behaved population; usually we assume that the good data are a fraction $0 < 1 - \varepsilon < 1$ of the overall data). Similarly, let μ_0 be the $(p \times 1)$-dimensional location associated with the good data. Accordingly, we sometimes call the remaining observations (i.e., the fraction ε of the overall data) *bad data*. There is supposed to be no implication that the bad data are necessarily errors – they may just arise from a distinct subpopulation. Moreover, and because we confine ourselves to methods that are affine equivariant (with certain exceptions), Mahalanobis distances play a key role. Thus, the Mahalanobis distance between points $\mathbf{y}$ and $\mathbf{x}$ in $\mathbb{R}^p$ w.r.t. Σ_0 will be denoted by $d_{\Sigma_0}(\mathbf{x}, \mathbf{y})$.

For the multivariate outlier detection task, we consider only contaminated completely at random (CCAR) mechanisms (though our final proposal lets depend the contamination weakly on the true data for practical reasons). According to Rocke and Woodruff (1996), the hardest kind of outliers to find (for Mahalanobis-distance based methods), is the kind that has a covariance matrix, Ω, with the same shape as the good data (i.e., $\Omega = \lambda\Sigma_0$, for $\lambda \in \mathbb{R}^1$). This follows from the fact that $d_{\Sigma_0}(\mathbf{x}, \mu_0)$ is least for a bad point, $\mathbf{x}$, from μ_0 for $\Omega = \lambda\Sigma_0$. If this kind of outlier can be detected, then other kinds should be as well. Thus, we intend to focus on a situation in which the bad data are drawn from the same distribution as the good data and then displaced (shift outliers). That is, the location of the bad data is shifted by μ, where $|\mu| = \eta$. In particular, we first consider some CCAR set-ups: let the bad data comprise a fraction of ε of the overall data and let these be distributed according to the following laws,

1. $N_p(\mu_0 + \mu, \Omega)$, where $\Omega = \lambda\Sigma_0$, and $0 < \lambda \ll 1$. For η large, and the extreme case when $\lambda = 0$ (the contamination forms a point mass), then $\mathbb{E}[d_{\Sigma_0}(\mathbf{x}, \mu_0 + \varepsilon\eta)]$ is less for a bad point than a good point whenever $\varepsilon > 1/(p+1)$ (Rocke and Woodruff, 1996). As a result, the detection of a bad point becomes very difficult (if not infeasible) for Mahalanobis-distance based methods.

2. $N_p(\mu_0 + \mu, \Omega)$, where $\Omega = \lambda\Sigma_0$, $\lambda = 1$ (i.e., pure shift outliers) and μ_0, Ω are the same as before. Altough bad data might seem to be easily detectable under this contamination scheme, no method is known to find the outliers with complete assurance. This is because the overlap of the distributions of distances can be very substantial.

3. $N_p(\mu + \mu_0, \Omega)$, where (in this case) $\mu_0 := \text{median}(\mathbf{y})$, and $\mu = \lambda \cdot \text{median}(\mathbf{y}) \cdot \text{mad}(\mathbf{y})$, where $\lambda = 1.5$, and $\Omega = \text{diag}(\max[0.5 \min(\mu), \text{median}(\mu)])$ (Hulliger and Schoch, 2009).

4. Let the bad data be defined as $\tilde{\mathbf{y}}_c = (\mathbf{y}^* - \mu_0)\lambda + \mu_0 + \mu$, where $\lambda = 0.1$, and μ is the location shift parameter. The ith contaminated observation is obtained from $\mathbf{y}_{ic} = \tilde{\mathbf{y}}_{ic} \cdot \mathbb{1}\{\mathbf{y}_i^* \neq 0\}$. (The second operation replaces the contaminated observation by a zero if the true value was zero). Strictly speaking the contamination is not at random because it depends on Y^*, in particular the knowledge of the zero's. However, the dependence on the true data Y^* is weak when $\lambda << 1$. We propose to

© http://ameli.surveystatistics.net/ - 2011

use $\lambda = 0.1$ to obtain a scattered contamination with the same actual form of the original data but more concentrated. The advantage is that the original scatter Σ_0 must not be known.

The last two contamination set-ups are more data-driven.

In all of the above models the choice of the shift μ is crucial. We propose to choose as the contamination shift direction the eigenvector of $\mathbf{\Sigma_0}$ corresponding to a small eigenvalue. To avoid computational problems with eigenvalues close to zero it may be wise not to use the smallest eigenvalue. The length of the contamination shift should be chosen such that the location of the center of the contamination $\mu_\mathbf{0} + \mu$ lies on a high quantile of the original distribution (e.g. 99%).

We propose to use only contamination 4 in the simulations. The contamination direction and the length of the shift will have to be defined in a later version of this proposal.

4.4 Schemes

In an earlier version of this paper (June 3, 2010), we have proposed 5 OC-Schemes in first priority (see Table 4.6).

		Univariate			Multivariate
	Propensity	CCAR	CAR	NCAR	CCAR*
OCAR	$\epsilon = 0.01$	1	2	1	
OCAR	$\epsilon = 0.15$	1	3	3	
OAR	$\alpha = 1$	2	1	3	1
OAR	$\alpha = 2$	2	2	2	2

Table 4.6: Outlier and contamination schemes: 1 indicates first priority, 2 second, 3 third

In connection with the discussion during the AMELI consortium meeting in Trier (October 4 and 5, 2010), we recommend the following scenarios

- Univariate setup
  - OCAR-CCAR-0.01
  - OCAR-CCAR-0.001
  - OCAR-NCAR
  - OAR-CCAR
  - OAR-NCAR
- multivariate setup
  - OCAR-CCAR
  - OAR-CCAR

It goes without saying that one may simulate all other settings from Table 4.6 (e.g., one may simulate the OCAR-CCAR setting with $\varepsilon = \texttt{OTypeEpsilon} = 0.15$, which then may serve as an absolutely extreme case because of the exceptionally heavy number of outliers). However, the relevant scenarios are included in the above list.

Bibliography

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Hulliger, B. and **Schoch, T.** (**2009**): *Robust multivariate imputation with survey data*. Proceedings of the 57th Session of the International Statistical Institute, Durban.

Rocke, D. M. and **Woodruff, D. L.** (**1996**): *Identification of Outliers in Multivariate Data*. Journal of the American Statistical Association, 91 (435), pp. 1047–1061.

© http://ameli.surveystatistics.net/ - 2011

Chapter 5

Missing Data Models for Simulation

5.1 Introduction

This is a short description of the missing data scenarios for the simulations, including `R` (R Development Core Team, 2010) code specification for simulation studies using package `simFrame` (Alfons, 2010; Alfons et al., 2010b). From exploring the Austrian EU-SILC data with `R` package `VIM` (Templ et al., 2010) we derived simple rules for setting missing values (see also Templ et al., 2009). The purpose of this set of rules is to insert missing values into income components in order to be able to evaluate imputation and outlier detection methods. As discussed in the Trier meeting, missing values will only be treated in a multivariate setting, setting missing values in the equivalized household income directly is not considered.

To simplify the multivariate setting for the simulations, the derived rules are based on aggregated income components. More precisely, 4 income components are considered:

- Personal income from employment (employed and self-employed)
- Personal income from transfers
- Household income from capital
- Household income from employment (of children or not attributed) and transfers

In the case of the AAT-SILC population, which has been generated with the data simulation framework described in Alfons et al. (2010a) and implemented in the `R` package `simPopulation` (Alfons and Kraft, 2010), 16 income components are originally available. These income components are listed in Table 5.1. More information on the income components in EU-SILC can be found in Eurostat (2004). For most components, it is clear to which of the aggregated components mentioned above they should be assigned. Nevertheless, it is not completely clear in the following case:

Table 5.1: Variables selected for the simulation of the Austrian EU-SILC population data.

Name	Variable
py010n	Employee cash or near cash income
py050n	Cash benefits or losses from self-employment
py090n	Unemployment benefits
py100n	Old-age benefits
py110n	Survivor's benefits
py120n	Sickness benefits
py130n	Disability benefits
py140n	Education-related allowances
hy040n	Income from rental of a property or land
hy050n	Family/children related allowances
hy070n	Housing allowances
hy080n	Regular inter-household cash transfer received
hy090n	Interest, dividends, profit from capital investments in unincorporated business
hy110n	Income received by people aged under 16
hy130n	Regular inter-household cash transfer paid
hy145n	Repayment/receipts for tax adjustment

- `hy145n` (repayment/receipts for tax adjustment): Taxes are collected for both income and capital, therefore it is in principle not clear to which aggregated component it should be assigned. However, we propose to include it in the income from employment and transfers.

In short, this is our proposal for computing the aggregated income components from the components available in AAT-SILC:

- Personal income from employment: `pye <- py010n + py050n`
- Personal income from transfers:
  `pyt <- py090n + py100n + py110n + py120n + py130n + py140n`
- Household income from capital: `hyc <- hy040n + hy090n`
- Household income from employment and transfers:
  `hyet <- hy050n + hy070n + hy080n + hy110n - hy130n - hy145n`

The aggregated components `pye`, `pyt`, `hye`, `hyc` and `hyt` have been added to the AAT-SILC population data available in the AMELI `svn` repository (http://svn.uni-trier.de/AMELI/WORK_PACKAGES/WP6/general/aatsilc.zip).

The rest of the paper is organized as follows. Chapter 5.2 discusses different missing data mechanisms for the simulation study. The missing value rates to be used in the simulation study are then presented in Chapter 5.3. However, to obtain the missing value rates we

© http://ameli.surveystatistics.net/ - 2011

investigated different EU-SILC data sets from Austria: the public use data from 2004 (a subsample released by Statistics Austria for research purposes), and the data as reported to Eurostat from 2004, 2005 and 2006. All these data sets contain flags that indicate missing values for each income component (more precisely, fully or partially imputed values are flagged). While the public use data set indicates rather low proportions of missingness, the other data sets indicate very high proportions of imputed values. We discussed this with subject matter specialists, who believe that the high proportions of imputed values may be artefacts resulting from technical issues. A thorough analysis of the missing value rates is thus difficult and we propose to investigate three different scenarios.

5.1.1 Missing data mechanisms

In the missing data literature, three important cases of processes generating missing values are commonly distinguished, based on ideas by Rubin (1976). For a more detailed discussion on these missing data mechanisms, the reader is referred to Little and Rubin (2002).

Let $\boldsymbol{X} = (x_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ denote the data, where n is the number of observations and p the number of variables, and let $\boldsymbol{M} = (M_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ be an indicator whether an observation is missing ($M_{ij} = 1$) or not ($M_{ij} = 0$). Furthermore, let the observed and missing parts of the data be denoted by $\boldsymbol{X}_{obs}$ and $\boldsymbol{X}_{miss}$, respectively. The missing data mechanism is then characterized by the conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{X}$, denoted by $f(\boldsymbol{M}|\boldsymbol{X}, \phi)$, where ϕ denotes unknown parameters (see Little and Rubin, 2002).

The missing values are *missing completely at random* (MCAR) if the missingness does not depend on the data $\boldsymbol{X}$, i.e., if

$$f(\boldsymbol{M}|\boldsymbol{X}, \phi) = f(\boldsymbol{M}|\phi). \tag{5.1}$$

Note that there may still be a certain pattern in the missing values, depending on the unknown parameters ϕ, but such a pattern will be independent of the actual data. A more general scenario is given if the missingness depends on the observed information $\boldsymbol{X}_{obs}$. In this case, the missing values are *missing at random* (MAR), which translates to the equation

$$f(\boldsymbol{M}|\boldsymbol{X}, \phi) = f(\boldsymbol{M}|\boldsymbol{X}_{obs}, \phi). \tag{5.2}$$

On the other hand, the missing values are said to be *missing not at random* (MNAR) if Equation (5.2) is violated. This can be written as

$$f(\boldsymbol{M}|\boldsymbol{X}, \phi) = f(\boldsymbol{M}|(\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}), \phi). \tag{5.3}$$

Hence, in the latter case, the missing values cannot be fully explained by the observed part of the data.

5.2 Missing Values in Aggregated Income Components

Since the evaluation of imputation methods or outlier dectection for incomplete data is not the major task in the AMELI project, we propose to consider only two missing data mechanisms: MCAR and MAR. In general, missing values are generated on the personal level. For the household income components, however, it does not make sense to set the value of only some persons in a household to `NA`. Consequently, missing values should then be set in a second step for all persons in a household that contain `NA`s in the respective household income component. For this second step, the function `collectNA()` has been implemented. The `R` code of `collectNA()` along with a short description of its arguments is given in(Hulliger et al., 2011, Chapter G).

5.2.1 Setup for MCAR

The code for generating a control object for the insertion of missing values on the personal level that corresponds to an MCAR situation in simulations using `simFrame` is given in Listing 5.1. Please note that the choice for the missing value rates is discussed in Chapter 5.3.

```
NArate <- matrix(c(0.1, 0.1, 0.3, 0.025, 0.1, 0.3,
    0.15, 0.1, 0.3, 0.025, 0.1, 0.3), 3, 4)
nc <- NAControl(target = c("pye", "pyt", "hyc", "hyet"),
    NArate = NArate)
```

Listing 5.1: Code listing for an MCAR situation.

5.2.2 Setup for MAR

A graphical exploration of the public use EU-SILC data from 2004 using the `R` package `VIM` showed a strong dependecy of the missing values on the variables `r007000` (main activity status) and `age`.

Figure 5.1 contains spine plots of the variable `r007000` (main activity status) with missing values in `pye` (personal income from employment; *top left*), `pyt` (personal income from transfers; *top right*), `hyc` (household income from capital; *bottom left*) and `hyet` (household income from employment and transfers; *bottom right*) highlighted in red. In addition, Figure 5.2 shows parallel boxplots of the variable `age` grouped according to observed and missing data in each of the aggregated income components.

We thus propose to construct a MAR situation as described in the following. For each income component, a vector of probability weights for each individual is computed. The selection of persons whose values are set to `NA` is then based on those probability weights. It should in particular be noted that we do not consider a stratified design for selecting the individuals, as proposed by Hulliger and Schoch (2010) for the selection of observations to be contaminated, due to the more complex dependencies on the two auxiliary variables.

© http://ameli.surveystatistics.net/ - 2011

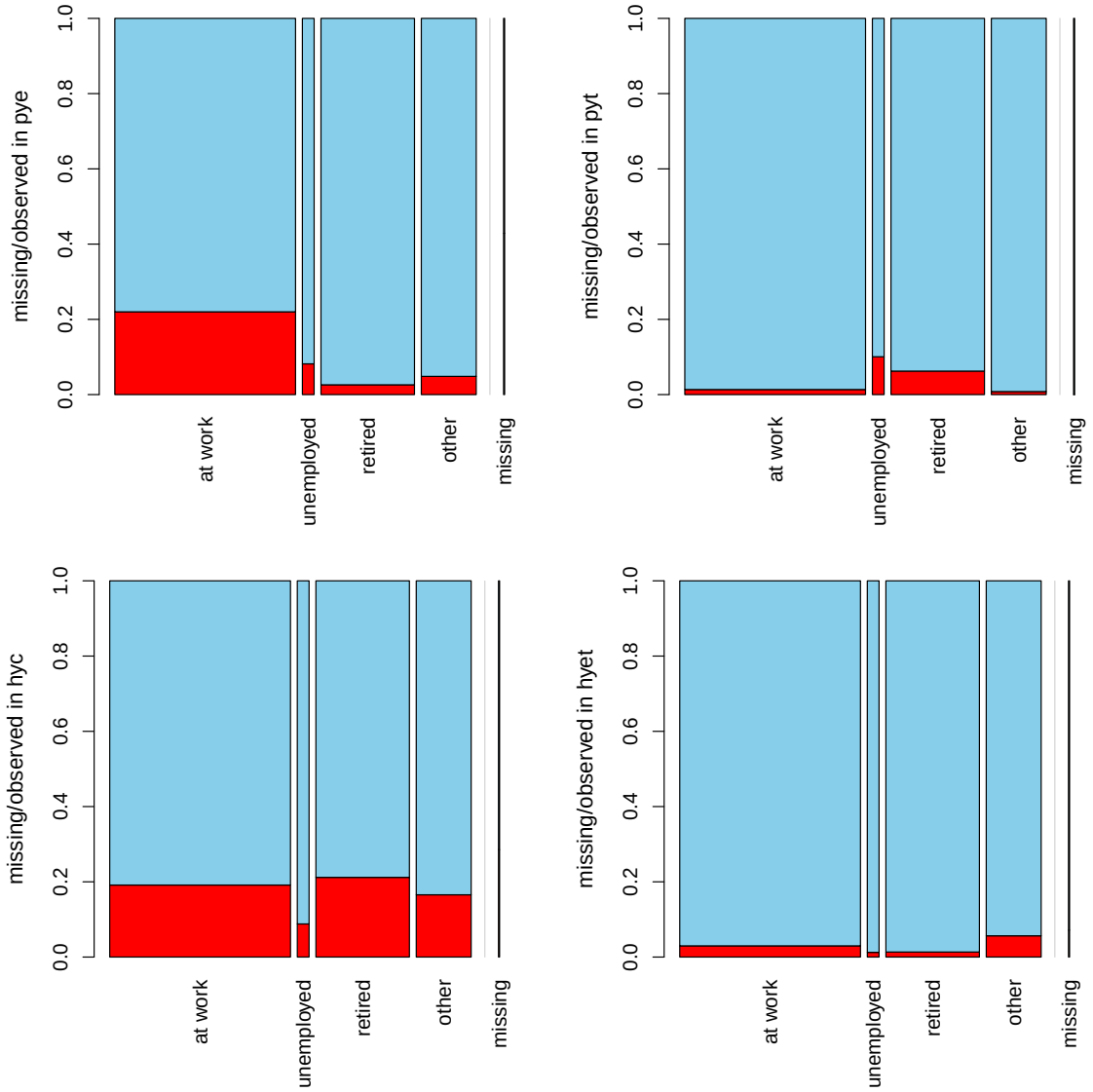


Figure 5.1: Spine plots of the variable `r007000` (main activity status) in the Austrian public use EU-SILC data from 2004, with missing values in `pye` (personal income from employment; *top left*), `pyt` (personal income from transfers; *top right*), `hyc` (household income from capital; *bottom left*) and `hyet` (household income from employment and transfers; *bottom right*) highlighted in red.

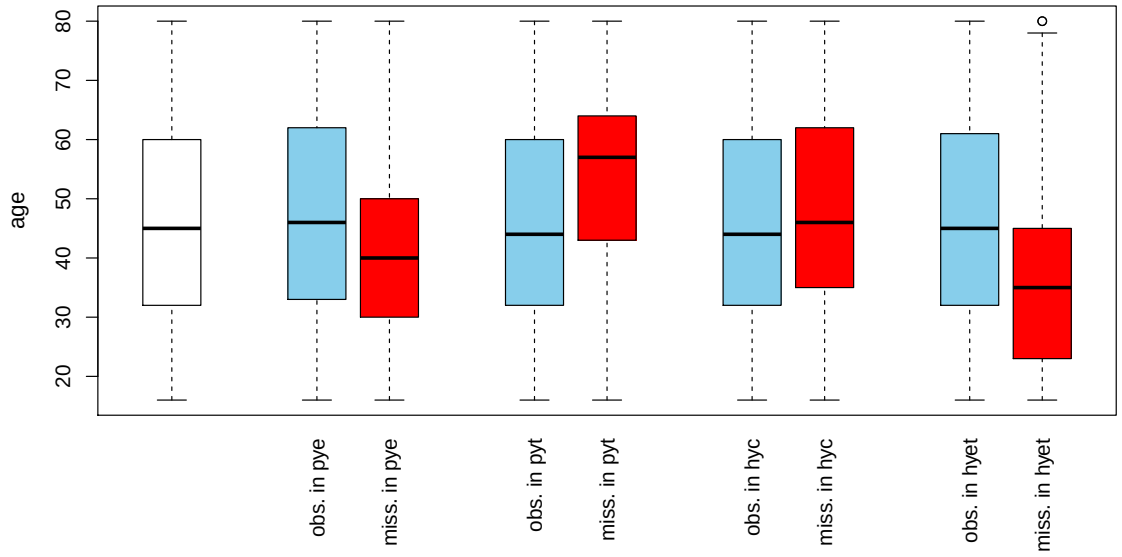


Figure 5.2: Parallel boxplots of the variable `age` in the Austrian public use EU-SILC data from 2004, grouped according to observed and missing data in each of the aggregated income components.

In any case, the proposed vectors of probability weights are constructed in the following manner:

- For each category of main activity status, the percentages of missing values in the four aggregated income components (with respect to the total number of missing values in the corresponding component) have been determined for the public use data set from 2004 and the data sets reported to Eurostat. Note that for the household income components, these percentages correspond to the percentages of households rather than individuals, since `NA`s always occur for all household members. From this analysis, we propose the probability weights listed in Table 5.2 for observations of the four possible outcomes of main activity status (`rb170` in AAT-SILC, `RB210` in AMELIA).
- Age is divided into five categories: $[-1, 16)$, $[16, 25)$, $[25, 50)$, $[50, 65)$ and $[65, 96]$ (cf. Eurostat, 2004). Based on a similar analysis of the percentages of missing values in the four aggregated income components, we propose the probability weights listed in Table 5.3 for observations in the five age categories.
- Final probability weights should then be derived by combining the two vectors of weights for each of the four components. Economic status is considered to be more important, therefore a coefficient 2/3 is used, while 1/3 is used for the probability weights corresponding to age category.

For the AAT-SILC data, the resulting probability weights for the four components are available as variables `pMARpye`, `pMARpyt`, `pMARhyc` and `pMARhyet`. With `simFrame`, a

© http://ameli.surveystatistics.net/ - 2011

Table 5.2: Proposed probability weights by income component for observations of the four possible outcomes of main activity status (`rb170` in AAT-SILC, `RB210` in AMELIA).

		Probability weight			
Category	Description	`pye`	`pyt`	`hyc`	`hyet`
1	At work	0.85	0.20	0.45	0.45
2	Unemployed	0.05	0.10	0.05	0.05
3	Retired	0.05	0.65	0.20	0.10
4	Other	0.05	0.05	0.15	0.15
	`NA`	0	0	0.15	0.25

Table 5.3: Proposed probability weights by income component for observations in the five age categories.

	Probability weight			
Age category	`pye`	`pyt`	`hyc`	`hyet`
$[-1, 16)$	0	0	0.20	0.25
$[16, 25)$	0.15	0.10	0.10	0.15
$[25, 50)$	0.60	0.15	0.35	0.40
$[50, 65)$	0.20	0.40	0.20	0.15
$[65, 96]$	0.05	0.35	0.15	0.05

control class for missing values can then be defined as follows:[1]

```
NArate <- matrix(c(0.1, 0.1, 0.3, 0.025, 0.1, 0.3,
    0.15, 0.1, 0.3, 0.025, 0.1, 0.3), 3, 4)
nc <- NAControl(target = c("pye", "pyt", "hyc", "hyet"),
    NArate=NArate,
    aux = c("pMARpye", "pMARpyt", "pMARhyc", "pMARhyet"))
```

Listing 5.2: Code listing for MAR situation

This MAR scenario is based on simplified rules, but those are determined by an extensive data analysis of different EU-SILC samples from Austria.

5.3 Missing value rates

As mentioned in Chapter 1, the flag variables for missing values in the public use data and in the data submitted to Eurostat draw a very different picture of the missing value rates. In particular the flags from the data submitted to Eurostat are rather cryptic. We discussed this issue with the subject matter specialists from Statistics Austria in order to determine suitable missing value rates for the simulation study. The subject matter

[1] The functionality to specify an auxiliary variable with probability weights for each target variable was added in `simFrame` version 0.3.

specialists believe that the high proportions of imputed values in the data sets reported to Eurostat may be artefacts resulting from technical issues.

We propose to investigate three scenarios: one with realistic proportions of missing values (as in the public use data), one with rather low proportions of missing values (equal proportions in all components), and one with rather high proportions of missing values (equal proportions in all components). The proposed missing value rates and their priorities for the simulation study are listed in Table 5.4.

Table 5.4: Missing value rates for the insertion of missing values into the aggregated income components.

	Missing value rate				
Scenario	`pye`	`pyt`	`hyc`	`hyet`	Priority
1	10%	2.5%	15%	2.5%	1
2	10%	10%	10%	10%	2
3	30%	30%	30%	30%	2

Bibliography

Alfons, A. (**2010**): `simFrame`: Simulation Framework. `R` package version 0.3.7.
URL http://CRAN.R-project.org/package=simFrame

Alfons, A. and **Kraft, S.** (**2010**): `simPopulation`: Simulation of synthetic populations for surveys based on sample data. `R` package version 0.2.1.
URL http://CRAN.R-project.org/package=simPopulation

Alfons, A., **Kraft, S.**, **Templ, M.** and **Filzmoser, P.** (**2010**a): *Simulation of synthetic population data for household surveys with application to EU-SILC*. Research Report CS-2010-1, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf

Alfons, A., **Templ, M.** and **Filzmoser, P.** (**2010**b): *An Object-Oriented Framework for Statistical Simulation: The `R` Package `simFrame`*. Journal of Statistical Software, 37 (3), pp. 1–36, to appear.

Eurostat (**2004**): *Description of Target Variables: Cross-sectional and longitudinal*. EU-SILC 065/04, Eurostat, Luxembourg.

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

© http://ameli.surveystatistics.net/ - 2011

Hulliger, B. and **Schoch, T.** (**2010**): *Report on Contamination Settings*. AMELI internal report, University of Applied Sciences Northwestern Switzerland.

Little, R. and **Rubin, D.** (**2002**): Statistical Analysis with Missing Data. New York: John Wiley & Sons, 2nd ed., ISBN 0-471-18386-5.

Rubin, D. (**1976**): *Inference and missing data*. Biometrika, 63 (3), pp. 581–592.

Templ, M., **Alfons, A.** and **Filzmoser, P.** (**2009**): *An application of VIM, the R package for visualization of missing values, to EU-SILC data*. Research Report CS-2009-2, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2009-2complete.pdf

Templ, M., **Alfons, A.** and **Kowarik, A.** (**2010**): VIM: Visualization and Imputation of Missing Values. R package version 1.4.2.
URL http://CRAN.R-project.org/package=VIM

R Development Core Team (**2010**): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Chapter 6

Simulation Criteria

6.1 Introduction

The criteria that will be used to compare settings and methods are listed in this chapter. It is based on the criteria in Milestone 4.2 and in Deliverable D6.1 and additional criteria developed during the analysis. Not all criteria will be calculated in every simulation. However, if a criteria is used the notation and definition in this chapter is applied.

6.2 Criteria

Let θ^* be the parameter to estimate, usually a Laeken indicator, evaluated at the true, complete population where no additional contamination has been added, y^*. Let $\hat{\theta}_k$ denote the kth estimate of θ^*, for $k = 1, \ldots, r$ replications. Similarly, let $\hat{V}(\hat{\theta}_k)$ denote the variance estimate, $k = 1 \ldots, r$. The variance to estimate is determined by the simulation variance of the point estimator. This is a deviation from D6.1 where the variance of the point estimator at the true data is taken as the estimand.

6.2.1 Univariate Criteria

6.2.2 Point estimator

1. average of the point-estimates,

$$avgT := (1/r) \sum_{k=1}^{r} \hat{\theta}_k \tag{6.1}$$

2. variance of the point-estimates,

$$varT := 1/(r-1) \sum_{k=1}^{r} (\hat{\theta}_k - avgT)^2 \tag{6.2}$$

© http://ameli.surveystatistics.net/ - 2011

3. bias of the point estimates,

$$biasT := (1/r)\sum_{k=1}^{r}(\hat{\theta}_k - \theta^*) \tag{6.3}$$

or relative bias

$$relbiasT\% := 100 \cdot biasT/\theta^* \tag{6.4}$$

4. median point-estimate,

$$medT := \text{med}\{\hat{\theta}_k; k = 1, \ldots, r\} \tag{6.5}$$

5. median absolute deviation about the median point-estimate (Rousseeuw and Croux, 1993, p.1273),

$$madT := b \cdot \text{med}_k\{|\hat{\theta}_k - medT\}|\}, \quad \text{where } j, k = 1, \ldots, r;\ b = 1.4826 \tag{6.6}$$

6. median error of point estimates, (cf. Richardson and Welsh, 1995, p.1436),

$$medeT := \text{med}_k(\hat{\theta}_k - \theta^*), \quad k = 1, \ldots, r \tag{6.7}$$

7. root mean square error of the point estimates,

$$rmseT := \sqrt{(1/r)\sum_{k=1}^{r}(\hat{\theta}_k - \theta^*)^2} \tag{6.8}$$

or relative root mean square error of the point-estimates,

$$relrmseT\% := 100 \cdot rmseT/\theta^* \tag{6.9}$$

8. median absolute error of the point estimates (analog to MSE; Richardson and Welsh (1995, p.1436)),

$$medaeT := 1.4826 \cdot \text{med}_k|\hat{\theta}_k - \theta^*| \tag{6.10}$$

9. maximum absolute relative error, (This measure may be useful to assess the sensitivity of an estimator to the presence of influential units in the sample; Beaumont and Alavi (2004, p.12)),

$$relmaxeT := \max_k |(\hat{\theta}_k - \theta^*)/\theta^*| \tag{6.11}$$

6.2.3 Variance estimator

We write $\hat{V}_k$ as a shorthand for $\hat{V}(\hat{\theta}_k)$.

1. average of the variance-estimates

$$avgV := (1/r)\sum_{k=1}^{r}\hat{V}_k \tag{6.12}$$

2. variance of the variance-estimates

$$varV := (1/r)\sum_{k=1}^{r}(\hat{V}_k - avgV)^2 \tag{6.13}$$

3. bias of the variance estimates:

$$biasV := (1/r)\sum_{k=1}^{r}(\hat{V}_k - varT) \tag{6.14}$$

or relative bias of the variance estimates:

$$relbiasV\% := 100 \cdot biasV/varT \tag{6.15}$$

4. median of the variance-estimates,

$$medV := \mathrm{med}_k\hat{V}_k \tag{6.16}$$

5. median absolute deviation about the median variance-estimate (Rousseeuw and Croux, 1993, p.1273),

$$madV := 1.4826 \cdot \mathrm{med}_k|\hat{V}_k - medV| \tag{6.17}$$

6. median error of the variance estimates:

$$medeV := \mathrm{med}(\hat{V}_k - varT) \tag{6.18}$$

7. root mean squared error of the variance estimates

$$rmseV := \sqrt{(1/r)\sum_{k=1}^{r}(\hat{V}_k - varT)^2} \tag{6.19}$$

or relative root mean squared error of the variance estimates

$$relrmseV\% := 100 \cdot rmseV/varT \tag{6.20}$$

8. median absolute error of the variance estimates

$$medaeV := \mathrm{med}_k|\hat{V}_k - varT| \tag{6.21}$$

9. maximum absolute relative error of the variance estimates

$$relmaxeV := \max_k|\hat{V}_k - varT|/varT \tag{6.22}$$

© http://ameli.surveystatistics.net/ - 2011

6.3 Confidence interval

Confidence level is fixed at 0.95.

Confidence intervals may be defined explicitely for certain procedures by upper and lower limit: L_k and U_k. Or the limits are derived by a normal approxiamtion as $L_k = \hat{\theta}_k - 1.96\sqrt{\hat{V}_k}$ and accordingly U_k.

1. average half length of confidence interval I:

$$avglCI = (1/r)\sum_{k=1}^{r}(U_k - L_k) \tag{6.23}$$

2. Coverage probability:

$$covprobCI = (1/r)\sum_{k=1}^{r} 1\{L_k \leq \hat{\theta}_k \leq U_k\} \tag{6.24}$$

6.4 Outlier Criteria

Detection: False negatives and false positives.

Outlier detection is assumed to result in an indicator u_i which is 0 when observation i is an outlier and 1 if not.

Assume the outlier mechanism creates an outlier indicator vector o_i where $o_i = 1$ if observation i is an outlier and $o_i = 0$ if not. The sampling weight is w_i.

1. average proportion of false negatives (undetected outliers):

$$avepfn = (1/r)\sum_{k=1}^{r}\frac{\sum_{i\in S_k} w_i o_i u_i}{\sum_{i\in S_k} w_i o_i} \tag{6.25}$$

2. average proportion of false positives (nominated non-outliers):

$$avepfp = (1/r)\sum_{k=1}^{r}\frac{\sum_{i\in S_k} w_i (1-o_i)(1-u_i)}{\sum_{i\in S_k} w_i (1-o_i)} \tag{6.26}$$

6.5 Multivariate Criteria

Outliers are detected in p-dimensional raw data y_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The true complete data is denoted y^*_{ij}. The true population mean of variable j is $\mu^*_j = \sum_{i\in U} y^*_{ij}/N$. The true population standard deviation of variable j is $\sigma^{*2}_j = \sum_{i\in U}(y^*_{ij} - \mu^*_j)^2/(N-1)$.

Assume that together with an imputed value $\hat{y}_{ij}$ a flag is created with $v_{ij} = 1$ if y_{ij} has been imputed and $v_{ij} = 0$ otherwise. Let $v_i = 1 - \prod_{j=1}^{p}(1 - v_{ij})$ indicate imputed observations.

1. average of the absolute relative error of detection (similar to (EUREDIT, 2003, Measure 13) and corresponding to the continuous case of the distance measure in Section 3.5.4 of D6.1)

$$avgare = (1/r)\sum_{k=1}^{r}(1/p)\sum_{j=1}^{p}(1/\sigma_j^*)\left|\frac{\sum_{i\in S_k} w_i u_i y_{ij}/\sum_{i\in S_k} w_i u_i}{\mu_j^*} - 1\right| \tag{6.27}$$

   Note: Only applicable for $y_{ij} > 0$ and $y_{ij}^* > 0$.

2. average mean squared error of mean of imputed data

$$\begin{aligned} avgmse \;=\; & (1/r)\sum_{k=1}^{r}\left[\frac{\sum_{i\in S_k} w_i(\hat{y}_i - m_{S_k}(\hat{y}))^\top C^{*-1}(\hat{y}_i - m_{S_k}(\hat{y}))}{\sum_{i\in S} w_i \sum_{i\in S} w_i(1-v_i)}\right] \\ & +(1/r)\sum_{k=1}^{r}\left[(m_{S_k}(\hat{y}) - \theta^*)^\top C^{*-1}(m_{S_k}(\hat{y}) - \theta^*)\right], \end{aligned} \tag{6.28}$$

   where $y_i = (y_{i1}, \ldots, y_{ip})^\top$ and $m_S(\hat{y}) = \sum_{i\in S} w_i\hat{y}_i/\sum_{i\in S} w_i$ and θ^* is the vector of true population means and where C^* is the covariance of the true data y^* in the population.

3. average difference in variation between true and imputed data:

$$avgdv := (1/r)\sum_{k=1}^{r}\frac{2}{p(p-1)}\sum_{h=1}^{p-1}\sum_{j=h+1}^{p}|\hat{C}_{hj,k} - C_{hj}^*|, \tag{6.29}$$

   where $\hat{C}_{hj,k}$ is element h, j of the covariance matrix of $\hat{y}$ of S_k.

4. average Mahalanobis distance between imputed and true data

$$avgmd := (1/r)\sum_{k=1}^{r}\frac{\sum_{i\in S_k} w_i v_i(\hat{y}_i - y_i^*)^\top C^{*-1}(\hat{y}_i - y_i^*)}{\sum_{i\in S_k} w_i v_i} \tag{6.30}$$

The criteria 1, 2 and 3 are formulated such that it is not necessary to maintain the true data of the replicate samples at the same time as the resulting data $\hat{y}$. Criterion avgmd needs the true and the resulting (imputed) data for the sample.

Bibliography

Beaumont, J.-F. and **Alavi, A.** (**2004**): *Robust Generalized Regression Estimation.* Survey Methodology, 30 (2), pp. 195–208.

EUREDIT (**2003**): *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project (Volume 1).* Technical report, http://www.cs.york.ac.uk/euredit/.

© http://ameli.surveystatistics.net/ - 2011

Richardson, A. M. and **Welsh, A. H.** (**1995**): *Robust Restricted Maximum Likelihood in Mixed Linear Models*. Biometrics, 51 (4), pp. 1429–1439.

Rousseeuw, P. J. and **Croux, C.** (**1993**): *Alternatives to the Median Absolute Deviation*. Journal of the American Statistical Association, 88 (424), pp. 1273–1283.

Part II

Simulation Reports

© http://ameli.surveystatistics.net/ - 2011

Chapter 7

WP2: Parametric Estimation

7.1 Parametric Estimation of Income Distributions and Derived Indicators Using the GB2 Distribution

We present methodologies for the parametric estimation of a certain set of indicators of poverty and social exclusion computed within the EU-SILC survey, and in particular the median, the at-risk-of-poverty rate (ARPR), the relative median poverty gap (RMPG), the quintile share ratio (QSR) and the Gini index.

We are interested in fitting the GB2 distribution on the variable ëquivalized incomeüsing different methods of estimation, estimating the variance of the fitted parameters and of the fitted indicators, which are expressed as functions of the parameters of the distribution, whenever this is possible.

The methods of parametric estimation of the GB2 distribution, developed in the AMELI project, are described in Deliverable D2.1 (Graf et al., 2011).

7.1.1 Simulation setup

The simulations use the AMELIA universe. Simulations are run using the sampling designs described in Section 2 and are processed at the global level as well as for the 4 AMELIA regions.

Data source: AMELIA income data set

Study variable: Equivalized disposable income

Sample Size: 6,000 households

Number of samples: 1,000 per design

Indicators: median, ARPR, RMPG, QSR, Gini index, GB2 parameters

Estimation techniques: Estimation of indicators directly from the sample (*empirical estimation*), pseudo maximum likelihood estimation and method of non-linear fit for indicators (see Graf et al., 2011, chap. 2), left tail decomposition (see Graf et al., 2011, chap. 5)

Estimation level: Households and Persons

There are two general schemes.

- Estimation of the GB2 parameters and derived indicators.
- Variance estimation of the fitted parameters and indicators.

7.1.2 Simulation objectives

Our simulations have for objectives to

1. study the efficiency and the bias of the indicators' estimators under the different methods of estimation on the global and regional level. Efficiency and bias are compared with the standard empirical estimators.
2. study the effect of the robustification of the sampling weights on the indicators.
3. study the effect of the sample design on the estimators.
4. study the quality of the variance estimators for the cases where variance estimators are possible.

For description of the methods of estimation, the procedure of robustification of the sampling weights and the variance estimation (see Graf et al., 2011, chap. 4 and 5)

7.1.3 Simulation bed

The first type of simulations we have performed are on the global level (samples from the whole AMELIA universe). For each sample, drawn by one of the five sampling designs described in Section 2, we have done the following:

1. Fit the GB2 distribution on the variable equivalized income on the personal level (only positive values) using the method of ML estimation based on the full pseudo log-likelihood. The original and robustified sampling weights are applied.
2. Calculate the empirical estimates and the GB2 estimates of the indicators.
3. Calculate variance estimates of the GB2 parameters and indicators through ML estimation, either using the sampling weights only, or using the full design information.

© http://ameli.surveystatistics.net/ - 2011

4. Estimate the GB2 parameters and the indicators using the method of non-linear fit for indicators.

The second type of simulations focus on the behaviour of the estimates on the regional level. Samples were drawn from AMELIA universe. For each sample, drawn by one of the five sampling designs described in Chapter 3, we have done the following:

1. Fit the GB2 distribution on the variable equivalized income on the personal level (only positive values) using the method of ML estimation based on the profile pseudo log-likelihood. The used sampling weights are robustified. Parameters are estimated on the global and on the regional level.

2. Calculate the empirical estimates and the GB2 estimates of the indicators on the global level and for each region.

3. Calculate variance estimates of the GB2 parameters and indicators obtained through ML estimation, either using the sampling weights only, or using the full design information. Again, variance estimates are calculated on the global level and on the regional level.

4. Estimate the GB2 parameters and the indicators using the method of non-linear fit for indicators on the global and on the regional level.

5. Only on the regional level, fit the left tail decomposition of the GB2 (see Graf et al., 2011, chap. 5)

A summary is presented in Table 7.1.

Table 7.1: Summary of the different simulations

	Method of estimation	Sampling weights		AMELIA level	
		Original	Robust	Global	Regional
1	ML full	+	+	+	-
2	ML prof	-	+	+	+
3	NLS	+	-	+	+
4	Compound	-	+	-	+

7.1.4 Analysis of the simulation results

The main results of the simulations with the GB2 distribution are tabulated in Appendix (Hulliger et al., 2011).

In this section we will comment on the different aspects of the results of our simulation study, i.e. quality of the GB2 model fitted on the AMELIA universe and comparison with the EU-SILC data, quality of the variance estimation, quality of the estimation of the indicators.

The GB2 model for simulated data

On Figure 7.1 we can see the cumulative distribution plot for a sample drawn using the sampling design 1.4a of the AMELIA universe (see Table 2.1 for description of the sampling designs). A comparison can be made with the GB2 fit on real data (see Graf et al., 2011, Figure 4.6.2). We can see that the GB2 is capable to adapt to various income distributions.

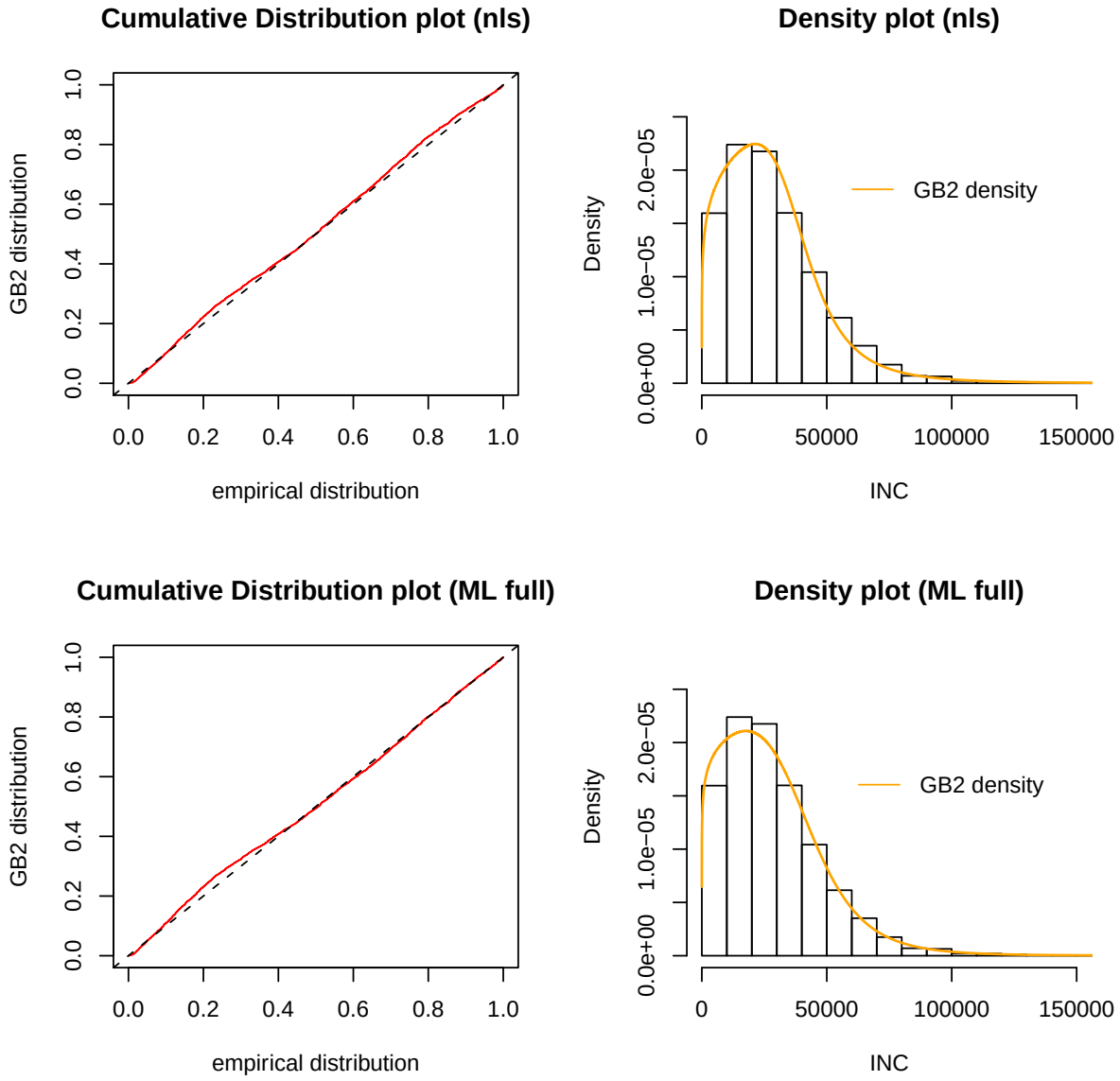


Figure 7.1: GB2 distribution and density plots, 1st sample of design d1.4a

Estimation of the GB2 parameters

On Figure 7.2 is presented a boxplot of the fitted GB2 parameters of the 26 participating countries in the EU-SILC survey, 2006. On Figure 7.3 is the corresponding plot for the simulated data set (AMELIA sampling design d1.4a, simulation type 1 of Table 7.1). On both figures we can see that the parameters obtained using the full or the profile log-likelihood are really close to each other. We can also see that the parameters p and q governing the left and right tale of the distribution, respectively, have much higher values for the EU-SILC survey.

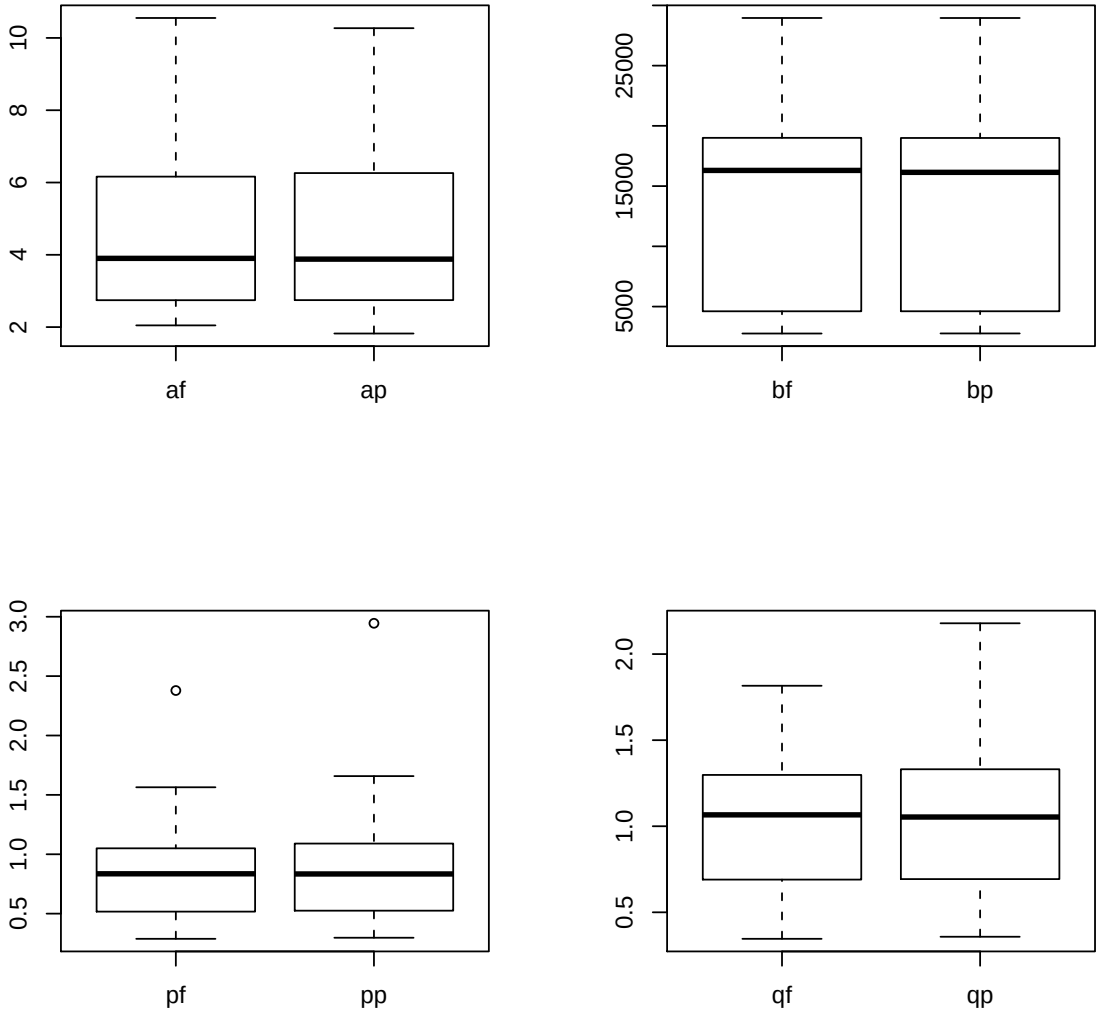


Figure 7.2: Fitted parameters af, bf, pf, qf (ML full log-likelihood on the left) and ap, bp, pp, qp (ML profile log-likelihhod on the right), robustified weights, EU-SILC countries 2006

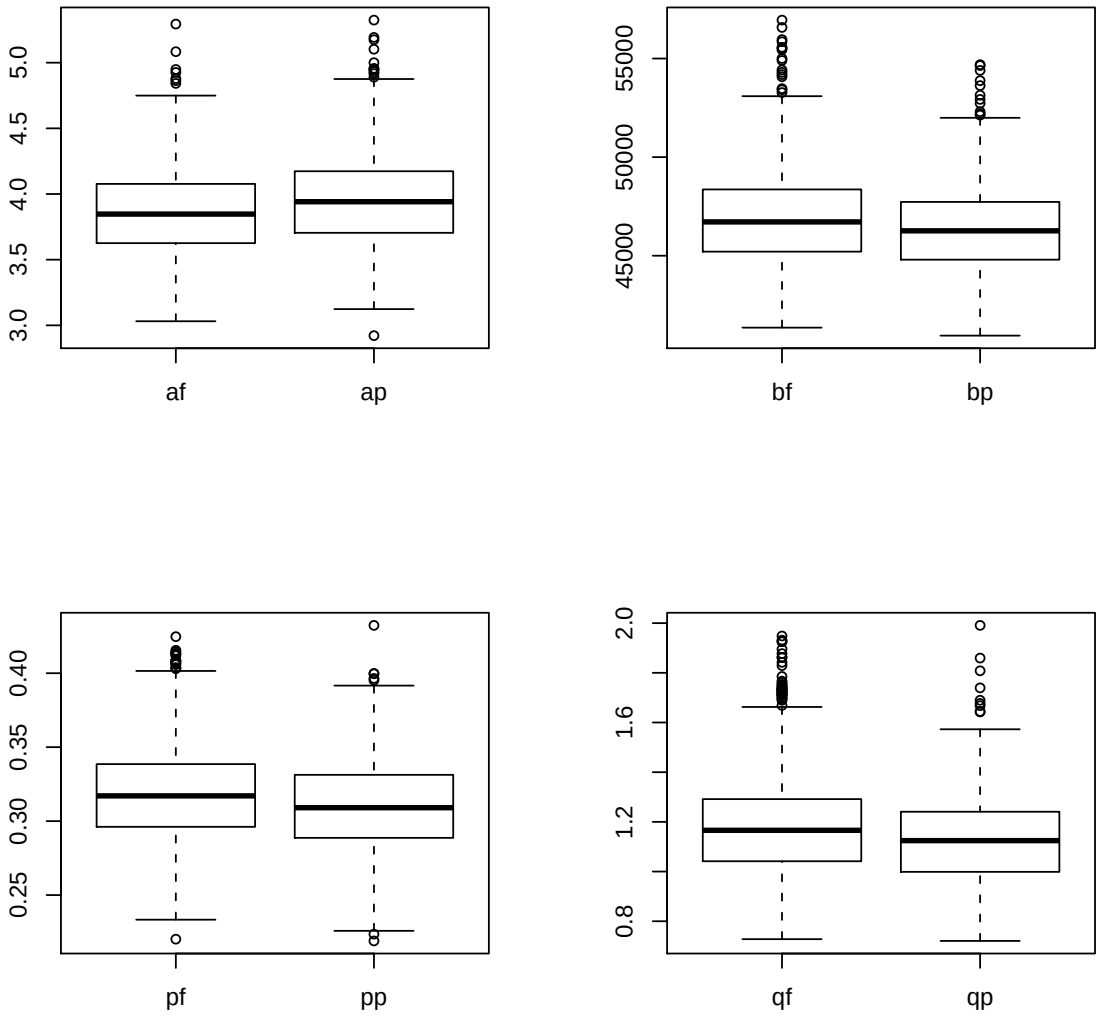


Figure 7.3: Fitted parameters af, bf, pf, qf (ML full log-likelihood on the left) and ap, bp, pp, qp (ML profile log-likelihood on the right), robustified weights, d1.4a, global level

© http://ameli.surveystatistics.net/ - 2011

Effect of the robustification on the GB2 parameters, AMELIA sample d1.4a

On Figure 7.4 we can see what the effect of the robustification of the sampling weights is on the fitted GB2 parameters with the method of maximum likelihood estimation using the full log-likelihood. The parameters calculated with the robustified (adjusted) weights are denoted by the suffix *adj*. Only converging samples have been taken for producing the boxplots.

We can note that the scale parameter b is almost not affected by the adjustment. The asymptotic behaviour of the GB2 when the income tends to 0 is similar to a Pareto distribution with parameter ap, and to a Pareto distribution with parameter aq, when the income tends to infinity. On the bottom of the figure we observe the two products of parameters ap and aq, respectively. Note that the notation ap used here is different from the notation used on Figures 7.2 and 7.3. These plots show that, in the case of the AMELIA data set, our adjustment affects essentially the left tail of the distribution.

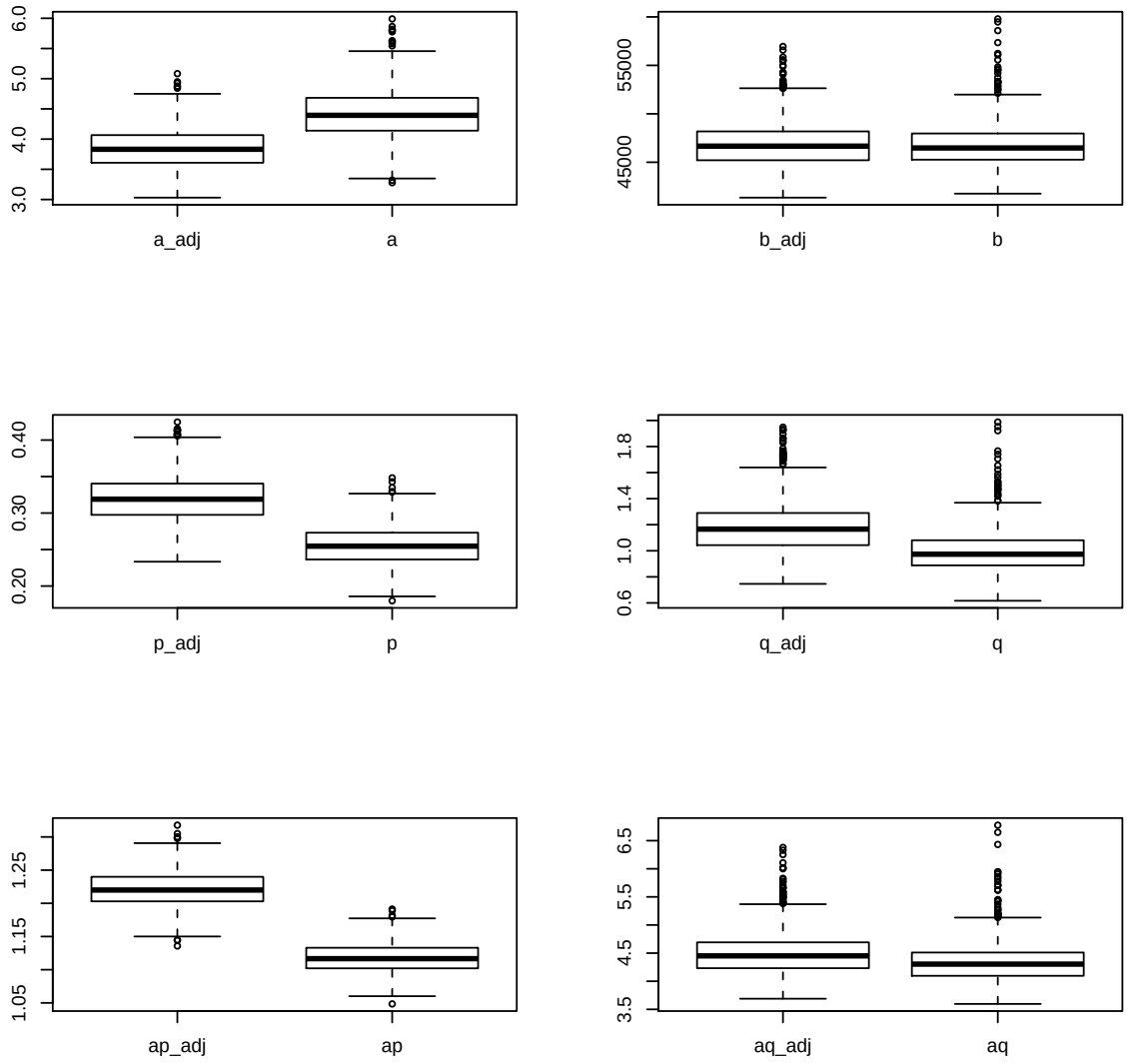


Figure 7.4: Fitted parameters, ML full, d1.4a, global level

© http://ameli.surveystatistics.net/ - 2011

Estimation of the derived indicators

Table 7.2 presents the simulation results for the GB2 estimate of the RMPG as well as for its empirical estimate, denoted eRMPG. The used notation is as follows:

T: point estimator of the fitted GB2 indicator,

adj: if the weights were robustified (yes) or not (no),

design: the used sampling design for the sample drawn from the AMELIA universe,

cvsT: coefficient of variation of the fitted parameter over the 1000 simulations,

cvT: mean over all simulations of the coefficient of variation, calculated using the variance estimator with sampling weights only,

cvdT: mean over all simulations of the coefficient of variation, calculated using the variance estimator with the full design information .

Remark: Details on the variance estimators can be found in Deliverable D2.1 (see Graf et al., 2011, sec. 4.4).

Effect of the robustification on the derived indicators, AMELIA sample d1.4a

Table 7.2 shows that, for all sampling designs, the GB2 estimate of RMPG has larger values than the empirical estimate. However, the adjustment of the sampling weights tends to reduce the value of the estimator. It is also interesting to notice that, in general, our estimates have smaller variance than the empirical estimate. Another advantage of the weight adjustment is that, in all cases, it reduces significantly the relative root mean squared error (RRMSE) of our estimates. For RMPG, we see that the RRMSE of the estimator using the robustified weights is smaller than the RRMSE of the empirical estimator. Corresponding tables for the median, ARPR, QSR and Gini can be found in the Appendix.

On Figure 7.5 the fitted indicators with robustified weights (adj) and original weights are compared with the empirical estimates (first in the plot) and the estimates obtained with the method of non-linear fit for indicators (the last in the plot). We can see that the adjustment of the sampling weights improves considerably the quality of estimation. It gives best results for the RMPG, and we can see that for QSR the adjustment was probably too strong. As described in our procedure (see Graf et al., 2011, sec. 4.3), we can modify the value of the constant for the correction factor and choose different quantiles. The plot shows also that, as expected, the indicators' estimates by the method of non-linear fit for indicators are close to the empirical estimates of the indicators.

Table 7.2: Point estimator of RMPG

T	adj	design	avgT	madT	varT	cvsT	cvT	cvdT	relbiasT%
RMPG	no	d1.2	4.667e+01	6.522e-01	4.027e-01	1.360e-02	1.368e-02	1.368e-02	7.016e+00
RMPG	yes	d1.2	4.409e+01	7.071e-01	4.409e-01	1.506e-02	1.293e-02	1.293e-02	9.757e-01
eRMPG	no	d1.2	4.367e+01	-	2.038e+00	3.269e-02	-	-	-
RMPG	no	d1.4a	4.664e+01	6.163e-01	3.711e-01	1.306e-02	1.377e-02	1.370e-02	6.848e+00
RMPG	yes	d1.4a	4.407e+01	6.439e-01	3.970e-01	1.430e-02	1.290e-02	1.283e-02	8.677e-01
eRMPG	no	d1.4a	4.369e+01	-	1.992e+00	3.230e-02	-	-	-
RMPG	no	d1.5a	4.667e+01	6.624e-01	4.723e-01	1.473e-02	1.466e-02	1.459e-02	6.837e+00
RMPG	yes	d1.5a	4.408e+01	6.973e-01	4.749e-01	1.563e-02	1.313e-02	1.305e-02	8.692e-01
eRMPG	no	d1.5a	4.369e+01	-	1.784e+00	3.057e-02	-	-	-
RMPG	no	d2.6	4.658e+01	8.605e-01	7.811e-01	1.897e-02	1.473e-02	0.000e+00	7.024e+00
RMPG	yes	d2.6	4.397e+01	8.746e-01	8.109e-01	2.048e-02	1.322e-02	0.000e+00	8.531e-01
eRMPG	no	d2.6	4.356e+01	-	2.041e+00	3.279e-02	-	-	-
RMPG	no	d2.7	4.657e+01	8.163e-01	7.001e-01	1.797e-02	1.375e-02	1.639e-02	6.885e+00
RMPG	yes	d2.7	4.392e+01	8.600e-01	8.038e-01	2.042e-02	1.295e-02	1.551e-02	8.430e-01
eRMPG	no	d2.7	4.354e+01	-	2.175e+00	3.387e-02	-	-	-

© http://ameli.surveystatistics.net/ - 2011

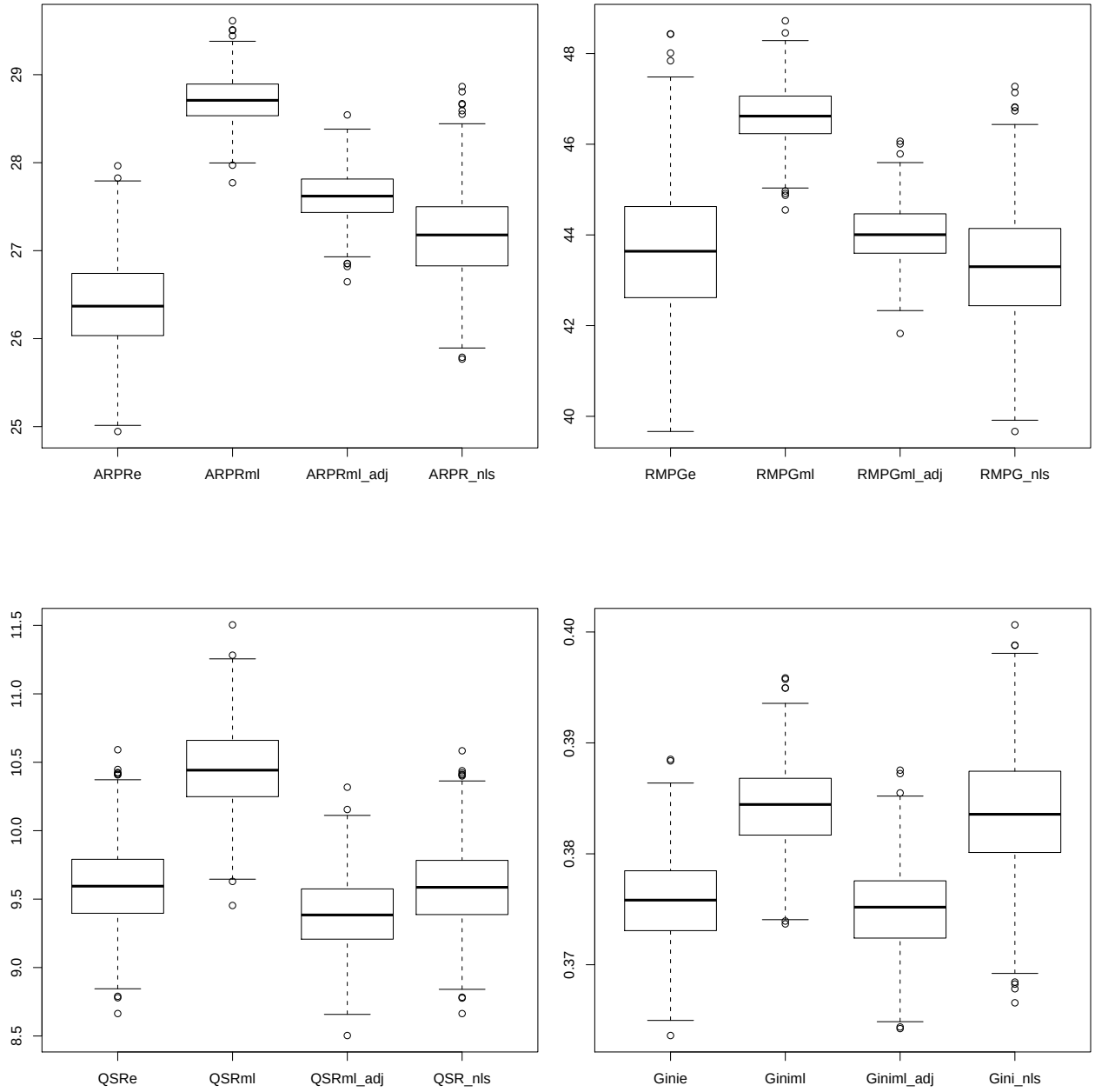


Figure 7.5: Fitted indicators, ML full, d1.4a, global level

Variance estimation of the GB2 parameters and derived indicators

The variance of the fitted GB2 parameters is calculated by linearization and using the so-called sandwich variance estimator (see Graf et al., 2011, sec. 4.4). The middle term of the sandwich variance estimator, i. e. the variance of the scores (the first derivatives of the pseudo log-likelihood), in our simulation study is calculated in two ways:

- In the first case, all the information we need is the income variable and the sampling weights.
- In the second case, inclusion probabilities, sample strata sizes can be considered when calculating the variance of the scores, through the design definition (using the R package survey, Lumley (2010)).

For example, for design d1.4a, we have the following definition:

```
dsktr = svydesign(id=~HID, strata=~NUTS2, fpc=NULL, weights=~aw, data=sktr),
```

where `sktr` denotes the sample of positive income values, `id` specifies the cluster ID (in our case the household id), `strata` defines the stratification variable and `weights` are the sampling weights. Then the variance of the scores is easily calculated on this design using the command `vcov`:

```
DV2 <- vcov(svytotal(~scores[,1]+scores[,2]+scores[,3]+scores[,4], design=dsktr))
```

We have implemented this second formula in our simulation study with success, except for the design d2.6, which is a two stage design with pps-sampling on the second stage and for which the *survey* formula was not adaptable. We have seen that our variance estimate using the sampling weights only ($def = 1$) is close or equal to the design variance calculated with the package *survey* for the one-stage sampling designs ($def = 2$) (see Table 7.3). For design d2.7 however, the bias in the variance is divided by a factor ranging from 2 to 100 (with def=2).

The variance of the derived indicators being calculated using the variance-covariance matrix of the scores and the first derivatives of the indicators with respect to the fitted parameters, the effect of the two different formulas can also be observed for the derived indicators (see e.g. Table 7.4).

Effect of the different variance formulae and of robustification of the sampling weights on the variance estimate of ARPR

From Table 7.4 we can learn that there is almost no difference between the two variance formulae ($def = 1$ and $def = 2$) for the one-stage designs. For all sampling designs, the use of the robustified weights reduces, but also underestimates, the variance of ARPR. In fact, we should take into account in the calculation of our variance estimate the additional variance due to the adjustment of the sampling weights.

© http://ameli.surveystatistics.net/ - 2011

Table 7.3: Variance estimator of a

V	def	adj	design	avgV	biasV	relbiasV%
a	1	no	d1.2	1.972e-01	1.744e-02	9.703e+00
a	2	no	d1.2	1.971e-01	1.740e-02	9.680e+00
a	1	yes	d1.2	2.039e-01	8.355e-02	6.943e+01
a	2	yes	d1.2	2.039e-01	8.356e-02	6.944e+01
a	1	no	d1.4a	3.326e-01	1.526e-01	8.480e+01
a	2	no	d1.4a	3.313e-01	1.514e-01	8.410e+01
a	1	yes	d1.4a	1.328e-01	1.168e-02	9.651e+00
a	2	yes	d1.4a	1.320e-01	1.096e-02	9.049e+00
a	1	no	d1.5a	1.528e-01	-1.017e-02	-6.241e+00
a	2	no	d1.5a	1.517e-01	-1.134e-02	-6.954e+00
a	1	yes	d1.5a	1.078e-01	-9.880e-03	-8.391e+00
a	2	yes	d1.5a	1.069e-01	-1.083e-02	-9.197e+00
a	1	no	d2.6	1.462e-01	-6.517e-02	-3.082e+01
a	1	yes	d2.6	1.345e-01	-1.933e-02	-1.257e+01
a	1	no	d2.7	1.707e-01	-7.311e-02	-2.998e+01
a	2	no	d2.7	2.172e-01	-2.661e-02	-1.091e+01
a	1	yes	d2.7	1.253e-01	-4.230e-02	-2.525e+01
a	2	yes	d2.7	1.671e-01	-4.200e-04	-2.500e-01

Table 7.4: Variance estimator of the ARPR

V	def	adj	design	avgV	biasV	relbiasV%
ARPR	1	no	d1.2	7.664e-02	1.170e-03	1.546e+00
ARPR	2	no	d1.2	7.665e-02	1.180e-03	1.562e+00
ARPR	1	yes	d1.2	6.540e-02	-2.003e-02	-2.344e+01
ARPR	2	yes	d1.2	6.541e-02	-2.002e-02	-2.343e+01
ARPR	1	no	d1.4a	7.859e-02	1.001e-02	1.459e+01
ARPR	2	no	d1.4a	7.760e-02	9.020e-03	1.315e+01
ARPR	1	yes	d1.4a	6.564e-02	-1.063e-02	-1.393e+01
ARPR	2	yes	d1.4a	6.460e-02	-1.167e-02	-1.530e+01
ARPR	1	no	d1.5a	8.620e-02	1.900e-04	2.188e-01
ARPR	2	no	d1.5a	8.506e-02	-9.500e-04	-1.108e+00
ARPR	1	yes	d1.5a	6.581e-02	-2.369e-02	-2.647e+01
ARPR	2	yes	d1.5a	6.465e-02	-2.485e-02	-2.776e+01
ARPR	1	no	d2.6	8.701e-02	-8.009e-02	-4.793e+01
ARPR	1	yes	d2.6	6.663e-02	-1.241e-01	-6.506e+01
ARPR	1	no	d2.7	7.744e-02	-7.368e-02	-4.876e+01
ARPR	2	no	d2.7	1.313e-01	-1.984e-02	-1.313e+01
ARPR	1	yes	d2.7	6.633e-02	-1.239e-01	-6.514e+01
ARPR	2	yes	d2.7	1.220e-01	-6.829e-02	-3.589e+01

AMELIA regions

This section gives an overview of the simulation results for region 1 of the AMELIA samples drawn with design d1.4a, simulation types 2, 3 and 4 of Table 7.1. For the simulations on the regional level we have used only the robustified weights.

On Figure 7.6 are shown the GB2 parameters for region 1 obtained through ML estimation, denotes respectively $amr1, bmr1, pmr1$ and $qmr1$ and NLS estimation. For the method of non-linear fit for indicators (NLS) we have used the 2-step estimation procedure, based on two different sets of initial values - ML fitted parameters on the one side and $a = 1/Gini, b = empiricalmedian, p = q = 1$ on the other side (see Section 4.5 of Deliverable 2.1). The obtained parameters are denoted respectively by, $anr1, bnr1, pnr1, qnr1$ and $an3r1, bn3r1, pn3r1, qn3r1$. We can see that all methods give similar results for the parameters' estimates, but the third method is more unstable. On Figure 7.7 we can compare the empirical indicators (denoted e.g. $eARPR$) with the indicators' estimators obtained through ML (e.g. $mlARPR$), NLS estimation two-step procedure with initial values from the ML fit (e.g. $nlsARPR$) and the method of decomposition of the GB2 (e.g. compARPR). We can see that all three methods of estimation give rather good results. The method of decomposition of the GB2 (GB2 compound) has not been developed for the estimation of the Gini, but, as seen on the plot, works well for the other indicators.

Non-convergence rate by regions

Table 7.5 shows, for each sampling design and for four methods of estimation, the number of samples per AMELIA region which have not converged. We can see that:

1. The method of ML estimation using the profile log-likelihood has a good convergence rate for all regions (there is only a small number of non-convergent samples for region 3);
2. the method of non-linear fit for indicators with initial values for the parameters coming from the ML fit gives the best convergence rate;
3. the compound fit with initial $p_l = (0.1, 0.7, 0.2)$ does not at all work for region 1, but the other partition seems to be a good compromise.

© http://ameli.surveystatistics.net/ - 2011

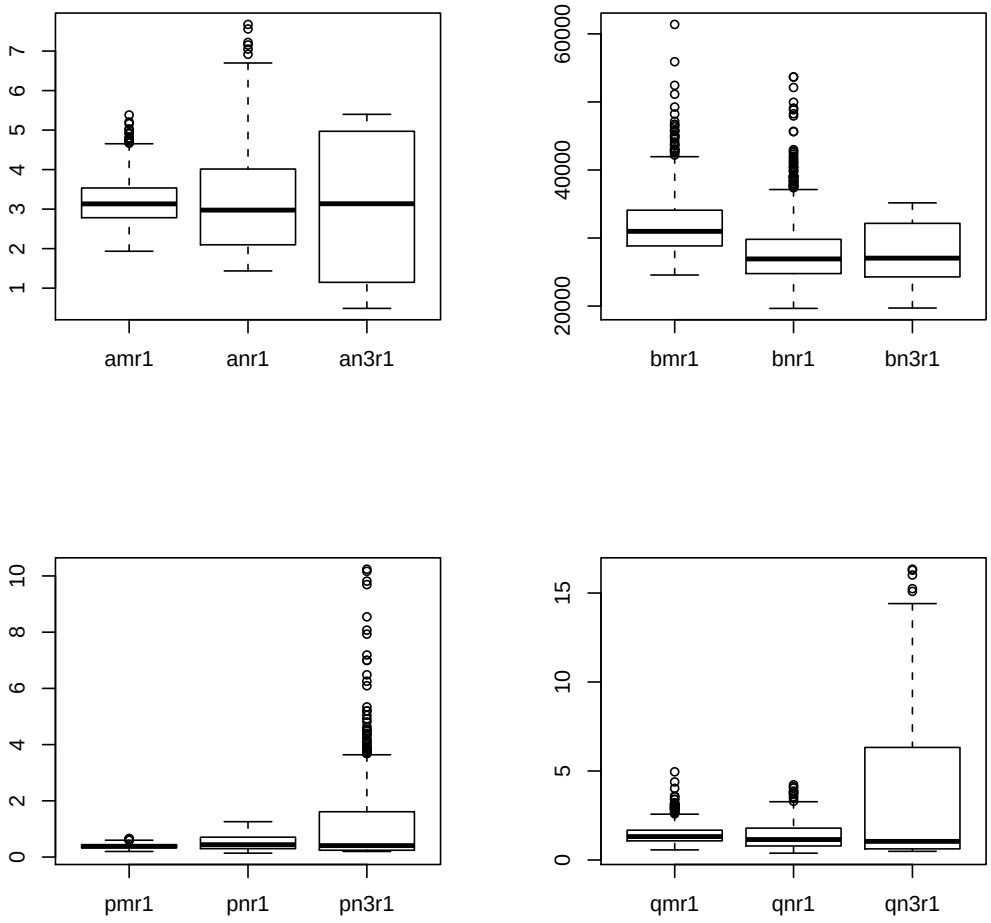


Figure 7.6: Fitted parameters, ML prof, d1.4a, region 1 (r1)

Table 7.5: Number of non-convergent samples by region over 1000 simulations

Method of estimation	design	reg. 1	reg. 2	reg. 3	reg. 4
ML fit profile log-likelihood	d1.2	0	0	34	0
	d1.4a	0	4	30	1
	d1.5a	0	0	27	1
	d2.6	0	0	25	0
	d2.7	0	1	36	0
2-step non-linear fit for indicators, initial values ML, fitted parameters a, p and q	d1.2	0	1	3	0
	d1.4a	0	0	1	0
	d1.5a	0	0	2	2
	d2.6	2	0	0	0
	d2.7	1	0	1	1
2-step non-linear fit for indicators, initial values ML, fitted parameter b	d1.2	0	0	0	0
	d1.4a	0	0	0	0
	d1.5a	0	0	0	0
	d2.6	0	0	0	0
	d2.7	0	0	0	0
2-step non-linear fit for indicators, initial values $a = 1/Gini$, $b = median$, fitted parameters a, p and q	d1.2	149	4	53	59
	d1.4a	169	4	43	59
	d1.5a	139	2	44	52
	d2.6	159	1	44	41
	d2.7	152	7	49	65
2-step non-linear fit for indicators, initial values $a = 1/Gini$, $b = median$, fitted parameter b	d1.2	0	0	0	0
	d1.4a	0	0	0	0
	d1.5a	0	0	0	0
	d2.6	0	0	0	0
	d2.7	0	0	0	0
compound fit, left tail decomposition, initial $p_l = (1/3, 1/3, 1/3)$	d1.2	0	75	0	0
	d1.4a	1	70	0	0
	d1.5a	0	92	0	0
	d2.6	10	45	0	0
	d2.7	10	34	0	0
compound fit, left tail decomposition, initial $p_l = (0.1, 0.7, 0.2)$	d1.2	999	401	0	0
	d1.4a	1000	399	0	0
	d1.5a	1000	389	0	0
	d2.6	999	356	0	0
	d2.7	999	388	0	0

© http://amelie.surveystatistics.net/ - 2011

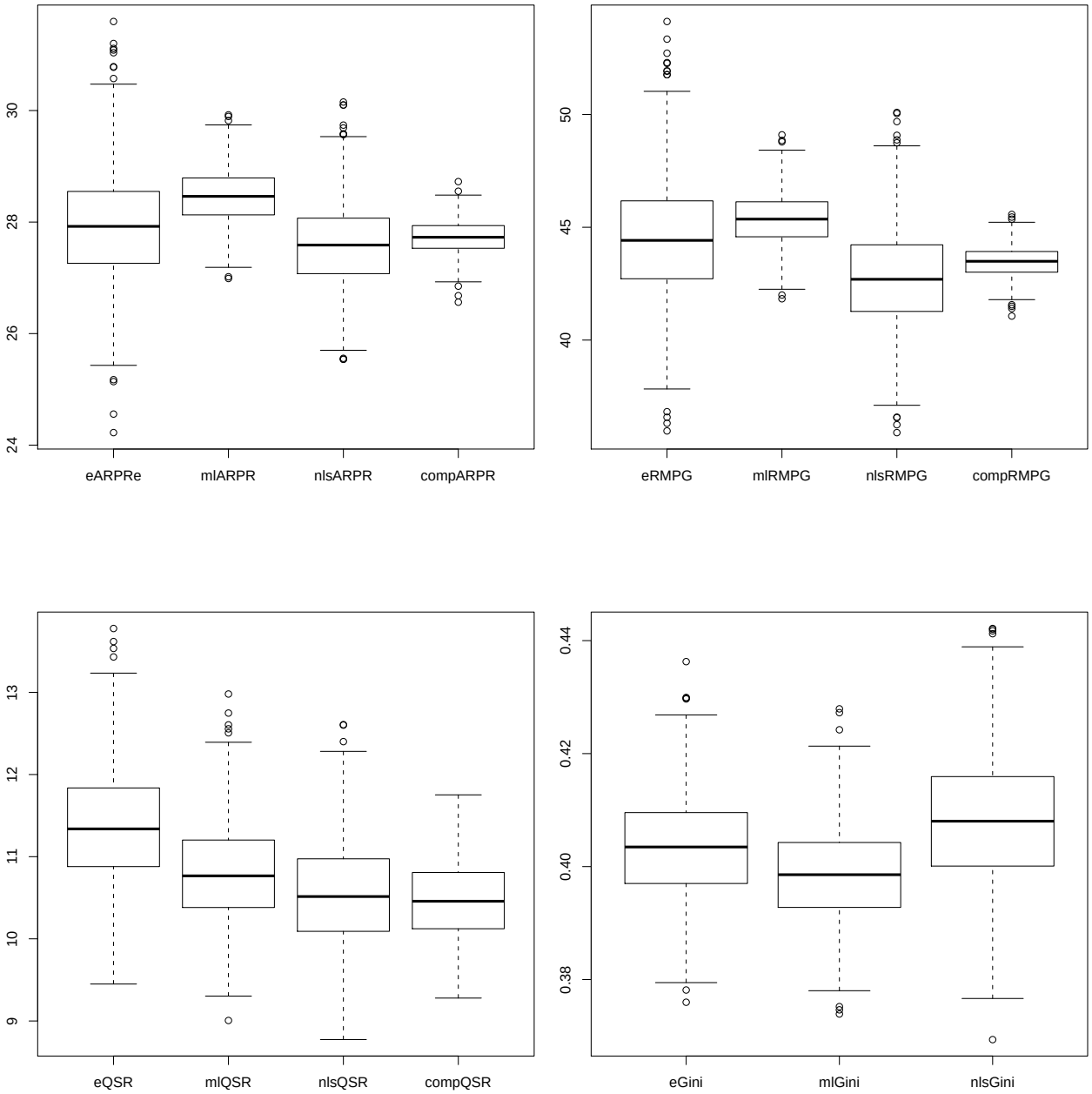


Figure 7.7: Fitted indicators, ML prof, d1.4a, region 1 (r1)

Discussion

Our simulations show that the GB2 model fits successfully various empirical income distributions, as is the case for the AMELIA synthetic universe, which is a very heterogeneous population.

We have developed various methods of estimation of the model parameters. The simulations clearly show the method of pseudo maximum likelihood estimation using the profile log-likelihood fits good both model parameters and derived indicators (except for the ARPR) on the global and regional level. However, we have seen that if the sampling weights are robustified using the ad-hoc procedure we have developed, this improves considerably the quality of the estimation. Variance estimation using the sampling weights only and the full design information is applicable for this method. Thus, we can recommend its use for the calculations of selected indicators as for the RMPG.

The method of non-linear fit for indicators gives good results in reproducing the empirical estimates of the derived indicators, however for the moment, no variance estimation of the obtained estimates is developed and for that matter the method of maximum likelihood estimation is preferable

For sub-populations, the decomposition of the GB2 seems to give promising results, based on the simulations with the AMELIA regions. It also gives various possibilities for adapting to the data, in choosing different partitions and different fitted parameters for the initial decomposition, i.e. left or right tail decomposition. We have seen that for certain regions in our simulation study it works better than for others. Thus it still needs to be tested and developed, e.g. for the calculation of the Gini.

© http://ameli.surveystatistics.net/ - 2011

7.1.5 Recommendations

Full and profile likelihood Quality of results is the same for full and profile likelihood. However, the profile likelihood algorithm converges faster. So, we recommend to fit the GB2 using the profile likelihood algorithm.

Variance estimation by linearization The sandwich variance estimator using the linearization of the indirect indicators works well. With one-stage designs, the simplified formula involving only the sample weights gives good results and can be used with a sample size in the order of magnitude of the EU-SILC country level. With two-stage designs (like design 2.7) the full design information should be included, in the lines of (Graf et al., 2011, ch.4).

Robustified weights Comparison between the results with the survey weights and the robustified weights shows that the bias in the indirect (GB2) indicators' estimates using the GB2 fit is greatly reduced when the weights are robustified. The variance estimation however underestimates the true variance when the weights are robustified. The additional variability due to the weight adjustment should be taken into account, but has not been developed yet.

Left and right tail decomposition of the GB2 When applying a compound fit, one has to choose between the left tail or the right tail decomposition (Deliverable 2.1, Chapter 5). The left tail decomposition is appropriate when the focus is on the group differences in the distribution of the poor. If the study aims at comparing the rich, the right tail decomposition is a better tool. For the analysis of indicators of poverty, the first is preferred. Indicators of inequality can be scrutinized by the use of both.

Weights of the components in the mixture In our developments, the components in a GB2 decomposition are given in advance. With the Amelia universe, when there is a large discrepancy between the regions, it seems preferable to evenly distribute the components, which amounts to specify equal initial probabilities p_l. On the contrary, our tests with the Austrian data, taking the four NUTS1 as regions, showed better results with an uneven distribution of components, with one component for the very poor ($p_1 = 0.1$), a large middle component ($p_2 = 0.7$) and a third component for the rich ($p_3 = 0.2$).

7.2 Parametric Estimation Using Dagum Distributions

This section presents the results of the simulation study regarding a mixture of two Dagum distributions (TCD), presented in deliverable 2.1 chapter 6 (see Graf et al. (2011)). The denotations and abbreviation are also the same as in the cited deliverable. Subsection 7.2.1 describes the analysis of the distribution parameters. The results of the indicator estimation is the subject of subsection 7.2.2. The final subsection concludes the section.

7.2.1 Analysis of the Distribution Parameter

The fitting of the TCD is a highly non-trivial process. The resulting distribution parameters of some samples suggest that very different distribution parameters can still lead to similar estimators of the povery and inequality measures. Table 7.6 demonstrates this observation. It shows the parameters and indicators of two different samples (denoted as sample A and B), both taken according to design 1.2. It can be seen that the shape

sample	a	b	p	a_2	b_2	p_2	α	QSR	ARPR	Gini
A	5.22	63402	0.137	3.84	35353	0.484	0.389	10.09	0.269	0.381
B	66.0	70629	0.0088	3.57	37969	0.459	0.225	10.01	0.270	0.381

Table 7.6: Parameters and Measures of two Different Samples

parameters of the first component as well as the mixture parameter differ a lot between these two samples. On the other hand, the estimators for all three poverty/inequality measures are very close for the two samples. Figure 7.8 illustrates the density of sample A (red curve) and sample B (green curve). The two curves have very similar shapes

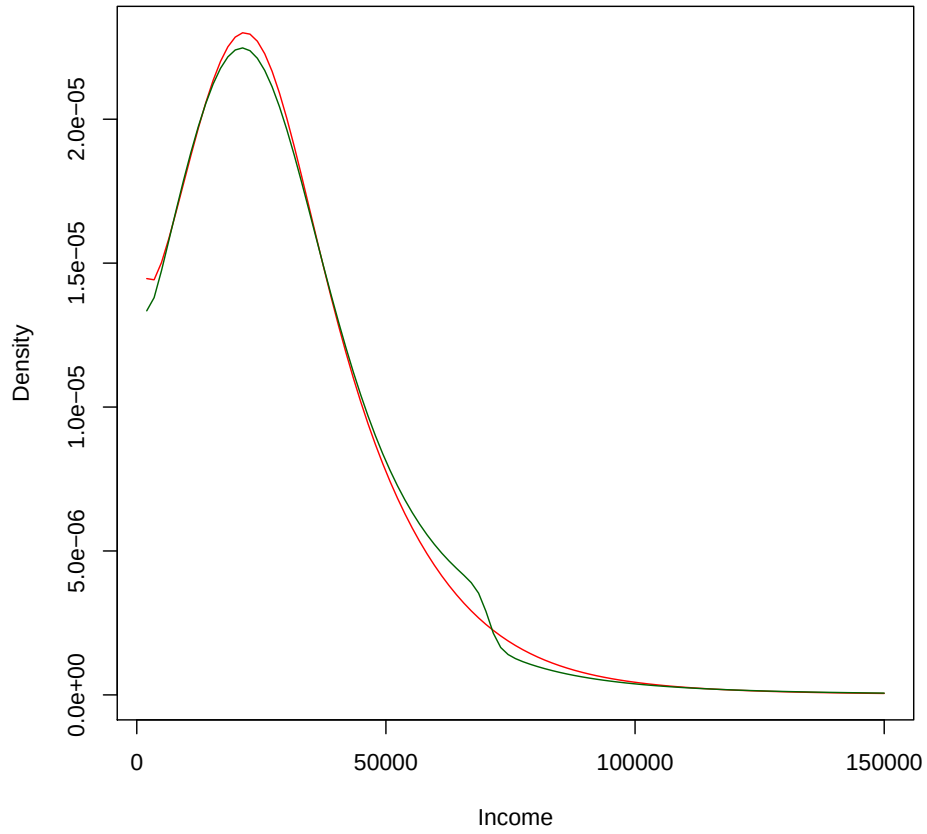


Figure 7.8: Two TCDs with Different Parameters

and therefore the indicator values do not vary much. This may also indicate that the

© http://ameli.surveystatistics.net/ - 2011

log-likelihood function of a TCD might have several local maxima in its parameter space. Because of this is seems reasonable to conduct further analysis on the fitting method.

7.2.2 Results of the Indicator Estimation

This subsection presents the estimation results of the QSR, the ARPR and the Gini. Due to numerical problems of the variance estimation of the indirect approach, the number of evaluated samples had to be reduced to 699. Figure 7.9 shows the results for the point estimators of the QSR. The figure includes histograms of the estimators, the estimated densities (black curves) of their distributions and the denity of a normal distributions with mean and standard deviation as the empirical values from the simulation (green curves). Its legends provide the mean values of the distributions, the true value of the universe (TV) and the 95% confidence interval coverage rates (CR).

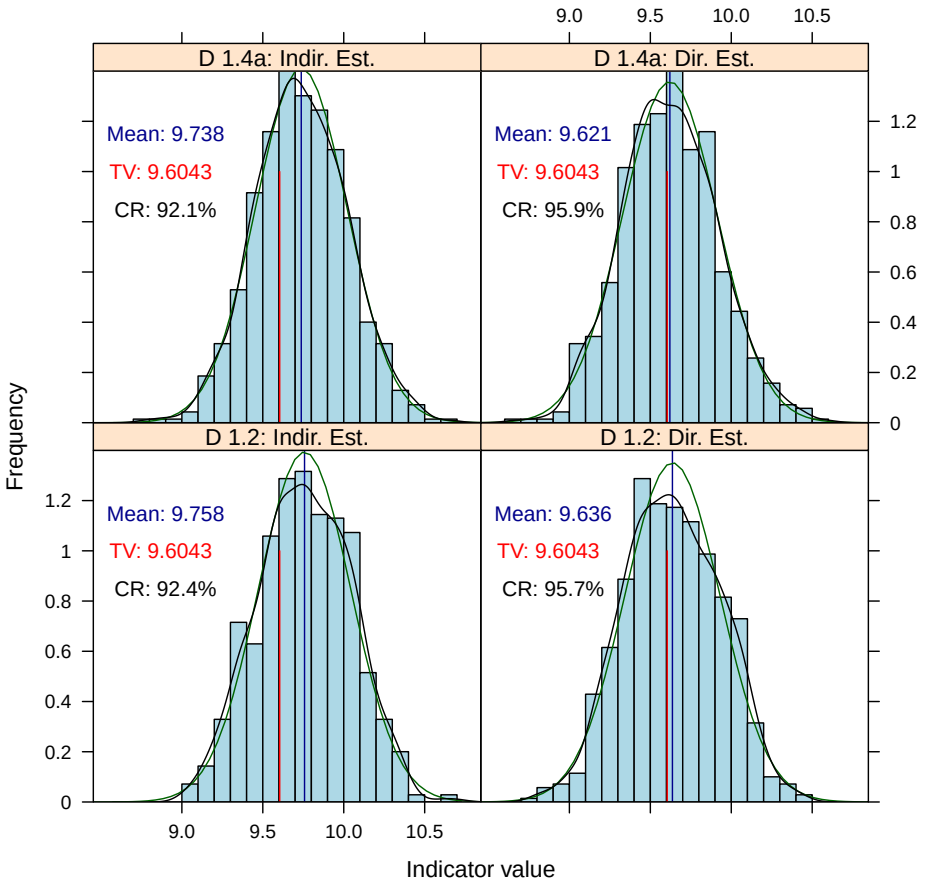


Figure 7.9: Point Estimators of the QSR

As a matter of fact the indirect estimation is biased, with a relative bias of 1.3% for D 1.2 and 1.6% for D 1.4a, whereas the direct estimation (with standard design-based estimators) can be interpreted as unbiased. In addition to that, the coverage rates are

considerably lower for the indirect estimation for both designs. The variance estimators of the QSR are diagrammed in figure 7.10. One key observation is that the benchmark (i.e.

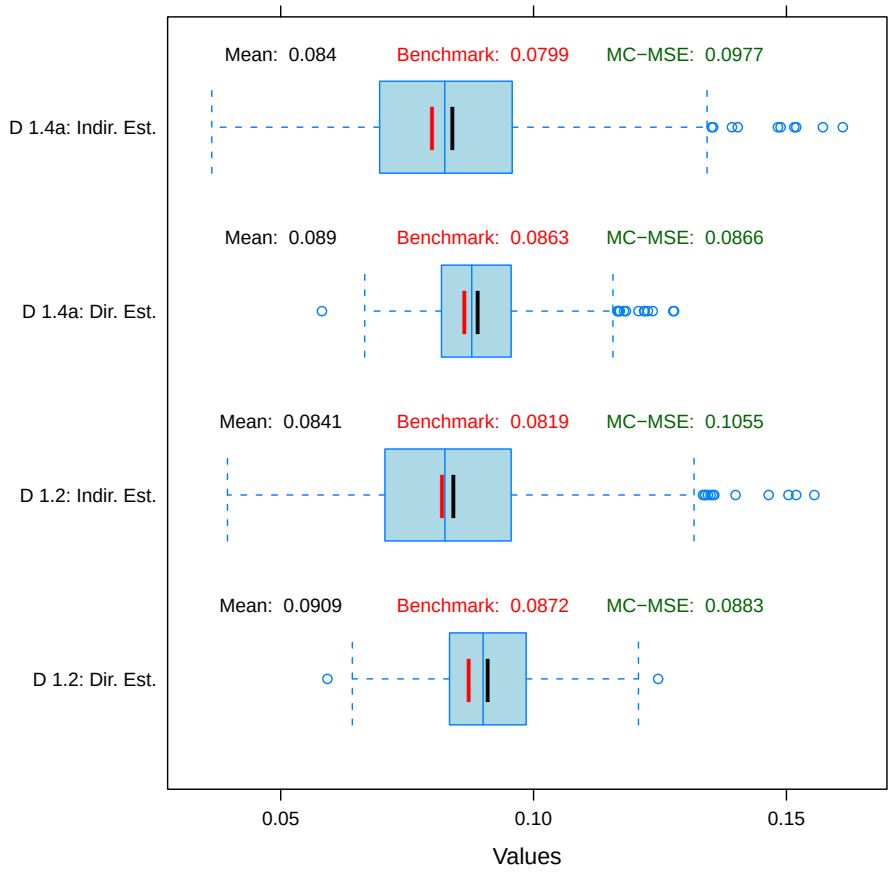


Figure 7.10: Variance Estimators of the QSR

the variance of the point estimators) is lower than the mean of the variance estimators, independent of design and estimation method. Although benchmark and mean of the variance estimator are lower for the indirect estimation for both designs, the Monte Carlo mean squared error (MC-MSE) signals that it cannot compensate the bias. Thus, one could conclude that the estimation of the QSR is better done directly.

The results for the ARPR give a very different impression. They are presented in figure 7.11 (point estimators) and figure 7.12.

Regardless of the design, all estimators seem to be unbiased and the indirect estimator leads to clearly more efficient results. This is backed up by the MC-MSE criterion which is far lower for the indirect method. For the ARPR the indirect estimation method is very recommendable. One possible trade-off is the high Monte Carlo variance of the distribution of variance estimators. Some samples yield variance estimators of double the mean of all variance estimators. On that account it seems reasonable to aspire the development of linearization methods also for the TCD.

© http://ameli.surveystatistics.net/ - 2011

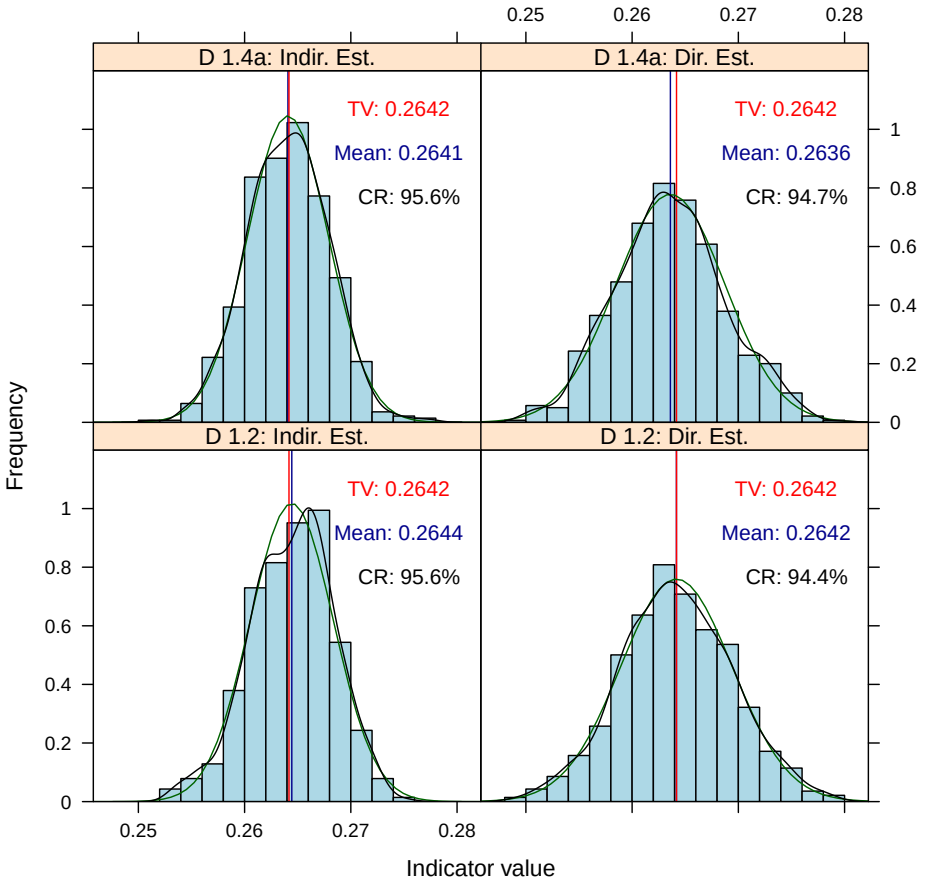


Figure 7.11: Point estimators of the ARPR

The third and final indicator investigated in the study is the Gini. Its estimation results are illustrated in figures 7.13 and 7.14.

The point estimators seem to be unbiased for both methods and both designs. Aside from that, the variance of the point estimator distribution of the indirect estimation is lower than the respective value of the direct method. Nevertheless, a few samples produce extremely high variance estimators for the indirect procedure for both designs. Some of the outliers are up to 16 times higher than the mean of the respective distribution. As already mentioned before, the variance estimation method for the indirect estimation has to be rethought.

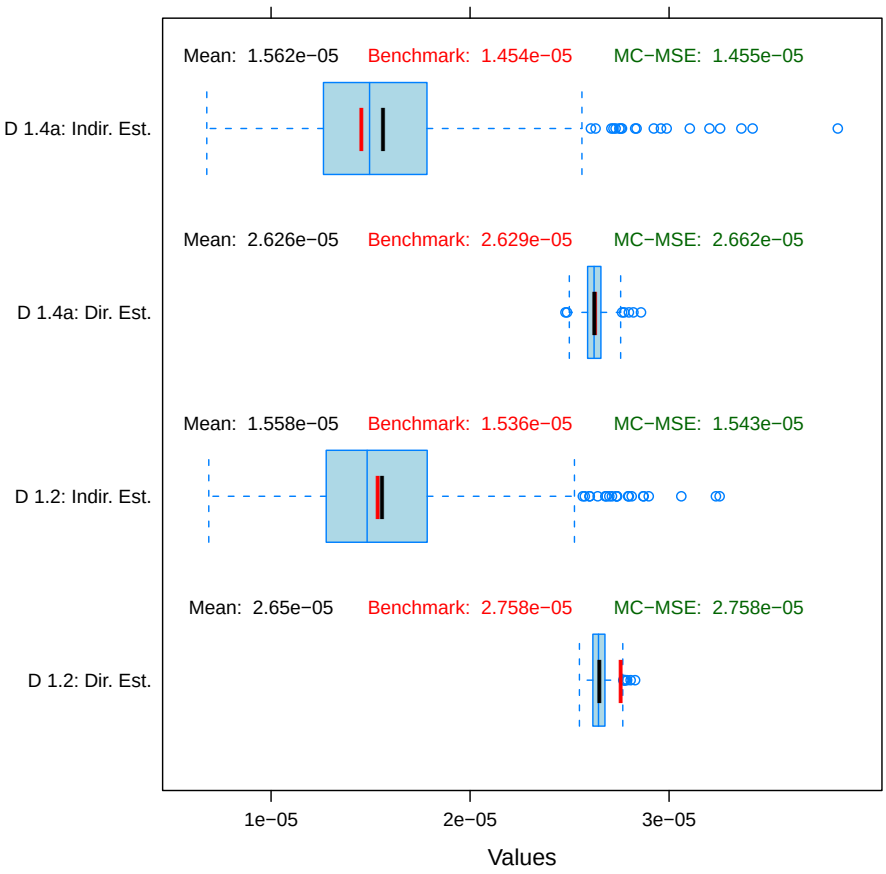


Figure 7.12: Variance estimators of the ARPR

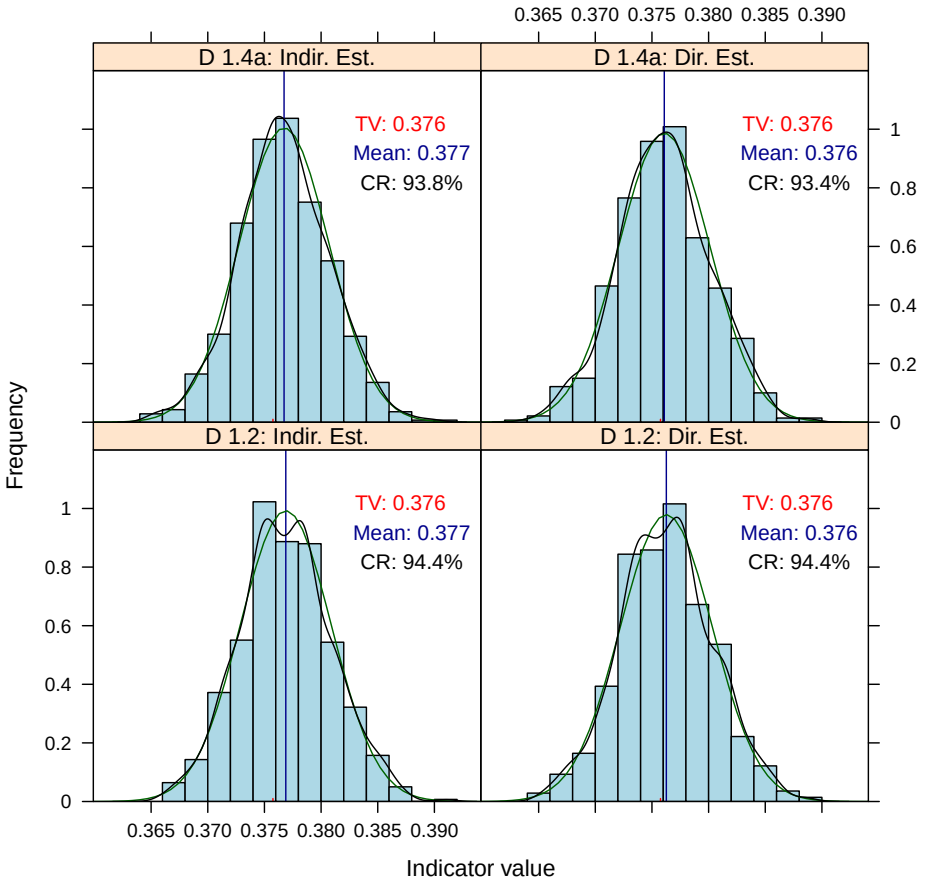


Figure 7.13: Point estimators of the Gini

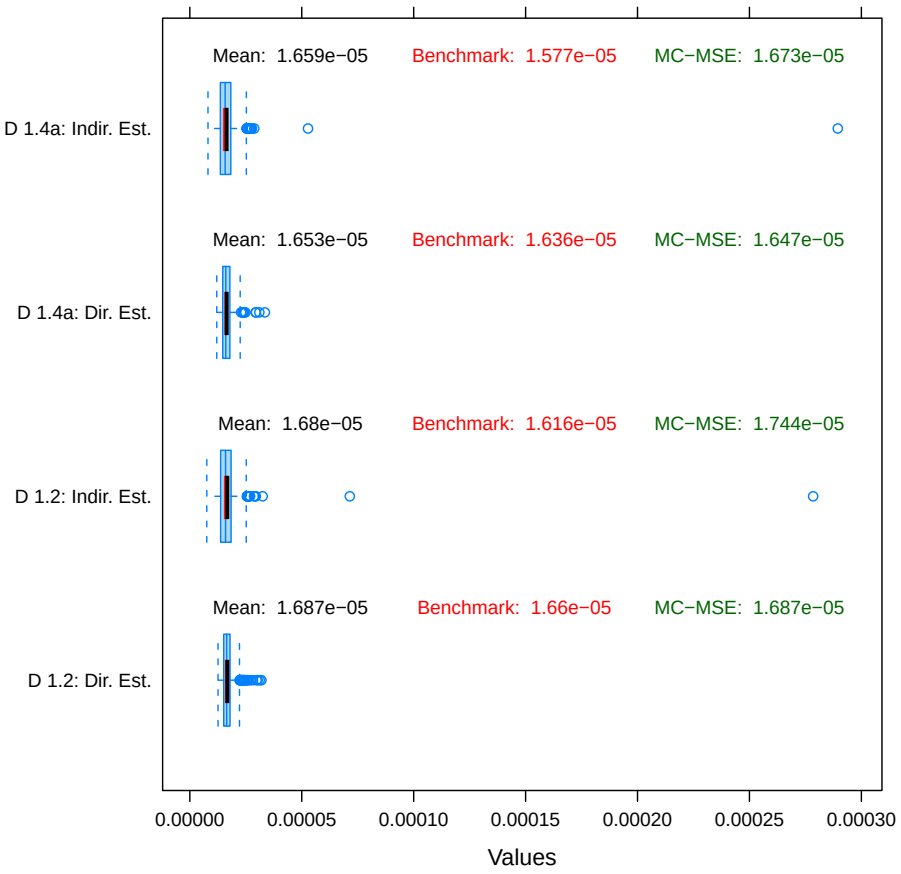


Figure 7.14: Variance estimators of the Gini

7.2.3 Conclusions and recommendations

All in all, the indirect estimation with a TCD provides decent estimation results, but cannot outperform the much faster direct estimation on all accounts. Due to this, it seems reasonable to enhance the fitting methodology of the TCD in the future, especially to tackle the global optimization problem. Furthermore, the need for the development of linearization methods for the variance estimation of the indirect method can be confirmed. With respect to the poverty- and inequality measures, the indirect method seems to be better for the ARPR, while the direct way of estimation seems to be beneficial for the QSR. The results for the Gini coefficient are rather ambiguous, but tend to favor the indirect estimation. In summary it can be said that the usage of a mixture of two Dagum components in this field looks promising but requires further investigation.

Bibliography

Graf, M., **Nedyalkova, D.**, **Münnich, R.**, **Seger, J.** and **Zins, S.** (**2011**): *Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion*. Research Project Report WP2 – D2.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Lumley, T. (**2010**): *survey: analysis of complex survey samples*. R package version 3.23-3.
URL http://cran.r-project.org/web/packages/survey/index.ht%ml

Chapter 8

WP2: Small Area Estimation

8.1 Simulation objectives

There are increasing demands for accurate statistics on poverty and social exclusion (poverty indicators for short) calculated for different population subgroups such as regional areas and demographic groups. One of the aims of the AMELI project was to investigate the current (standard) methods and develop new methods where appropriate. Estimators of selected poverty indicators (so-called Laeken indicators) for population subgroups or domains and small areas developed in the AMELI project are described in Deliverable 2.2 of Work Package 2 (Estimation). The simulation experiments with domain and small area estimation methods had the following objectives:

1. Investigation of the statistical properties (bias and accuracy) of the standard direct estimators of the selected poverty indicators for population domains and small areas. The standard estimators do not use auxiliary data or modelling.

2. Investigation of bias and accuracy of the new estimators, which use statistical models and auxiliary data at the unit level.

3. Investigation of bias and accuracy of the new estimators, which use statistical models and auxiliary data at an aggregated level.

4. Implementation of points 1 to 3 under equal and unequal probability sampling schemes.

5. For studying robustness of methods, the implementation of points 1 to 4 under various outlier contamination schemes.

6. Study of applicability of method incorporating a novel transformation of predictions.

7. Implementation of points 1 to 5 for two different populations, the register-based Finnish population and the more artificial Amelia population.

The indicators considered are the following:

© http://ameli.surveystatistics.net/ - 2011

- At-risk-of poverty rate
- The Gini coefficient
- Relative median at-risk-of poverty gap
- Quintile share ratio (S20/S80 ratio).

The indicators are nonlinear and are constructed using non-smooth functions such as medians and quintiles. They can be divided into two groups with respect to the selected estimation approach. For the estimation of at-risk-of poverty rate (poverty rate for short), we used generalized regression (GREG) and model calibration (MC) type estimators (as examples of design-based model assisted methods) and synthetic (SYN) and empirical best prediction (EBP) type estimators (featuring model-based methods). In all these estimators, logistic models were used because the underlying study variable is binary. Direct estimators, such as Horvitz-Thompson (HT) type estimators, were used as standard ("default") estimators.

Relative median at-risk-of poverty gap (poverty gap) and quintile share ratio (S20/S80 ratio) are examples of indicators that rely on medians or quantiles of the cumulative distribution function (CDF) of the underlying continuous equivalized income variable. For these indicators, direct design-based estimators were used as the standard estimators. The more advanced model-based indirect synthetic estimators use predictions calculated by using linear mixed models and auxiliary data at the unit level. In constructing the estimators, we use logarithmic transformation to correct for the skewness of the distribution of the study variable. Simple predictors were deemed substantially biased, so we developed more elaborate transformations aimed at improving the histogram of transformed predictions. Composite estimators were constructed as a linear combination of a design-based direct estimator and a model-based indirect estimator. In addition, for poverty gap we have studied the estimation of conditional expectations by simulation-based methods, resembling methods introduced in Molina and Rao (2010).

In many cases we assumed an access to unit-level auxiliary data on population elements, which is becoming an increasingly realistic assumption in statistical infrastructures of the EU countries. In addition, we developed frequency-calibrated prediction estimators that use aggregated auxiliary data. With our frequency-calibrated predictors, it is possible to use predictors when unit-level auxiliary information is not available. We also examine the properties of the estimators under different equal and unequal sampling designs and under outlier contamination. The underlying theory and derivations are presented in the Deliverable 2.2.

8.2 Simulation bed

Design bias and accuracy of estimators were examined by design-based simulation experiments. We used two populations: partially register-based Finnish population and synthetic Amelia population (Alfons et al., 2011, Deliverable 6.2, Report on Outcome of Simulation Study, March 2011). Programs written in R language have been supplied to apply the methods in practice.

8.2.1 Finnish population

The artificial Finnish population of one million persons was constructed from income data of seven NUTS3 -regions in Western Finland. The household properties, such as demographic composition and equivalized income were obtained from registers. Values of auxiliary variables of the household head were obtained from a household survey. Some personal auxiliary variables, most notably education level, had to be imputed for members other than household head, but the population was realistic enough for our simulation study.

In the simulations, $K = 1000$ samples of $n = 5000$ persons were drawn from the unit-level population. The sampling design was simple random sampling without replacement (SRSWOR) or PPS. For PPS, an artificial size variable was generated as a function of a qualitative variable. Then the PPS is approximately identical with stratified sampling. PPS was defined so that people with low income appear in samples with larger probability than people with large income. Therefore low education levels and certain socio-economic classes were given the largest inclusion probabilities. In PPS based on education level, the classes and relative inclusion probabilities are as follows (p is a constant depending on class frequencies):

Table 8.1: Education-class-specific inclusion probabilities

Education class	0	3	4	5	6	7	8
Inclusion probability	$5p$	$5p$	$4p$	$3p$	$2p$	p	p

Table 8.2: Socioeconomic-class-specific inclusion probabilities for PPS (by socstrat)

Socioeconomic class	1	2	3	4	5	6
Inclusion probability	$p/2$	$p/3$	$p/5$	p	p	p
Mean income	85069	68328	76491	58520	62448	56862

Our domains were 36 NUTS4 regions or 70 cells in the cross-classification of NUTS3 region, gender and age class (0-15, 16-24, 25-49, 50-64, and 65- years). These domains were classified by the expected sample size to size-classes with class boundaries at 50 and 100.

We created indicators for each class of a qualitative variable. The most commonly used model had auxiliary variables age and gender with interactions, socstrat and lfs-code. The corresponding linear fixed-effects model fitted to logarithms of income in the population had coefficient of determination $R^2 = 0.101$. When auxiliary variables house ownership and educ-thh were added to the model, the R^2 increased to 0.164.

8.2.2 Amelia population

From the synthetic Amelia data set constructed using SILC data (Alfons et al., 2011), we drew samples with SRSWOR ($n = 2000$) and PPS ($n = 6000$) based on a size variable with

© http://ameli.surveystatistics.net/ - 2011

Table 8.3: Specifcation of the auxiliary variables

Variable	**Label**	**Codes**
Age class	Age	0-15, 16-24, 25-49, 50-64, and 65-years
Gender	Gender	1 Males, 2 Females
House ownership	Indicator showing when the household owns the dwelling	0 (No), 1 (Yes)
Educ-thh	The number of household members having tertiary educational level	Count
Education	Education level of the household head	0 (Lowest) to 8 (Highest)
Empmohh	The total number of months of all household members being employed	Count
Socstrat	Socio-economic status of HH head	1 Wage and salary earners; 2 Farmers; 3 Other entrepreneurs; 4 Pensioners; 5 Other categories; 6 Not specified
Lfs-code	Employment status of HH member	1 Employed; 3 Unemployed; 3 Not in workforce

value 3 for education levels (ISCED) 0-3 and 2 for others. Forty regions (variable DIS) were classified by expected sample size with class boundaries at 45 and 55. Demographic domains were defined by age, gender and NUTS2 regions. For poverty rate, these domains were classified by size with breakpoints 50 and 100, for poverty gap with breakpoints 20 and 30. Our models fitted to the logarithm of the equivalized income variable EDI2 incorporated age class and gender with interactions, attained education level (ISCED), activity (working, unemployed, retired, or otherwise inactive) and degree of urbanisation (three classes).

8.3 Methods

We estimated the indicators by methods shown in Tables 8.4 and 8.5. The equations are in Deliverable 2.2.

Table 8.4: Poverty rate estimators.

Estimator	Description	Equations
Default	The default (direct) estimator of the poverty rate	(24)
Design-based estimators		
GREG	Generalized regression estimator assisted by a linear fixed-effects model	(26)
LGREG	Logistic GREG estimator assisted by a logistic fixed-effects model	(26)
MLGREG	GREG estimator (26) assisted by a logistic mixed model	(26)
MC	Model calibration; equation in parentheses e.g. MC(10)	(10), (12), (13)
Model-based estimators		
LSYN	Synthetic estimator based on a logistic fixed effects model	(14)
EBP	Empirical Best Predictor type estimator based on a logistic mixed model	(15)
EBP(Y)	Alternative EBP type estimator based on a logistic mixed model	(16)

8.4 Report on Simulations

From each sample, the following quality indicators were calculated for each domain estimator: mean, bias

$$Bias = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d),$$

absolute relative bias

$$ARB = \frac{|\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)|}{\theta_d}$$

© http://ameli.surveystatistics.net/ - 2011

Table 8.5: Estimators used in Gini coefficient, poverty gap and quintile share.

Estimator	Description	Equations
Default	The default (direct) estimator of the Laeken indicator	(27), (30), (33)
Model-based estimators		
Predictor	Estimator calculated from predicted values	(28), (31), (34)
Expanded predictor	Estimator (28), (31), or (34) from transformed predictions; used equation in parentheses	(17) or (18)
n-calibrated predictor	Predictor type estimator based on calibrated frequencies of fitted values	(18) and (19)
Composite estimators		
Composite	Composite estimator incorporating default estimator and expanded predictor	(17) or (18), (20)
n-calibrated composite	Composite estimator incorporating default estimator and frequency-calibrated predictor	(18), (19), (20)

and relative root mean squared error

$$RRMSE = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)^2}}{\theta_d}.$$

We present their averages over domain classes defined by domain size.

Outlier and contamination experiments were carried out as proposed in Hulliger and Schoch (2010), p. 7. Outliers were created in each sample without modifying the population. In OCAR, one percent of sampled persons were declared as outliers, chosen completely at random. In OAR, the probability of being an outlier varied as a function of labour force status and pensioner status as follows: 0.04 for employed people, 0.02 for the unemployed, 0.03 for people not in workforce but 0.01 for pensioners. In the Finnish population, the equivalized income of the outlier's household was the target of contamination, whereas in Amelia, the personal cash or near-cash income of an outlier was contaminated. Under CCAR contamination, a normally distributed value from $N(500000, 10000^2)$ was added to the target income variable. Under NCAR, the outlier's income value was multiplied by 1000. Under OAR, the expectations of contamination $N(\mu, 10000^2)$ were 5,000,000 for the employed, 4000 for the unemployed, 90000 for people not in workforce but 200 for pensioners. In Amelia, the equivalized income in the outlier's household was calculated anew using other personal and household-level income components. OAR contamination may sometimes result in a negative income when the original income is small. In simulations these were unfortunately left out from model fitting, as R replaces logarithms of negative values by missing values.

Most of the mixed models were fitted by R package nlme using maximum likelihood. Design weights were then not used. For Tables 8.8 and 8.9, we incorporated design weights into model fitting by glmer function of R package lme4. The linear and logistic fixed-effects models were fitted with GWLS and maximum pseudolikelihood methods incorporating design weights.

In the n-calibrated estimator, we treated socstrat as a variable unknown in the population. The marginal frequencies of socstrat classes were imputed with GREG assisted by multinomial logistic model (R module nnet).

8.5 Results

8.5.1 Poverty rate

Table 8.6 compares poverty rate estimators assisted by fixed effects models. Section a) shows results for a common model formulation where the model does not account for domain differences. NUTS3 indicators are included in Section b) to account for regional variation. Section c) includes domain-specific fixed effects. In this case the model-based LSYN and model-assisted LGREG coincide. Under SRSWOR, it was not necessary to include design weights in model fitting.

The default estimator, model calibration (MC), and GREG estimators are nearly design unbiased. Among these methods, model calibration based on (13) has the smallest RRMSE. In (13), the sums of fitted values were calibrated at NUTS3 level. Therefore there is not much difference between models (a) and (b). LSYN had the smallest RRMSE but it was design biased.

A logistic mixed model is used next to compare model-based EBP with model-assisted MLGREG (Table 8.7). Domain differences are accounted for by regional-level (Section a) or domain-specific (Section b) random intercepts in the model. In both cases, the EBP estimator has large negative design bias, especially for small domains, and MLGREG appears nearly design unbiased as expected. However, EBP shows better accuracy than MLGREG and other nearly unbiased methods of Table 8.6. MLGREG has somewhat larger bias than LGREG.

From tables 2 and 3 we see that random intercepts or fixed effects associated with NUTS3 regions yield better results than domain-specific effects.

Tables 8.8 and 8.9 show the effect of incorporating the design weights in fitting a mixed model. If the variable socstrat determining the size variable in PPS is not included in the model (Table 8.8), using design weights in fitting (EBP(Y)-W, no socstrat) results in smaller bias and RRMSE than model fitting without weights (EBP(Y), no socstrat). When socstrat was included in the model, EBP(Y)-W had smaller design bias than EBP(Y) but slightly larger RRMSE. MLGREG did not yield as small RRMSE as EBP(Y), but it had smaller bias. MLGREG-W benefitted slightly from using design weights in model fitting. We draw similar conclusions from Table 8.9.

Table 8.10 shows how contamination affects poverty rate estimators. A robust method of fitting the logistic mixed model was not available. Nevertheless, the poverty rate estimators are fairly robust. Only when the proportion of outliers is 15%, bias especially is large. EBP(Y) has the smallest RRMSE in this experiment. It was also least affected by contamination.

© http://ameli.surveystatistics.net/ - 2011

Table 8.6: Poverty rate estimators assisted by logistic and linear fixed effects models (Finnish data set).

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
a) Common model formulation									
Default	−0.04	0.03	-0.06	1.23	0.94	0.71	51.83	32.0	22.9
LSYN	-1.11	-0.32	0.54	13.95	12.39	5.87	18.16	16.54	10.15
LGREG	0.03	0.03	-0.05	0.68	0.87	0.65	48.55	30.55	20.66
GREG	0.03	0.03	-0.05	0.76	0.88	0.65	48.89	30.86	20.89
MC(12)	0.03	0.03	-0.05	0.68	0.86	0.65	48.56	30.55	20.65
MC(13)	0.03	0.03	-0.05	0.75	0.85	0.67	48.39	30.51	20.63
MC(10)	-0.12	0.01	-0.06	1.73	0.89	0.68	52.94	31.3	20.88
b) NUTS3 indicators added to x-variables									
Default	-0.06	0.02	-0.07	1.21	0.93	0.73	51.82	31.98	22.29
LSYN	-0.01	0.12	-0.14	7.98	8.05	4.9	19.35	15.99	11.41
LGREG	0.02	0.02	-0.06	0.71	0.84	0.68	48.4	30.5	20.66
GREG	0.02	0.02	-0.06	0.79	0.86	0.67	48.74	30.81	20.88
MC(12)	0.02	0.02	-0.06	0.72	0.84	0.68	48.42	30.51	20.64
MC(13)	0.02	0.02	-0.06	0.73	0.83	0.69	48.39	30.5	20.64
MC(10)	-0.15	0.0	-0.07	1.83	0.89	0.7	52.85	31.29	20.88
c) Domain indicators added to x-variables									
Default	–	–	–	1.21	0.93	0.73	51.82	31.98	22.29
LSYN	–	–	–	1.18	0.83	0.7	50.98	30.9	20.81
LGREG	–	–	–	1.18	0.83	0.7	50.98	30.9	20.81
GREG	–	–	–	1.08	0.85	0.67	50.84	31.1	20.98
MC(12)	–	–	–	1.15	0.82	0.7	51.04	30.93	20.81
MC(13)	–	–	–	1.09	0.84	0.7	50.7	30.95	20.83
MC(10)	–	–	–	1.89	0.88	0.7	52.45	31.26	20.88

Design: SRSWOR. Qualitative x: house ownership, age class, gender, lfs-code. Domains: NUTS3 by age by gender (D = 70 domains)

Table 8.11 shows how contamination affects estimators under PPS. The bias of EBP(Y) is larger than in Table 8.10, with the exception of contamination of 15%. The RRMSE of other methods are larger than under SRSWOR.

Table 8.7: Poverty rate estimators assisted by a logistic mixed model (Finnish data set).

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
a) NUTS 3 level random intercepts									
EBP(Y)	-1.47	-0.53	0.02	14.85	10.75	4.07	20.83	17.22	10.81
MLGREG	0.01	0.03	-0.05	0.66	0.87	0.68	48.66	30.72	20.75
b) Domain-specific random intercepts									
EBP(Y)	-1.43	-0.55	0.16	14.75	8.96	3.99	22.49	19.26	14.54
MLGREG	0.28	0.13	-0.27	2.2	3.44	2.76	55.67	39.87	40.44

Design: SRSWOR. Qualitative x: house ownership, lfs-code, age class, gender. Domains: NUTS3 by age by gender. Mixed model with NUTS3 random intercepts was fitted by nlme.

Table 8.8: Poverty rate estimators with design weights incorporated in model fitting (lme4) in methods with suffix "W". (Finnish data set)

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
Default	1.60	1.13	0.54	1.17	54.18	30.21	20.95	36.79
EBP(Y)	11.84	8.21	5.01	8.82	19.73	15.61	11.63	16.23
EBP(Y), no socstrat	13.40	9.88	7.37	10.60	20.94	16.94	12.93	17.51
EBP(Y)-W	9.33	8.04	5.57	7.97	20.00	16.23	12.39	16.76
EBP(Y)-W, no socstrat	9.58	8.27	5.47	8.14	20.01	16.38	12.43	16.83
MLGREG	1.56	1.13	0.59	1.17	53.95	30.22	20.89	36.69
MLGREG-W	1.57	1.14	0.58	1.17	53.64	30.12	20.82	36.53

Design: PPS based on socstrat. Qualitative x: age and gender with interactions, lfs-code and socstrat. Domains: NUTS3 by age by gender. Logistic mixed model with NUTS3 random intercepts was fitted by lme4.

Table 8.9: Poverty rate estimators in Amelia. Design weights are incorporated in model fitting (lme4) in methods with suffix "W".

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
Default	0.76	0.61	0.32	0.67	29.14	23.08	17.36	26.09
EBP(Y)	8.29	9.25	7.78	8.56	13.50	13.77	10.92	13.36
EBP(Y), no ISCED	8.67	10.01	7.88	9.04	13.98	14.52	11.51	13.93
EBP(Y)-W	8.35	8.93	7.92	8.50	13.61	13.65	11.13	14.40
EBP(Y)-W, no ISCED	8.30	8.96	7.77	8.47	13.68	13.69	11.02	13.44
MLGREG	0.74	0.57	0.29	0.64	28.13	22.34	16.90	25.21
MLGREG-W	0.74	0.57	0.29	0.64	28.12	22.34	16.89	25.21

Design: PPS based on ISCED. Qualitative x: age and gender with interactions, ISCED, activity, and degree of urbanisation. Domains: NUTS2 by age by gender. Logistic mixed model with NUTS2 random intercepts was fitted by lme4.

© http://ameli.surveystatistics.net/ - 2011

Table 8.10: Poverty rate in contaminated Finnish data.

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
Baseline (no contamination)								
Default	1.11	1.04	0.50	0.94	51.94	31.79	22.04	36.76
MLGREG	1.41	0.98	0.42	1.01	48.83	30.81	20.93	34.99
EBP(Y)	9.00	8.21	5.36	7.84	19.91	17.51	12.51	17.23
OCAR-CCAR 1%								
Default	1.69	1.29	0.50	1.25	52.13	31.93	22.11	36.90
MLGREG	1.91	1.25	0.45	1.30	49.04	30.94	21.00	35.13
EBP(Y)	8.47	8.52	5.33	7.77	19.68	17.73	12.54	17.28
OCAR-NCAR 1%								
Default	1.65	1.31	0.50	1.25	52.16	31.91	22.11	36.90
MLGREG	1.94	1.27	0.44	1.32	49.09	30.94	21.00	35.15
EBP(Y)	8.48	8.53	5.34	7.78	19.77	17.76	12.54	17.28
OAR-CAR								
Default	1.73	1.10	0.63	1.22	52.17	31.93	22.09	36.91
MLGREG	1.88	1.10	0.58	1.26	49.06	30.93	20.95	35.12
EBP(Y)	8.66	8.58	5.45	7.89	19.67	17.76	12.60	17.26
OCAR-CCAR 15%								
Default	23.36	15.81	4.93	16.02	63.25	39.23	23.45	44.20
MLGREG	23.72	15.84	4.92	16.16	60.43	38.25	22.46	42.56
EBP(Y)	21.24	20.04	6.17	17.30	28.87	27.20	13.72	24.71

Design: SRSWOR. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS3 by gender and age class (70 domains). Logistic mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Table 8.11: Poverty rate in contaminated data (Finnish data set).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
Baseline (no contamination)								
Default	1.60	1.13	0.54	1.17	54.18	30.21	20.95	36.79
MLGREG	1.53	1.15	0.57	1.16	53.99	30.20	20.83	36.69
EBP(Y)	11.85	8.76	6.97	9.48	20.26	16.20	12.76	16.91
OCAR-CCAR 1%								
Default	2.04	1.39	0.63	1.46	54.33	30.28	20.94	36.87
MLGREG	2.13	1.41	0.64	1.50	54.14	30.26	20.83	36.77
EBP(Y)	11.41	8.68	7.24	9.35	20.01	16.15	12.92	16.84
OCAR-NCAR 1%								
Default	2.01	1.37	0.68	1.45	54.38	30.29	20.97	36.90
MLGREG	2.10	1.39	0.70	1.50	54.17	30.29	20.85	36.79
EBP(Y)	11.41	8.71	7.26	9.36	20.04	16.21	12.96	16.88
OAR-CAR								
Default	2.35	1.20	0.86	1.54	54.29	30.14	20.94	36.79
MLGREG	2.33	1.23	0.81	1.53	54.04	30.13	20.82	36.67
EBP(Y)	11.49	8.81	7.61	9.51	20.00	16.25	13.18	16.93
OCAR-CCAR 15%								
Default	21.53	14.86	10.08	16.22	63.75	36.99	26.08	44.21
MLGREG	21.97	14.78	10.16	16.36	63.66	36.94	25.99	44.14
EBP(Y)	17.73	16.87	12.57	16.26	25.95	23.20	18.57	23.19

Design: PPS by socio-economic status. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS3 by gender and age class (70 domains). Logistic mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

© http://ameli.surveystatistics.net/ - 2011

8.5.2 The Gini coefficient

Table 8.12 shows an experimental comparison of the expanded predictor (17) of the Gini coefficient, the default estimator and the ordinary predictor (28). Benefits from the expansion (17) are obvious.

Table 8.12: Estimators of Gini coefficient assisted by linear mixed model (Finnish data set).

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default	-.007	-0.004	-.002	2.92	1.57	0.66	14.09	11.42	7.66
Predictor	-.066	-0.066	-.063	27.96	28.14	26.18	28.12	28.30	26.34
Expanded predictor (17)	-.004	-0.003	-.005	3.97	3.04	3.44	4.43	3.56	3.86
Composite	-.005	0.001	-.004	3.46	2.11	2.56	5.79	4.29	3.91

Design: SRSWOR. Quantitative x: educ-thh, empmohh. Qualitative x: house ownership, lfs-code, socstrat. Domains: 36 NUTS4 regions. Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Tables 8.13 and 8.14 summarize experiments with contamination. The expanded predictor and frequency-calibrated predictor are better methods than the default one. They are also fairly robust. Composite estimators have large design bias in the most contaminated data. In OCAR-NCAR, the bias and RRMSE of expanded predictor and frequency-calibrated estimator are larger under PPS than under SRSWOR.

Table 8.13: Gini coefficient in contaminated data (Finnish data set).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	3.27	1.74	0.66	1.56	14.28	11.36	7.57	10.40
Expanded predictor (18)	4.55	6.37	3.12	4.94	5.10	6.68	3.72	5.39
Composite	2.27	3.59	2.18	2.90	6.28	5.94	3.70	5.18
Predictor	49.38	50.15	48.74	49.53	49.72	50.49	49.05	49.86
n-calibrated predictor	3.06	4.64	2.95	3.81	5.06	5.64	3.70	5.18
n-calibrated composite	2.46	2.84	2.09	2.51	6.09	5.37	3.64	4.85
OCAR-CCAR 1%								
Default	14.76	17.67	17.64	17.26	33.52	29.92	22.66	27.80
Expanded predictor (18)	13.15	14.98	9.38	12.70	13.34	15.14	9.63	12.90
Composite	13.00	15.66	12.00	13.97	18.59	18.36	13.10	16.49
Predictor	49.92	50.68	49.29	50.07	50.07	50.83	49.44	50.22
n-calibrated predictor	8.93	12.59	8.83	10.73	10.23	13.15	9.20	11.32
n-calibrated composite	10.50	14.11	11.66	12.72	16.53	16.95	12.84	15.41
OCAR-NCAR 1%								
Default	98.84	151	231	173	173	212	254	223
Expanded predictor (18)	15.31	17.08	11.26	14.73	15.73	17.49	11.77	15.18
Composite	68.85	111	193	135	116	153	212	169
Predictor	48.12	48.85	47.64	48.31	48.98	49.70	48.43	49.14
n-calibrated predictor	10.78	14.53	10.67	12.61	12.27	15.34	11.30	13.45
n-calibrated composite	63.22	107	192	131	108	148	211	165
OAR-CAR								
Default	88.07	118	141	122	139	152	151	150.05
Expanded predictor (18)	25.32	27.29	19.72	24.28	25.48	27.45	19.90	24.45
Composite	68.17	91.56	113	96.21	102.31	114	122	115
Predictor	32.54	33.10	31.96	32.61	34.44	34.99	33.66	34.43
n-calibrated predictor	20.67	24.79	19.07	22.15	21.59	25.25	19.35	22.61
n-calibrated composite	64.93	90.04	113	95.00	97.92	113	122	114

Design: SRSWOR. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS4. Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

© http://ameli.surveystatistics.net/ - 2011

Table 8.14: Gini coefficient in contaminated data under PPS (Finnish data set).

.

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	4.11	2.40	0.84	2.08	16.58	13.17	8.58	11.99
Expanded predictor (18)	4.55	6.33	3.11	4.92	5.11	6.65	3.74	5.38
Composite	1.88	3.19	2.15	2.63	7.55	6.33	3.96	5.64
Predictor	47.58	48.22	47.04	47.70	47.65	48.29	47.11	47.77
n-calibrated predictor	3.32	4.75	3.00	3.92	5.05	5.52	3.72	4.80
n-calibrated composite	2.72	2.61	2.09	2.44	7.27	5.76	3.91	5.30
OCAR-CCAR 1 %								
Default	13.37	17.00	17.37	16.63	33.59	30.93	23.34	28.56
Expanded predictor (18)	12.48	14.18	8.76	11.99	12.73	14.42	9.11	12.26
Composite	11.87	14.90	11.35	13.20	18.13	17.88	12.65	16.03
Predictor	47.96	48.59	47.44	48.09	48.04	48.66	47.52	48.16
n-calibrated predictor	7.90	12.01	8.48	10.17	9.38	12.50	8.91	10.77
n-calibrated composite	9.21	13.44	11.17	12.03	16.06	16.57	12.52	15.04
OCAR-NCAR 1 %								
Default	93.11	149.85	229.28	170.65	168.79	211.05	251.31	219.72
Expanded predictor (18)	21.90	23.84	17.22	21.18	24.74	26.66	20.09	24.02
Composite	69.35	113.65	194.25	136.60	120.81	157.17	212.86	172.23
Predictor	46.54	47.11	46.19	46.70	46.68	47.25	46.33	46.84
n-calibrated predictor	16.26	21.06	16.87	18.88	20.06	24.17	19.86	22.04
n-calibrated composite	64.69	110.71	193.96	134.38	114.04	153.48	212.57	169.34
OAR-CAR								
Default	69.91	100.96	132.73	108.12	127.63	144.19	148.67	143.51
Expanded predictor (18)	24.46	26.30	19.01	23.41	24.63	26.47	19.20	23.59
Composite	55.00	77.76	102.93	83.69	92.71	105.97	114.76	107.30
Predictor	30.66	31.06	30.22	30.70	30.78	31.18	30.34	30.82
n-calibrated predictor	19.27	23.86	18.62	21.33	20.29	24.28	18.89	21.78
n-calibrated composite	51.80	76.02	102.87	82.35	87.90	103.80	114.71	105.53

Design: PPS by socio-economic status. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS4. Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme without design weights.

Table 8.15: Gini coefficient in contaminated Amelia data under SRSWOR.

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	2.60	2.10	1.61	2.11	12.89	11.64	10.43	11.68
Expanded predictor (18)	10.69	8.37	7.71	8.89	11.51	9.33	8.69	9.81
Composite	5.46	4.21	4.01	4.53	8.05	6.76	6.14	6.98
Predictor	21.75	23.27	23.74	22.94	22.59	24.00	24.44	23.70
n-calibrated predictor	6.22	4.31	3.74	4.72	11.32	9.06	7.84	9.40
n-calibrated composite	3.70	2.33	2.03	2.66	8.58	7.10	6.15	7.28
OCAR-CCAR 1%								
Default	7.77	8.75	9.79	8.74	21.48	20.63	19.94	20.70
Expanded predictor (18)	12.65	10.29	9.63	10.81	13.38	11.12	10.50	11.63
Composite	10.82	9.59	9.51	9.94	13.32	11.91	11.46	12.21
Predictor	22.25	23.75	24.22	23.43	23.09	24.49	24.92	24.18
n-calibrated predictor	7.80	6.00	5.53	6.41	12.46	10.09	8.96	10.49
n-calibrated composite	7.66	6.71	6.70	6.99	12.13	10.67	9.99	10.92

Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation. Domains: districts (DIS). Mixed model with DIS random intercepts was fitted to log(income+1) by lme without design weights.

© http://ameli.surveystatistics.net/ - 2011

8.5.3 Poverty gap

Our experiments imply that poverty gap is the most difficult Laeken indicator to estimate, considering the large RRMSE of all estimators. Table 8.16 shows an experiment with a lot of auxiliary information. All poverty gap estimators, even the default estimator have design bias in small domains, probably due to the non-linear formulation of the indicator. The ordinary predictor (31) is far too biased to be useful. The expanded predictor and corresponding composite estimator are better than the default estimator especially in small domains.

Table 8.16: Poverty gap estimators assisted by a linear mixed model (Finnish data set).

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default	2.1	0.9	0.4	12.14	4.37	1.78	65.85	43.58	27.26
Predictor	-6.8	-9.8	-14.6	40.09	43.36	57.47	61.49	57.09	62.09
Expanded predictor (17)	-3.1	-3.0	-3.6	17.01	19.61	16.58	23.85	25.43	22.92
Composite	-1.7	-2.1	-2.5	10.91	14.41	11.90	25.63	22.39	18.63

Design: SRSWOR Quantitative x: educ-thh, empmohh. Qualitative x: house ownership, lfs-code, socstrat. Domains: NUTS3 by age by gender (70 domains). Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

The amount of auxiliary data seems to have an effect on the poverty gap estimation results: in Table 8.17 involving less auxiliary data than Table 8.16, the expanded predictor and the frequency-calibrated poverty gap estimator are significantly better than the default estimator only in the smallest domains (expected sample size smaller than 50). Moreover, they are severely biased. The corresponding composite estimators perform better, also in the large domains. Some composite estimators could not be calculated due to limited time. All estimators except the ordinary predictor are robust. Actually, contamination often seemingly improves the properties of estimators.

The simulation-based method (23) yields fairly good poverty gap estimates, although there seems to be systematic bias: estimates are too large in small domains and too small in large domains (Table 8.18). As a result, the poverty gap differences between domain size classes apparent in estimation by the default method are not seen in estimates based on the simulation-based method.

Although these results are promising, experiments with Gini coefficient and quintile share were disappointing due to large bias. The distribution of the equivalized incomes differs from assumed log-normal distribution: there are fewer rich people than expected. As a consequence, some of the simulated incomes were unrealistically large. However, in other countries, the distribution of equivalized incomes may be closer to log-normal, and then the method of Molina and Rao is probably the best method available, if minimization of MSE is required. Better results might also be obtained with a more realistic income distribution.

Table 8.19 compares two bootstrap techniques used in estimating the MSE of the synthetic component in a composite estimator. K=500 samples were used in the bootstrap and

Table 8.17: Poverty gap in contaminated data (Finnish data set).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination, SRSWOR								
Default	13.15	5.14	2.07	7.30	66.91	44.17	27.57	48.50
Expanded predictor (18)	45.85	40.04	44.42	43.11	51.91	43.92	47.69	47.64
Composite	28.58	24.24	22.33	25.35	43.28	32.65	29.22	35.66
Predictor	49.85	56.74	62.77	55.66	80.02	75.73	73.33	76.71
n-calibrated predictor	42.42	36.58	39.13	39.25	64.08	48.56	48.45	54.08
n-calibrated composite	23.34	21.74	19.72	21.85	47.37	34.83	29.17	38.02
No contamination, PPS								
Default	13.54	7.66	2.30	8.61	69.74	45.85	28.18	50.60
Expanded predictor (18)	45.03	40.06	45.92	43.09	52.50	45.41	48.96	48.70
Predictor	52.61	52.73	53.76	52.91	67.09	63.09	64.83	64.89
n-calibrated predictor	42.83	37.53	45.84	41.20	59.99	47.18	51.25	52.63
OCAR-CCAR 1%, PPS								
Default	13.11	7.50	2.06	8.34	69.24	45.68	28.17	50.34
Expanded predictor (18)	44.83	39.87	45.99	42.95	52.62	45.43	48.97	48.76
Predictor	55.16	56.38	57.39	56.16	69.04	66.34	67.88	67.64
n-calibrated predictor	42.51	37.28	45.48	40.91	59.79	47.04	50.97	52.44
OCAR-CCAR 15%, SRSWOR								
Default	9.68	6.92	4.20	7.28	59.46	41.08	27.71	44.59
Expanded predictor (18)	41.61	35.35	40.77	38.83	52.33	41.82	45.51	46.42
Composite	25.59	20.18	19.37	21.93	41.75	30.27	27.26	33.68
Predictor	92.76	94.22	95.28	93.94	103.19	101.02	99.66	101.49
n-calibrated predictor	41.18	34.02	37.06	37.27	62.97	46.69	46.28	52.41
n-calibrated composite	23.19	18.85	17.59	20.11	45.43	32.37	27.38	35.90
OCAR-NCAR 15%, PPS								
Default	10.57	6.82	5.48	7.87	64.08	42.27	27.75	46.95
Expanded predictor (18)	34.45	30.56	36.93	33.31	53.01	43.52	46.25	47.50
Predictor	99.27	99.38	99.42	99.35	99.52	99.55	99.58	99.55
n-calibrated predictor	34.57	29.55	37.48	33.04	59.73	45.36	48.22	51.11

Design: SRSWOR or PPS by socio-economic status. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS3 by gender and age class (70 domains). Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

RAST correction was applied. Estimating the MSE of the synthetic component in the composite estimator by parametric bootstrap may yield small benefits over the simple equation (21), but it requires much more computing time.

© http://ameli.surveystatistics.net/ - 2011

Table 8.18: Poverty gap estimation by the simulation-based method of Molina and Rao (2010). (Finnish data set.)

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Simulation-based	2.42	-0.41	-3.59	35.96	19.14	13.51	41.28	24.96	17.77
Default	0.72	1.02	0.37	10.09	4.82	1.85	69.66	44.18	27.54

Design: SRSWOR. Quantitative x: educ-thh, empmohh. Qualitative x: house ownership, lfs-code, socstrat. Domains: NUTS3 by age by gender Mixed model with NUTS3 random intercepts was fitted by nlme.

Table 8.19: Composite estimates (32) of poverty gap with MSE of synthetic component estimated by ordinary bootstrap (21) or by parametric bootstrap (22) (Finnish data).

Estimator	ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major
ordinary bootstrap	11.30	14.76	12.22	25.65	22.64	18.63
parametric bootstrap	11.25	13.98	12.56	25.22	22.60	18.69

Design: PPS by education level. Quantitative x: educ-thh, empmohh.Qualitative x: house ownership, lfs-code, socstrat. Domains: NUTS3 by age by gender. A mixed model with NUTS3 random intercepts was fitted by nlme without using design weights.

8.5.4 Quintile share

Table 8.20 shows experimental results with quintile share estimators assisted by a linear fixed-effects model. The ordinary predictor (34) is definitely design biased. The expanded predictor yields much better results than the default estimator in all domain size classes. It does not have much design bias.

Table 8.20: Quintile share estimators assisted by a linear fixed effects model (Finnish data set).

Estimator	BIAS			ARB(%)			RRMSE(%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default estimator	0.6	0.3	0.2	1.88	1.12	0.59	18.01	13.80	9.19
Predictor	13.2	13.5	12.8	44.63	45.47	45.49	44.95	45.78	45.81
Expanded predictor (17)	0.8	-0.2	1.4	5.63	4.18	6.17	6.25	5.11	6.88
Composite	0.7	0.0	1.0	4.57	3.22	4.27	7.22	5.53	6.14

Design: SRSWOR. Quantitative x: educ-thh, empmohh. Qualitative x: house ownership, lfs-code, socstrat. Domains: 36 NUTS4 regions. Model was fitted to log(income+1).

Tables 8.21-8.23 summarize our experiments with contaminated data under SRSWOR. The expanded predictor and frequency-calibrated predictor have the smallest RRMSE and not too much design bias. Moreover, they are more robust than the default estimator. Composite estimators suffer from bias in contaminated data.

Table 8.24 shows a contamination experiment with PPS. The PPS design seems to result in larger RRMSE of expanded predictor and frequency-calibrated estimator under OCAR-CCAR but other changes are small (compare to Table 8.21).

© http://ameli.surveystatistics.net/ - 2011

Table 8.21: Quintile share in contaminated data (Finnish data set)

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	2.31	1.23	0.57	1.14	18.17	13.77	9.17	12.72
Expanded predictor (18)	2.75	4.47	8.65	5.74	4.06	5.76	9.38	6.83
Composite	2.23	3.48	5.51	4.04	6.03	5.87	7.35	6.43
n-calibrated predictor	5.61	5.00	9.22	6.61	8.67	7.19	10.20	8.48
n-calibrated composite	4.85	3.97	5.78	4.74	8.77	6.75	7.68	7.36
OCAR-CCAR 1%								
Default	11.33	13.96	15.12	14.02	27.92	24.41	19.55	23.14
Expanded predictor (18)	8.52	10.25	4.63	7.98	9.06	10.79	5.99	8.82
Composite	9.09	10.88	7.60	9.45	12.74	13.07	9.84	11.86
n-calibrated predictor	3.67	7.93	4.31	6.03	7.89	9.58	6.23	8.14
n-calibrated composite	5.86	9.03	7.41	8.01	11.69	12.01	9.82	11.17
OCAR-NCAR 1%								
Default	31.91	49.01	80.10	57.87	59.10	70.22	87.79	75.02
Expanded predictor (18)	10.84	12.02	5.31	9.43	11.70	13.10	7.49	10.88
Composite	20.69	32.47	62.62	41.72	32.01	43.09	68.66	50.79
n-calibrated predictor	5.80	9.56	4.76	7.30	9.54	11.52	7.48	9.79
n-calibrated composite	17.85	30.86	62.40	40.44	30.27	41.72	68.46	49.78
OAR-CAR								
Default	35.59	50.82	67.85	54.85	58.37	65.41	71.64	66.68
Expanded predictor (18)	17.08	18.09	9.31	14.78	17.39	18.44	10.21	15.32
Composite	25.90	36.56	55.20	41.81	35.02	43.57	58.21	47.67
n-calibrated predictor	12.27	15.47	8.65	12.56	14.46	16.69	10.00	13.97
n-calibrated composite	23.45	35.40	55.16	40.88	33.33	42.68	58.20	46.98

Design: SRSWOR. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS4. Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Table 8.22: Unit-level quintile share estimators in contaminated data (Amelia).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
Expanded predictor	12.3	8.6	5.7	8.9	16.0	13.6	11.4	13.7
Composite	9.8	7.1	4.7	7.2	16.0	14.6	12.6	14.5
OCAR-CCAR 1%								
Direct	7.9	9.1	10.8	9.2	43.8	41.8	39.3	41.7
Expanded predictor	14.3	8.5	5.7	9.5	18.1	14.2	12.2	14.8
Composite	12.8	8.0	6.9	9.2	18.8	15.9	14.0	16.2
OCAR-NCAR 1%								
Direct	9.1	12.3	16.7	12.6	53.3	53.2	53.2	53.2
Expanded predictor	15.0	8.9	6.6	10.1	18.6	14.5	12.4	15.1
Composite	13.4	9.4	9.3	10.6	21.3	19.3	18.6	19.7

Design: SRSWOR. Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation. Domains: DIS regions. Mixed model with DIS random intercepts was fitted to log(income+1) by nlme.

Table 8.23: Quintile share estimators with aggregated auxiliary data in contaminated data (Amelia).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
n-calibrated predictor	11.1	13.3	10.6	11.9	31.3	29.6	25.9	29.1
n-calibrated composite	8.8	10.8	8.9	9.7	27.9	26.6	23.5	26.1
OCAR-CCAR 1%								
Direct	7.9	9.1	10.8	9.2	43.8	41.8	39.3	41.7
n-calibrated predictor	10.9	10.3	7.0	9.6	30.6	27.7	23.7	27.5
n-calibrated composite	9.0	7.0	4.9	7.0	27.2	24.5	21.1	24.4
OCAR-NCAR 1%								
Direct	9.1	12.3	16.7	12.6	53.3	53.2	53.2	53.2
n-calibrated predictor	11.0	9.6	6.3	9.1	30.3	27.1	23.0	26.9
n-calibrated composite	9.4	6.4	4.3	6.7	28.5	26.0	23.3	26.0

Design: SRSWOR. Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation. Domains: DIS regions. Mixed model with DIS random intercepts was fitted to log(income+1) by nlme.

© http://ameli.surveystatistics.net/ - 2011

Table 8.24: Quintile share in contaminated data under PPS (Finnish data set).

Estimator	ARB(%)				RRMSE(%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	3.13	1.69	0.66	1.52	20.66	15.58	9.86	14.22
Expanded predictor (18)	2.86	4.57	8.71	5.83	3.96	5.71	9.34	6.78
Composite	2.39	3.47	5.69	4.12	7.20	6.22	7.61	6.86
n-calibrated predictor	6.27	5.06	9.07	6.68	9.27	6.88	9.96	8.32
n-calibrated composite	5.43	3.99	5.86	4.86	9.95	6.86	7.87	7.65
OCAR-CCAR 1%								
Default	10.52	13.62	14.99	13.69	29.05	25.30	20.02	23.91
Expanded predictor (18)	7.76	9.61	4.62	7.55	8.39	10.23	6.00	8.45
Composite	8.18	10.20	7.06	8.78	13.01	12.68	9.69	11.65
n-calibrated predictor	3.51	7.49	4.56	5.88	7.99	9.00	6.30	7.88
n-calibrated composite	4.65	8.46	6.94	7.38	12.19	11.68	9.71	11.04
OCAR-NCAR 1%								
Default	29.55	49.00	80.02	57.50	58.55	70.38	87.64	74.97
Expanded predictor (18)	16.80	17.99	9.87	14.89	19.85	21.12	15.06	18.75
Composite	22.38	34.98	64.44	43.87	35.20	45.60	70.37	53.10
n-calibrated predictor	10.81	15.20	9.47	12.52	16.27	19.05	14.94	17.18
n-calibrated composite	19.52	33.57	64.40	42.75	33.24	44.39	70.35	52.22
OAR-CAR								
Default	27.28	42.44	63.48	47.93	53.99	61.18	69.71	63.26
Expanded predictor (18)	16.15	17.04	8.50	13.83	16.52	17.48	9.51	14.47
Composite	21.04	30.21	48.92	35.69	31.39	38.28	53.48	42.81
n-calibrated predictor	10.43	14.45	8.05	11.58	13.59	15.69	9.46	13.15
n-calibrated composite	18.05	28.79	48.83	34.54	29.44	37.07	53.43	41.92

Design: PPS by socio-economic status. Qualitative x-variables: age and gender with interactions, lfs-code and socstrat. Domains: NUTS4. Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme without using the design weights.

8.6 Recommendations

In general, results are not improved by adding domain-specific terms to the used model. We obtained better estimates by including terms such as random intercepts associated with NUTS3 levels when domains were defined by NUTS4, for example.

In estimation of the poverty rate, logistic mixed models are at least theoretically preferable to fixed effects models as they describe differences between domains parsimoniously. Of all the poverty rate estimators, EBP might be the best choice unless it is important to avoid design bias. Our findings are similar to the conclusions of Fabrizi et al. (2007) and Judkins and Liu (2000).

Ordinary predictors are substantially biased: poverty gaps and Gini coefficients were too small and quintile shares were too large. The expanded predictors of quintile share and Gini predictors had much smaller RRMSE. They were also more robust than the default method or the ordinary predictor. In small domains, the expanded predictor was nearly always better than the default estimator. In the largest domains, the default estimator may be preferred to the expanded predictor unless there are outliers. In contaminated data the expanded predictors of quintile share and Gini coefficient appear to be better than the default estimator in all domain size classes.

In poverty gap estimation, the expansion technique does not seem to work as well as in the case of quintile share and Gini coefficient. We might prefer composite poverty gap estimators over predictors. As only modest improvements were obtained with elaborate techniques, the default poverty gap estimator appears good enough.

The frequency-calibrated estimators (Eqs. 18 and 19) have similar robustness properties as the expanded predictor. However, in the case of the poverty gap, the frequency-calibrated method may perform poorly. The frequency-calibrated estimators should be used only if unit-level population data is not available.

A composite estimator consists of a default estimator and a corresponding expanded predictor. In the case of no contamination, these estimators had smaller bias than the expanded predictors, but RRMSE was usually slightly larger. If contamination yields bias in the default estimator, composite estimators consequently suffer from bias. Composite estimators of quintile share or Gini coefficient may not be a good choice if some contamination is suspected.

© http://ameli.surveystatistics.net/ - 2011

Chapter 9

WP3: Variance Estimation

9.1 Design-based Simulation Study on the Amelia Dataset

Variance estimation in survey sampling is essential for statistical inference. For indicators of poverty and social exclusion estimated from sample survey data it gives much needed information on the accuracy of the estimators. Further, it enables the statistician to construced valid confidence intervals (CI) for the estimated indicators $\hat{\theta}$. The two main problems in this context are

(i) that due to complex survey designs (unequal probability sampling) it is not practicable to estimate $\mathsf{Var}(\hat{\theta})$ directly,

(ii) calculating $\hat{\theta}$ involves the estimation of non-smooth statistics.

The goal of the simulation study in WP3 is to study the properties of different variance estimation techniques, namely approximative variance based on linearization techniques and resampling methods (for details on the relevant methodology see Bruch et al., 2011 for resampling methods and Münnich and Zins, 2011 for linearization methods).

The study includes three one-stage designs, 1.2, 1.4a, and 1.5a, and two two-stage designs, 2.6 and 2.7 (see table 2.1). It is based on 10,000 samples drawn from the Amelia synthetic dataset (see Alfons et al., 2011, chapter 4).

Sampling with equal probabilities within the designs 1.2, 1.4a and 2.7 is done via the `sample` function in `R` (cf. R Development Core Team, 2010). When sampling elements with unequal probabilities, as in design 1.5a and design 2.6, Midzuno sampling (cf. Tillé, 2006, p. 117) is applied using the `midzuno` function from the `simFrame` package (cf. Alfons, 2011; Alfons et al., 2010). For two-stage design 2.6 the `sample` function is used at the first stage to select primary sampling units (PSUs), and the `midzuno` function at the second stage to select secondary sampling units (SSUs) with unequal probabilities.

Following the terminology of the `survey` package (cf. Tillé and Matei, 2011), `survey.design` objects for the five designs can be specified in the following way.

```
#=================================================================
# Specification of survey.design objects
# for designs 1.2, 1.4a, 1.5a, 2.6, and 2.7.
#
#-----------------------------------------------------------------
# X          data.frame, Sample from Amelia on personal level
# N          integer, Number of persons in Amelia
# HID        Amelia variable, Household id
# CIT        Amelia variable, LAU2 id
# NUTS2      Amelia variable, NUTS2 id
# DOU        Amelia variable, Degree of urbanisation
# SIND       Cross classification between NUTS2 and DOU
# HHG        Amelia variable, Household size
#-----------------------------------------------------------
# Each person belongs to one HID only and each HID belongs to
# one CIT only. Further, each CIT belongs to one NUTS2 area
# and one SIND class only.
# With respect to this nested structure the following variables
# have to be included into X for each sampled person.
#-----------------------------------------------------------
# fpc1_I     Number of households in the corresponding NUTS2 area
# pik1_II    Inclusion probability of HID proportional
#            to HHG within the corresponding NUTS2 area
# fpc1_II    Number of CITs in the corresponding SIND class
# fpc2_II    Number of HIDs in the corresponding CIT
# pik1_II    Inclusion probability of CIT in the corresponding
#            SIND class
# pik2_II    Inclusion probability of HID proportional
#            to HHG within the corresponding CIT area

# Relevant R packages
library(survey)

# Design 1.2
d1.2 <- svydesign(id=~HID, fpc=~rep(N,nrow(X)), data=X)
# Design 1.4a
d1.4a <- svydesign(id=~HID, strata=~NUTS2, fpc=~fpc1_I, data=X)
# Design 1.5a
d1.5a <- svydesign(id=~HID, strata=~NUTS2, fpc=~pik1_I, pps="brewer", data=X)
# Design 2.6
d2.6 <- svydesign(id=~CIT+HID, strata=~SIND, fpc=~pik1_II+pik2_II, pps="brewer", data=X)
# Design 2.7
d2.7 <- svydesign(id=~CIT+HID, strata=~SIND, fpc=~fpc1_II+fpc2_II, data=X)
```

In the simulation study, variance estimators for the design-based estimators of the following five indicators are evaluated:

ARPR the At-risk-of-poverty Rate,

RMPG the Relative Median Poverty Gap,

QSR the Quintile Share Ratio,

GINI the Gini coefficient,

MEAN and the Income Mean. The functions used to compute both point and variance estimates for those five estimators are in the Appendix Volume Hulliger et al. (2011).

All indicators are estimated for the equivalised disposable income of each person in the sample. The definition of estimatiors of ARPR, RMPG, QSR and GINI and the equivalised disposable can be found in Eurostat (2009), whereas in this simulation we used

© http://ameli.surveystatistics.net/ - 2011

only the inverse inclusion probabilities as weights. The income mean was estimated via

$$\text{MEAN} = \sum_{i \in s} \frac{y_i \frac{1}{\pi_i}}{\frac{1}{\pi_i}} ,$$

where y_i is the equivalised disposable income of the i-th person and π_i is the inclusion probability of the i-th person into sample s (see Särndal et al., 1992, p. 182).

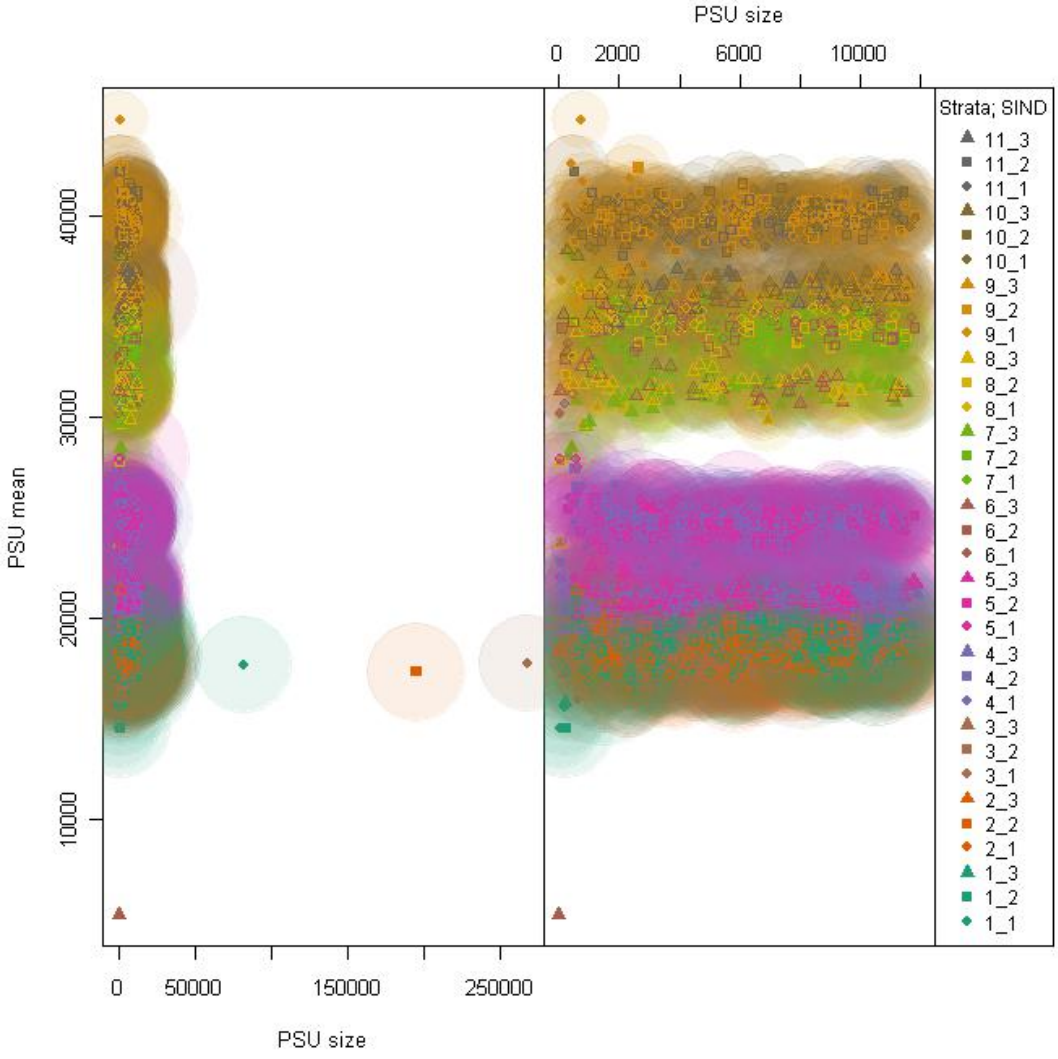


Figure 9.1: Distribution of PSUs for Designs 2.6 and 2.7

Because of their significance for the variance of point estimators from the two-stage samples, figure 9.1 displays the size, the mean and the standard deviation of the equivalised disposable income within each PSU (see also Bruch et al., 2011, chapter 4). The mean is displayed on the y-axis, the size on the x-axis and the radii of the circels correspond to the standard diviations. The left side of the graph includes all PSUs, whereas on the right side the three PSUs with the biggest size are not shown to have a more balanced overview. Different colours and plotting symbols indicate together the membership of a PSU to a certain stratum. The three plotting symbols correspond to the three denominations of DOU and the colours to the eleven categories of NUTS2. The figure shows that

the size of the PSUs, appart from the three biggest ones, seems to follow an uniform distribution over the range of 2 to 12,000 households. Further, it is visible from the values of the means and standard deviations that there are groups of two to three strata which are largely overlapping, whereas others appear to be separated. This is due to the grouping of the means and, compared to the range of the means, a moderate variance within PSUs.

9.2 Variance Approximation

Linearization

Here, the variance of an estimator $\hat{\theta}$ is approximated by the variance of the estimated total of the linearized or influence values $z_i \; \forall \; i \in s$, so that

$$\mathsf{Var}\left(\sum_{i\in s}\frac{z_i}{\pi_i}\right) \approx \mathsf{Var}(\hat{\theta}) \; . \tag{9.1}$$

The z_i corresponding to estimators ARPR, RMPG, QSR, GINI are given in Münnich and Zins (2011). For estimator MEAN z_i is given by $z_i = (y_i - \text{MEAN})\hat{N}^{-1}$, where $\hat{N} = \sum_{i\in s}\pi_i^{-1}$.

For the equal probability designs 1.2, 1.4a and 2.7 following standard variance estimators for the estimated total $Z = \sum_{i\in s}\frac{z_i}{\pi_i}$ are applied.

$$\widehat{\mathsf{Var}}_{1.2}\left(Z\right) = N^2\left(\frac{N-n}{N}\right)\frac{S^2}{n} \tag{9.2a}$$

$$\widehat{\mathsf{Var}}_{1.4a}\left(Z\right) = \sum_{h=1}^{H} N_h^2\left(\frac{N_h - n_h}{N_h}\right)\frac{S_h^2}{n_h} \tag{9.2b}$$

$$\widehat{\mathsf{Var}}_{2.7}\left(Z\right) = \sum_{h=1}^{H} N_h^{'2}\left(\frac{N_h^{'} - n_h^{'}}{N_h^{'}}\right)\frac{S_{he}^2}{n_h^{'}} + \sum_{h=1}^{H}\frac{N_h^{'}}{n_h^{'}}\sum_{q=1}^{n_h^{'}}\widehat{\mathsf{Var}}\left(Z_{hq}\right) \; , \tag{9.2c}$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \frac{\sum_{i=1}^{n} z_i}{n}\right)^2 \; ,$$

$$S_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(z_{hi} - \frac{\sum_{i=1}^{n_h} z_{hi}}{n_h}\right)^2 \; ,$$

$$S_{he}^2 = \frac{1}{n_h^{'} - 1}\sum_{q=1}^{n_h^{'}}\left(Z_{hq} - \frac{Z_h}{N_h^{'}}\right)^2 \; ,$$

$$\widehat{\mathsf{Var}}\left(Z_{hq}\right) = \frac{N_h^{''2}}{n_h^{''}}\left(\frac{N_h^{''} - n_h^{''}}{N_h^{''}}\right)\frac{1}{n_{hq}^{''} - 1}\sum_{i=1}^{n_{hq}^{''}}\left(z_{hqi} - \frac{\sum_{i=1}^{n_{hq}^{''}} z_{hqi}}{n_{hq}^{''}}\right)^2 \; ,$$

with $Z_{hq} = \frac{N_{hq}^{''}}{n_{hq}^{''}}\sum_{i=1}^{n_{hq}^{''}} z_{hqi}$ and $Z_h = \frac{N_h^{'}}{n_h^{'}}\sum_{q=1}^{n_h^{'}} Z_{hq}$ (see Chapter 1 in Bruch et al. (2011)). In equations 9.2b and 9.2c H refers to the number of different classes of NUTS2 and SIND, respectively. Tables 9.1 and 9.2 contain the relative bias and the coverage probability of CI (see section 6.2), corresponding to these variance estimators.

© http://ameli.surveystatistics.net/ - 2011

Table 9.1: Relative Bias (in %) of Direct Variance Estimators for Designs 1.2, 1.4a, and 2.7

	1.2	1.4a	2.7
ARPR	$-3.0305e-01*$	$-1.8197e+00$	$-1.5225e+01$
RMPG	$-2.7191e+00$	$-7.3391e-01*$	$-1.0303e+01$
QSR	$-9.2006e-01$	$9.0976e-01*$	$-1.1714e+01$
GINI	$-1.7382e+00$	$6.6028e-01*$	$-1.8795e+01$
MEAN	$-1.1183e+00*$	$1.2522e+00$	$-1.7069e+01$

stars (i.e., *) denote the closest value to zero in each row

Table 9.2: Coverage Probability of CI constructed with Direct Variance Estimators for Designs 1.2, 1.4a, and 2.7

	1.2	1.4a	2.7
ARPR	$9.5070e+01$	$9.4700e+01$	$9.0640e+01$
RMPG	$9.4640e+01$	$9.4790e+01$	$9.2650e+01$
QSR	$9.4620e+01$	$9.5260e+01$	$8.3690e+01$
GINI	$9.4440e+01$	$9.5090e+01$	$8.5550e+01$
MEAN	$9.4850e+01$	$9.5070e+01$	$7.9960e+01$

9.3 Approximation of Design Variance

For designs 1.5a and 2.6 the application of a direct estimator for 9.1 of kind (2.2) or (2.4) in Bruch et al. (2011) was, although theoretically possible, not workable because of excessive long computational times, (for the computation of second-order inclusion probabilities of Midzuno sampling see Tillé, 2006, p. 117). Thus, approximations for design variances are used, not only because of their computational simplicity but also because of their relevance in practice. A general approximation of variance 9.1 is given by

$$\widehat{\mathsf{Var}}_{\text{approx}}\left(\hat{\theta}\right) = \sum_{i \in s} \frac{\hat{b}_i}{\pi_i^2} \cdot e_i^2 \,, \tag{9.3}$$

with e_i^2 as defined in formula (2.32) in Bruch et al. (2011).

Two broad classes of variance estimators are considered: **type 1 estimators** that require knowledge about the first-order inclusion probabilities for the sampled elements only, and **type 2 estimators** requiring knowledge about the first-order inclusion probabilities for all the elements in the universe.

The first two type 1 estimators, *deville1* and *deville2*, are special cases of the variance estimator 9.3, also described by Matei and Tillé (2005). They differ only in terms of the value chosen for $\hat{b}_i$. Obtaining the value of $\hat{b}_i$ by means of fixed-point approximation yields the *fix* estimator (see Bruch et al., 2011). Two further type 1 estimators, *rose* and *deville3*, are given in Matei and Tillé (2005). The *brewer2* estimator, which is the only estimator of the Brewer family on sample length, completes the list of type 1 estimators considered in the study (see Matei and Tillé (2005)).

The estimators *berger*, *brewer1*, *brewer3*, *brewer4*, *approx2*, *approx3*, *approx4* belong to the category of type 2 estimators. Matei and Tillé (2005) proposed the five variance estimators to which we will refer to as *approx2* to *approx4* (for the *brewer* estimators see also Bruch et al., 2011).

Design 1.5a

The results of our simulation study for design 1.5a with respect to the criterion relative bias of the variance estimators *relbiasV* are given in table 9.3. Although estimator *approx4* has the smallest bias the differences between the variance estimators are negligible. Furthermore, noteworthy discrepancies between type 1 estimators on sample length and type 2 estimators on universe length cannot be detected.

Table 9.3: Relative Bias (in %) of Direct Variance Estimator for Design 1.5a

	approx2	approx3	approx4		
ARPR	$4.4605e-01$	$4.4605e-01$	$4.4597e-01*$		
RMPG	$-2.4094e+00$	$-2.4094e+00$	$-2.4095e+00$		
QSR	$1.3490e+00$	$1.3490e+00$	$1.3489e+00*$		
GINI	$7.5450e-01$	$7.5450e-01$	$7.5441e-01*$		
MEAN	$3.5557e+00*$	$3.5557e+00*$	$3.5557e+00*$		
	berger	brewer1	brewer2	brewer3	brewer4
ARPR	$4.4621e-01$	$4.4604e-01$	$4.4605e-01$	$4.4607e-01$	$4.4607e-01$
RMPG	$-2.4092e+00*$	$-2.4095e+00$	$-2.4094e+00$	$-2.4093e+00$	$-2.4093e+00$
QSR	$1.3492e+00$	$1.3489e+00*$	$1.3490e+00$	$1.3491e+00$	$1.3491e+00$
GINI	$7.5469e-01$	$7.5444e-01$	$7.5450e-01$	$7.5456e-01$	$7.5456e-01$
MEAN	$3.5559e+00$	$3.5557e+00*$	$3.5557e+00*$	$3.5558e+00$	$3.5558e+00$
	deville1	deville2	deville3	fix	rose
ARPR	$4.4605e-01$	$4.4605e-01$	$4.4605e-01$	$4.4606e-01$	$4.4605e-01$
RMPG	$-2.4094e+00$	$-2.4094e+00$	$-2.4094e+00$	$-2.4093e+00$	$-2.4094e+00$
QSR	$1.3490e+00$	$1.3490e+00$	$1.3490e+00$	$1.3491e+00$	$1.3490e+00$
GINI	$7.5450e-01$	$7.5450e-01$	$7.5450e-01$	$7.5455e-01$	$7.5450e-01$
MEAN	$3.5557e+00*$	$3.5557e+00*$	$3.5557e+00*$	$3.5558e+00$	$3.5557e+00*$

stars (i.e., *) denote the closest value to zero in each row

© http://ameli.surveystatistics.net/ - 2011

Design 2.6

The variance of Z estimated with the π_i's from design 2.6 can by the following expression:

$$\widehat{\mathsf{Var}}_{2.6}(Z) = \sum_{h=1}^{H} N_h^{'2} \cdot \left(\frac{N_h^{'} - n_h^{'}}{N_h^{'}} \right) \frac{S_{he}^2}{n_h^{'}} + \sum_{h=1}^{H} \frac{N_h^{'}}{n_h^{'}} \sum_{q=1}^{n_h^{'}} \widehat{\mathsf{Var}}(Z_{hq}) \ , \tag{9.4}$$

where

$$\widehat{\mathsf{Var}}(Z_{hq}) = \sum_{i=1}^{n_{hq}^{''}} \sum_{j=1}^{n_{hq}^{''}} \frac{\pi_{i\,j|hq} - \pi_{i|hq}\pi_{j|hq}}{\pi_{i\,j|hq}} \frac{z_{hqi}}{\pi_{i|hq}} \frac{z_{hqj}}{\pi_{j|hq}} \ . \tag{9.5}$$

The variance esitmator in 9.5 is approximated by one of the estimators mentioned in section 9.3. The results of the simulations regarding design 2.6 are given in table 9.4. A comparison of the different approximation methods reveals hardly any differences. The reason for the indifference between the approximation estimator is due to the fact that the variance attribute to the first stage of sampling is the dominant part of 9.4. This is a result of the huge difference of the means beween the PSUs and the strong varying PSU sizes, as displayed 9.1. Thus, the contribution of 9.5 to the variance of Z is negligible. Table 9.5 contains the coverage probability of CI constructed with using the *approx4* variance estimator for designs 1.5a and 2.6.

Table 9.4: Relative Bias (in %) of Direct Variance Estimates for Designs 2.6

	approx2	approx3	approx4		
ARPR	$-1.7307e+01$	$-1.7307e+01$	$-1.7308e+01$		
RMPG	$-9.3250e+00$	$-9.3250e+00$	$-9.3268e+00$		
QSR	$-1.1956e+01$	$-1.1956e+01$	$-1.1956e+01$		
GINI	$-2.0504e+01$	$-2.0504e+01$	$-2.0505e+01$		
MEAN	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$		
	berger	brewer1	brewer2	brewer3	brewer4
ARPR	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$
RMPG	$-9.3242e+00$	$-9.3262e+00$	$-9.3250e+00$	$-9.3239e+00$	$-9.3238e+00$
QSR	$-1.1955e+01$	$-1.1956e+01$	$-1.1956e+01$	$-1.1955e+01$	$-1.1955e+01$
GINI	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$
MEAN	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$
	deville1	deville2	deville3	fix	rose
ARPR	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$	$-1.7307e+01$
RMPG	$-9.3251e+00$	$-9.3250e+00$	$-9.3250e+00$	$-9.3240e+00$	$-9.3251e+00$
QSR	$-1.1956e+01$	$-1.1956e+01$	$-1.1956e+01$	$-1.1955e+01$	$-1.1956e+01$
GINI	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$	$-2.0504e+01$
MEAN	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$	$-1.7330e+01$

stars (i.e., *) denote the closest value to zero in each row

Table 9.5: Coverage Probability of CI constructed with Direct Variance Estimators for Designs 1.5a and 2.6

	1.5a	2.6
ARPR	$9.4950e+01$	$8.9340e+01$
RMPG	$9.4550e+01$	$9.2930e+01$
QSR	$9.4850e+01$	$8.3880e+01$
GINI	$9.5140e+01$	$8.4230e+01$
MEAN	$9.5320e+01$	$7.8720e+01$

9.4 Bootstrap

The boostrap method used in the simulation study is a Monte Carlo Bootstrap with 100 replications. The method was implemented by using the `boot` function of the `boot` package (see Canty and Ripley, 2010; Davison and Hinkley, 1997). In case of the two-stage sampling designs 2.6 and 2.7 only the first stage of sampling is considered, because, as mentioned above, it is the principal component of the variance. The relative bias and the coverage probability of CI for bootstrap variance estimators are given in tables 9.6 and 9.7.

Table 9.6: Relative Bias (in %) of Boostrap Variance Estimators for Designs 1.2, 1.4a, 1.5a, 2.6 and 2.7

	1.2	1.4a	1.5a	2.6	2.7
ARPR	$3.4170e+00$	$1.7062e+00*$	$3.0425e+00$	$-2.3091e+01$	$-3.1199e+00$
RMPG	$4.9318e+00$	$6.9235e+00$	$3.6407e+00*$	$-9.1982e+00$	$1.3912e+01$
QSR	$3.7710e-01*$	$2.9131e+00$	$1.9858e+00$	$-1.8887e+01$	$4.4255e+00$
GINI	$-1.5121e+00$	$1.4993e+00$	$4.3150e-01*$	$-2.6783e+01$	$-7.9915e+00$
MEAN	$-1.0235e+00*$	$4.9193e+00$	$3.7298e+00$	$-2.2048e+01$	$1.5894e+00$

stars (i.e., *) denote the closest value to zero in each row

Table 9.7: Coverage Probability of CI constructed with Bootstrap Variance Estimators for Designs 1.2, 1.4a, 1.5a, 2.6 and 2.7

	1.2	1.4a	1.5a	2.6	2.7
ARPR	$9.5100e+01$	$9.4910e+01$	$9.4810e+01$	$8.7850e+01$	$9.3070e+01$
RMPG	$9.4410e+01$	$9.4750e+01$	$9.4600e+01$	$9.2390e+01$	$9.4940e+01$
QSR	$9.4280e+01$	$9.5180e+01$	$9.4220e+01$	$8.2210e+01$	$8.8260e+01$
GINI	$9.4240e+01$	$9.4770e+01$	$9.4660e+01$	$8.1890e+01$	$9.0070e+01$
MEAN	$9.4620e+01$	$9.5260e+01$	$9.5090e+01$	$7.7630e+01$	$9.0340e+01$

9.5 Balanced Repeated Replication

The Group Balanced Repeated Replication (BRR) method is also tested, whereas the grouping is repeated as described in Shao and Rao (1993, p. 344). For the BRR, replicate weights are computed by the formula (5.2) in Davison and Sardy (2004, p. 18).

© http://ameli.surveystatistics.net/ - 2011

Like the bootstrap, the BRR is only applied at the first sampling stage. To select the PSU a Hadamard matrix is used, which is implemented in `R` with the `hadamard` function (see package, Lumley, 2010). The BRR is not used for design 1.2 due to the missing stratification. The relative bias and the coverage probability of CI for this approach are given in tables 9.8 and 9.9, respectively.

Table 9.8: Relative Bias (in %) of BRR Variance Estimators for Designs 1.4a, 1.5a, 2.6 and 2.7

	1.4a	1.5a	2.6	2.7
ARPR	$3.0199e+00*$	$4.2885e+00$	$-1.6374e+01$	$-1.4306e+01$
RMPG	$8.2757e+00$	$5.4446e+00*$	$-8.3269e+00$	$-6.5820e+00$
QSR	$1.8593e+00*$	$2.4236e+00$	$-1.4850e+01$	$-1.4224e+01$
GINI	$8.4603e-01*$	$1.0440e+00$	$-2.0591e+01$	$-1.9099e+01$
MEAN	$1.1381e+00*$	$3.4923e+00$	$-1.5421e+01$	$-1.5416e+01$

stars (i.e., *) denote the closest value to zero in each row

Table 9.9: Coverage Probability of CI constructed with BRR Variance Estimators for Designs 1.4a, 1.5a, 2.6 and 2.7

	1.4a	1.5a	2.6	2.7
ARPR	$9.5080e+01$	$9.5190e+01$	$8.8240e+01$	$8.9430e+01$
RMPG	$9.4880e+01$	$9.4850e+01$	$9.2390e+01$	$9.2250e+01$
QSR	$9.5110e+01$	$9.4670e+01$	$8.1440e+01$	$8.1250e+01$
GINI	$9.4870e+01$	$9.4910e+01$	$8.1650e+01$	$8.3070e+01$
MEAN	$9.4910e+01$	$9.5130e+01$	$7.6710e+01$	$7.8240e+01$

9.6 Comparison of the different variance estimation methods

For a better comparison of the different variance estimators figure 9.2 shows the distribution of the relative bias for each variance estimate with boxplots. Boxplots labeled *naive* relate to the estimators in equations 9.2a, 9.2b, and 9.2c, the ones labeled as *approx4* refer to the approximation estimator used for estimating the variance for designs 1.5a and 2.6. The resampling methods are labeled with *boot* and *rbrr* for the Bootstrap and Balanced Repeated Replication, respectively. The boxplots labeled *naive* for 1.5a and 2.6 related to estimators 9.2b and 9.2c, i.e. in this case designs 1.5a and 2.6 are treated as being drawn with equal probability.

For one stage designs (1.2, 1.4a and 1.5a) all are variance estimators are almost unbiased. In general, the direct estimators, *approx4* and *naive*, show a greater efficiency as the resampling methods. It is noteworthy that using the naive variance estimator for design 1.5a results also to a comparable small relative bias. For instance, in cases of the ARPR the relative bias using the naive variance estimator is only slightly higher `0.004825974` than using the best approximation method *approx4*, which has a relative bias of `0.004459717`. For other indicators the difference is even lower. Another notable observation is that the

bootstrap and the BRR estimators give also reasonable results for design 1.5a although this methods do not explicitly account for unequal probability sampling.

Finally it needs to by mentioned that for the two-stage designs all variance estimators, except the bootstrap for design 2.7, underestimated the variance. In general they have negative relative bias well above 10%. In case of the variance estimators based on linearization it is assumed that there might be a problem of convergence. Casually speaking for the linearization method to work we need that certain remainder term R_n to converge in probability to zero as the sample size increases to infinity (see section 2.1 Münnich and Zins, 2011). A sufficient condition for this to hold is met in case of iid observations y_i, but for estimator from general survey designs this is not easy to prove (see Demnati and Roa, 2004). Serfling (1980) makes some suggestions on how to analysis R_n in practice.

Figures 9.3 and 9.4 display the kernel density estimators for the distribution of the point and variance estimates, respectively. In figure 9.3 it can be seen that almost all point estimators are normal distributed. Exceptions are the non-robust statistics QSR, GINI and MEAN in case of the two-stage sampling designs 2.6 and 2.7. This shape of the distributions of point estimates and the underestimations of the variance estimators explain the low coverage probability of CI in case of the two-stage designs, which can be found in the tables 9.2, 9.5, 9.7, and 9.9.

The kernel density estimation of the different variance estimation methods in figure 9.4 shows that the distribution of the *naive* approach and the *approx4* method for the designs with only one stage are almost normal distributed for all indicators. In comparison to these methods the distribution of the resampling methods for the ARPR and the RMPG is very flat and indicates a high variance for this variance estimators.

For the two stage designs has the empirical distribution of most variance estimators a heavy positive skewness. This reflects also the problem with the variance estimation for these design.

© http://ameli.surveystatistics.net/ - 2011

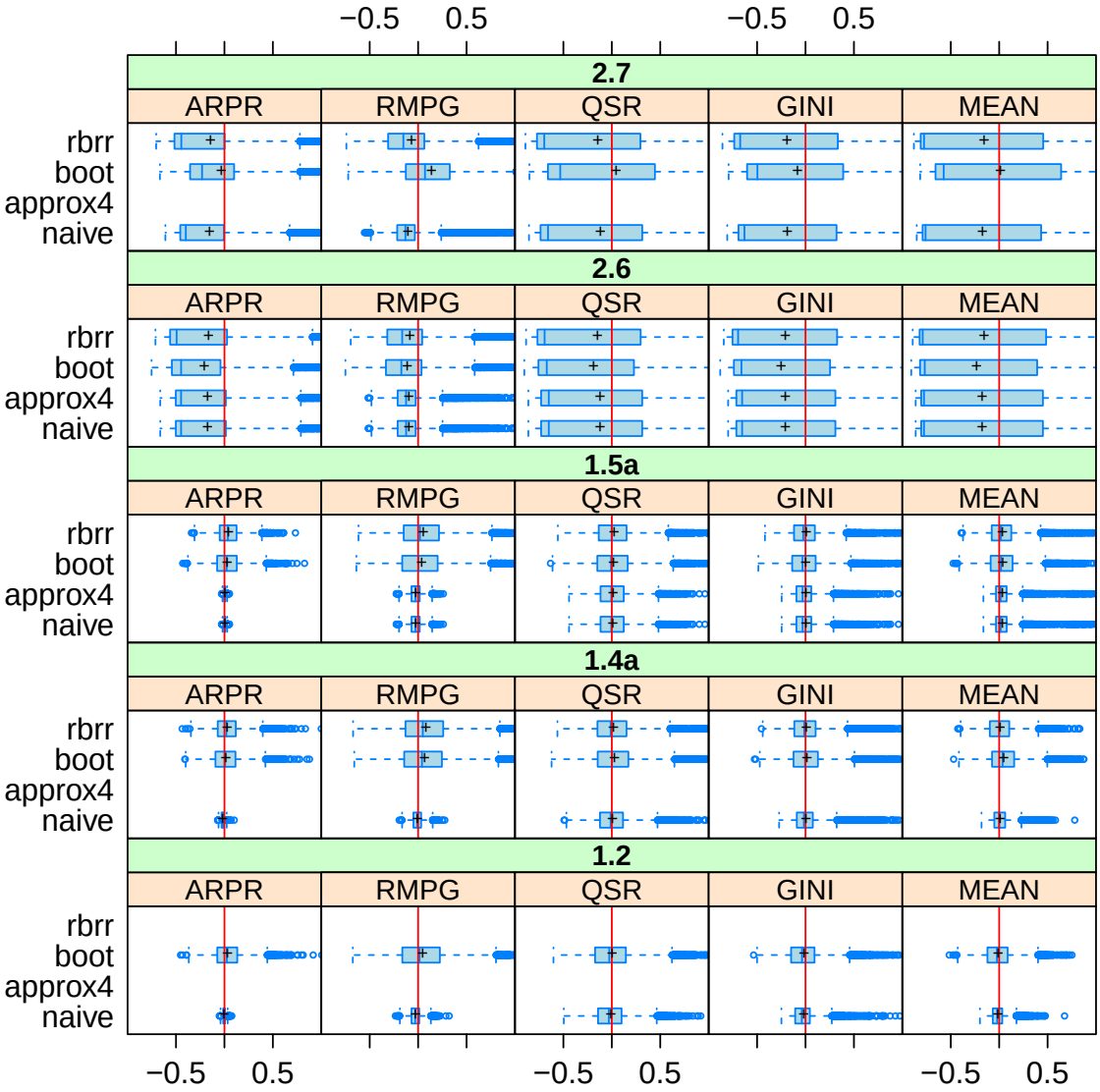


Figure 9.2: Relative Bias of Variance Estimates

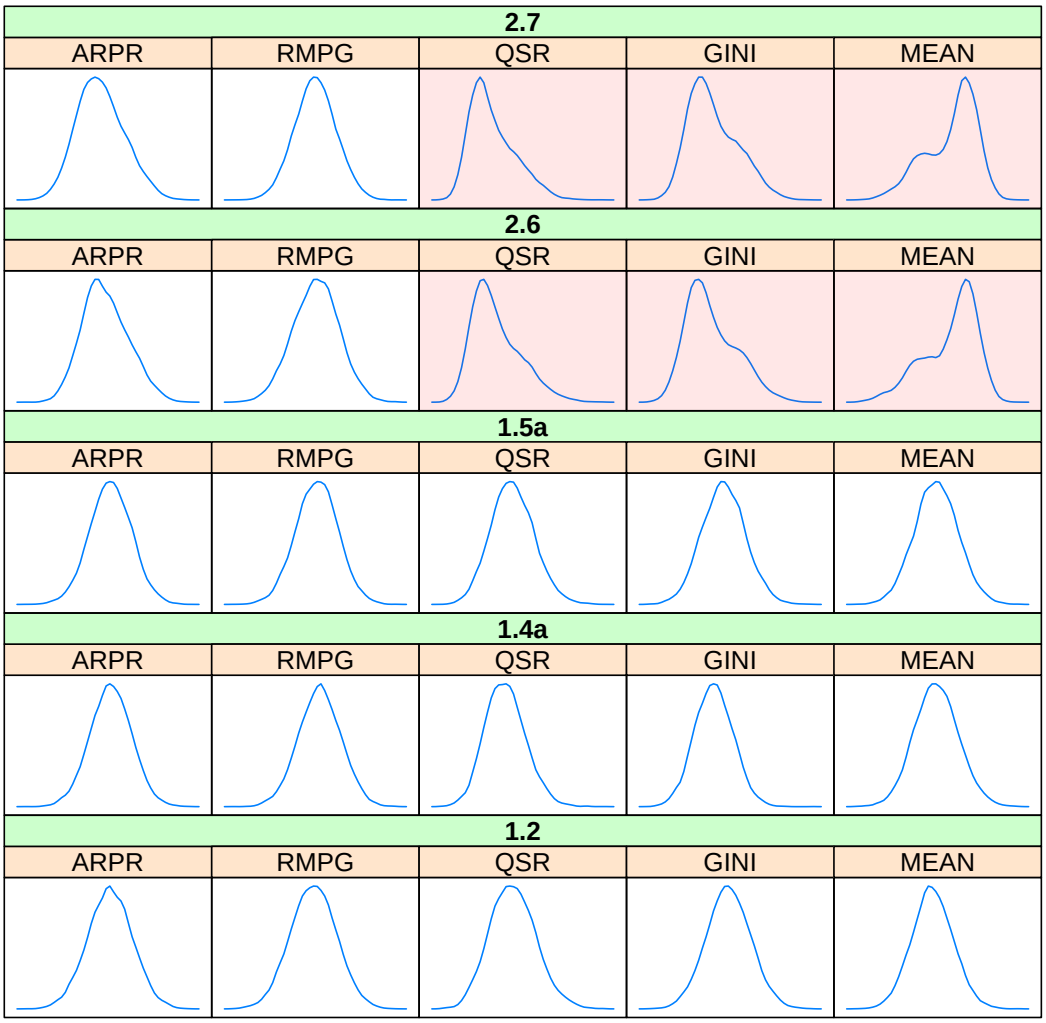


Figure 9.3: Density Estimation of Point Estimates

© http://ameli.surveystatistics.net/ - 2011

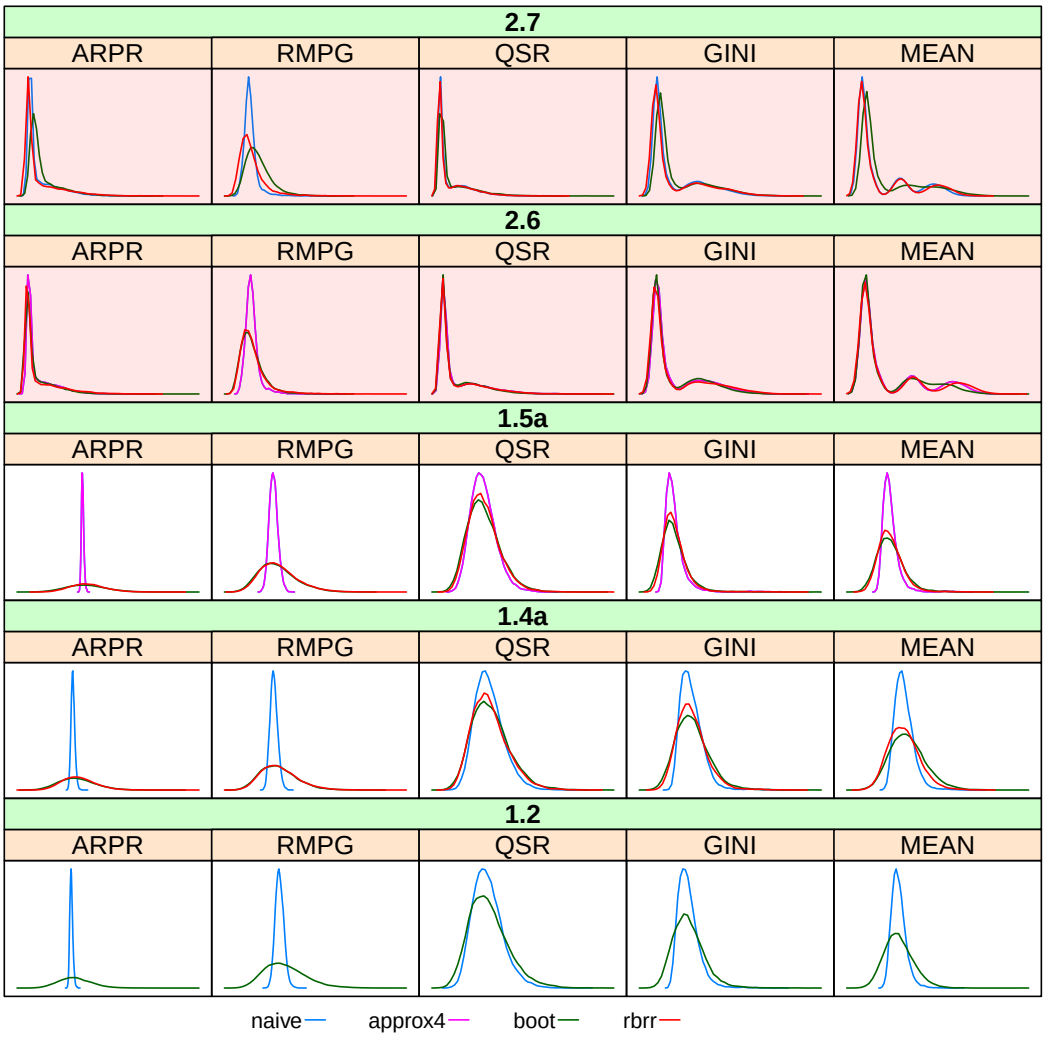


Figure 9.4: Density Estimation of Variance Estimates

9.7 Recommendations

In practice there is often a desire of accurate but also simple variance estimators. Therefore in the simulation study different techniques have been examined that simplify variance estimation under complex survey designs and statistics. To cope with complex statistics linearization and resampling methods have been studied.

In general we endorse the usage of linearization over that of resampling, because linearization is less computationally intensive and usually more efficient. Further, the linearizations for most indicators of poverty and income inequality are well known, which allows for the direct usage of variance estimation software that accommodates to many sampling designs. This is in contrast to resampling methods that are often not applicable to complex sampling designs.

However, because the results of the simulation study on the performance of the linearization methods are mixed, we can not recommend linearization without reservation. The simulation has shown that variance estimates based on the linearization technique perform very well for the one-stage design but are negatively biased for the two-stage designs. Because these methods estimate only the asymptotic variance of an estimator there may be problems of convergence. We presume that this might be caused by the highly skewed distribution of sampling elements in case of the two-stage samples.

Another issue considered was the simplicity of the variance estimator with regard to the sampling design. First, there are variance estimators that use approximations of the second-order inclusion probabilities in case of unequal probability sampling. Here we recommend the usage of the type 1 estimators, that require knowledge about the first-order inclusion probabilities for the sampled elements only. These estimators require less information than the type 2 estimators, where inclusion probabilities have to be known for each element in the sampling frame, but also they are in general less computationally intensive and the simulations have shown that they are almost as accurate as the type 2 estimators. Second, the resampling methods we used in our simulation only considered the first stage of the sampling process. This omission of the following stages is justified under certain conditions like small sampling fractions or homogeneous units in several primary sampling units. For resampling methods we refer also to Table 3.1 in Bruch et al. (2011), which contains an overview of the different resampling methods with regard to there applicability to some characteristics of (complex) sample surveys.

Bibliography

Alfons, A. (**2011**): `simFrame`: Simulation Framework. `R` package version 0.4.1.
URL http://CRAN.R-project.org/package=simFrame

Alfons, A., **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.** and **Templ, M.** (**2011**): *Synthetic Data Generation of SILC Data*. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Alfons, A., **Templ, M.** and **Filzmoser, P.** (**2010**): *An Object-Oriented Framework*

for Statistical Simulation: The R Package simFrame. Journal of Statistical Software, 37 (3), pp. 1–36, to appear.

Bruch, C., **Münnich, R.** and **Zins, S.** (**2011**): *Variance Estimation for Complex Surveys*. Research Project Report WP3 – D3.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Canty, A. and **Ripley, B.** (**2010**): `boot`: Bootstrap R (S-Plus) Functions. R package version 1.2-43.
URL http://CRAN.R-project.org/package=boot

Davison, A. C. and **Hinkley, D. V.** (**1997**): Bootstrap mehods and their application. Cambridge: Cambridge University Press.

Davison, A. C. and **Sardy, S.** (**2004**): *Resampling Methods for Variance Estimation*. Technical report, DACSEIS deliverable D5.1, http://www.dacseis.de.

Demnati, A. and **Roa, J.** (**2004**): *Linearization Variance Estimators for Survey Data*. Survey Methodology, 30 (1), pp. 17 – 26.

Eurostat (**2009**): *Algorithms to compute Overarching Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC)*. Technical report, Eurostat Doc LC-ILC/11/08/EN – Rev. 2.
URL http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/documents/tab/Tab/LC_ILC%2011-08%20Rev2%20Overarching%20indicators%20methodology%20Jun.pdf

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Lumley, T. (**2010**): `survey`: analysis of complex survey samples. R package version 3.23-3.
URL http://CRAN.R-project.org/package=survey

Matei, A. and **Tillé, Y.** (**2005**): *Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size*. Journal of Official Statistics, 21 (4), pp. 543 – 570.

Münnich, R. and **Zins, S.** (**2011**): *Variance Estimation for Indicators of Poverty and Social Exclusion*. Research Project Report WP3 – D3.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

R Development Core Team (**2010**): `R`: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org/

Serfling, R. J. (**1980**): Approximation theorems of mathematical statistics. New York: Wiley.

Shao, J. and **Rao, J. N. K.** (**1993**): *Standard errors for low income proportions estimated from stratified multi-stage samples*. Sankhyā. The Indian Journal of Statistics. Series B, 55 (3), pp. 393–414.

Särndal, C.-E., **Swensson, B.** and **Wretman, J.** (**1992**): Model Assisted Survey Sampling. New York: Springer-Verlag.

Tillé, Y. (**2006**): Sampling algorithms. Springer Series in Statistics, New York: Springer.

Tillé, Y. and **Matei, A.** (**2011**): `sampling`: Survey Sampling. R package version 2.4. URL http://CRAN.R-project.org/package=sampling

© http://ameli.surveystatistics.net/ - 2011

Chapter 10

WP4: Robustness

10.1 Introduction

Social inclusion indicators which are based on income data must take care of outliers in the data. The robust methods developed in the AMELI project are described in Deliverable D4.2 (Hulliger et al., 2011b). The simulations with robust methods have the following objectives

1. Study efficiency and bias of the robust estimators without contamination and with various settings of contamination. Efficiency and bias must be compared with the standard robust estimators.
2. Study the effect of tuning constants in order to give practical advice on their use.
3. Study the effect of the sample design on the robust estimators.
4. Study the quality of the variance estimators for the cases where variance estimators are possible.
5. For the multivariate methods, study the impact of missing values in the components on robust procedures.

The simulations use the AAT-SILC and AMELIA universe with various outlier and contamination settings (see 4.2.4). In addition, for the multivariate procedures, missing values were generated with the mechanisms described in Section 5.1. Most of the simulations were run with the one stage designs but also two stage designs were used (see 2).

10.2 Robust estimation of means

We present some results on simulations for estimating the population mean. This is useful to understand the behaviour of robust estimators and to compare the results with previous work on robust estimation.

We use the AAT-SILC population and simple random sampling only. The estimators are the one-sided robustified Horvitz-Thompson estimator with a asymmetric Huber ψ-function and the one-sided Trimmed Mean. Table 10.1 shows the performance in relative bias and relative mean squared error (both expressed as a percentage of the uncontaminated value, cf. Section 6.2). We chose a standard tuning constant for both estimators and the optimal tuning constant in terms of relative mean squared error. The standard tuning constant for RHT is $k = 6$ (measured in normalized mads, i.e. in mad*1.4826). This tuning constant yields approximately 0.5% of observations which are downweighted and it thus is comparable to the standard trimming proportion used for the trimmed mean, i.e. $\alpha_u = 0.005$. Note that it is very difficult to discuss different robust estimators that downweight a very different number of observations because it then is not clear, whether differences are due to the type of estimator or due to the choice of the tuning constants. Note that the tuning constants have been chosen on grids such that the actually evaluated tuning constant for RHT is $k = 6.2$ and the optimal tuning constant might be somewhat smaller still.

Table 10.1: Relative bias and root mean squared error (in %) of RHT and TM at AAT-SILC with simple random samples ($k = 6.2, \alpha_u = 0.005$).

			OCAR				OAR			
			CCAR		NCAR		CCAR		NCAR	
T	$\epsilon =\rightarrow$	0	0.001	0.01	0.001	0.01	0.001	0.01	0.001	0.01
HT	relbias	0.00	3.62	34.70	1.57	15.04	3.52	33.29	1.64	15.46
HT	relmse	0.74	3.70	34.71	1.76	15.08	3.60	33.30	1.81	15.49
RHT	relbias	-0.23	0.72	9.94	0.61	8.24	0.66	9.08	0.60	7.87
RHT	relmse	0.77	1.03	9.98	0.96	8.29	0.99	9.13	0.95	7.91
opt k		7.7	4.3	2	4.3	2	4.3	2	4.3	2
RHT	relbias	0.00	-0.21	-0.61	-0.25	-0.71	-0.26	-1.00	-0.28	-1.05
RHT	relmse	0.74	0.75	0.94	0.76	1.01	0.77	1.23	0.78	1.27
TM	relbias	-1.75	-0.73	25.62	-0.90	9.29	-0.79	23.87	-0.90	9.34
TM	relmse	1.89	1.04	25.64	1.16	9.33	1.08	23.89	1.16	9.38
opt α		0	0.05	0.031	0.003	0.026	0.003	0.028	0.003	0.026
TM	relbias	0.00	-0.73	0.11	0.00	0.15	0.63	0.43	0.02	-0.22
TM	relmse	0.74	1.04	0.84	0.74	0.78	0.98	0.96	0.75	0.79

Bias is an important driver for a large relrmse of the estimators. The Horvitz-Thompson estimator, which is the arithmetic mean in our case of simple random sampling of households, with full enumeration of household members, as expected has practically no bias it there is no contamination. The relrmse is therefore just the standard deviation, which is 0.74%. This is not astonishing with the large sample size. A light contamination with outlyingness completely at random (OCAR) and contamination completely at random (CCAR) with outlier rate 0.1% ($\epsilon = 0.001$) introduces a bias and mean squared error of 3.62%! In other words about 16 observations in our samples of about 16 000 have a 36-fold effect. If the outlier rate is 1% ($\epsilon = 0.01$) the mean estimator is biased by more than 30%. In other words a contamination of 160 observations yields a mean of 16 000 observations practically useless. This shows the inherent non-robustness of the classical non-robust estimators. The different scenarios of contamination do not differ by much

© http://ameli.surveystatistics.net/ - 2011

though the impact of the NCAR contamination is less heavy than CCAR.

The optimal robust estimators, i.e. the estimators with the least relative root mean squared errors, have very low bias in all contamination scenarios and very low relative root mean squared error. The highest relrmse occurs for the RHT estimator at an OAR-NCAR contamination with outlier rate $\epsilon = 0.01$: 1.27% with a bias of −1.05%. The optimal tuning constant of the RHT depends only on the outlier rate but not on the particular scenario. It is $k = 7.7$ for $\epsilon = 0$, $k = 4.3$ for $\epsilon = 0.001$ and $k = 2$ for $\epsilon = 0.01$. Obviously the standard tuning constant $k = 6.2$ is relatively good at the lower outlier rates but does not downweight enough at $\epsilon = 0.01$. The grid for the trimming proportion is finer than for k and thus there is a slight variability visible but roughly also here the optimal trimming proportion is $\alpha_u = 0$ for $\epsilon = 0$, as to be expected, $\alpha_u = 0.005$ for $\epsilon = 0.001$ $\alpha_u = 0.03$ for $\epsilon = 0.01$. The optimal tuning constants thus downweight more observations than the outliers.

The RHT estimator with the standard tuning constant $k = 6.2$ has slightly smaller bias and relrmse for outlier rate $\epsilon = 0.001$ than the trimmed mean TM and the robust estimators are much better than the classical non-robust estimator. In fact the highest relrmse of RHT and TM is 1.16% at the NCAR scenarios compared to the 3.7% of the classical HT at the OCAR-CCAR scenario. For the higher outlier rate $\epsilon = 0.01$ the standard estimators have too low tuning constants but nevertheless perform relatively well. Again the RHT estimator performs better than the trimmed mean at CCAR contamination and somewhat better at NCAR contamination.

Overall it seems advisable to use a robust estimator instead of the classical estimator because the loss with a slight robustification is minimal if the data behaves well (no contamination) while the gain is moderate to large if there is contamination in the data.

We now look at the results for the variance estimators for the robustified Horvitz-Thompson estimator and for the trimmed mean. Table 10.2 shows the performance of the linearized variance estimators for the estimators with standard tuning constant. The variance estimator for the RHT estimator has a positive bias and a moderate to large relrmse. In spite of the large bias at $\epsilon = 0.01$ the variance estimator remains useful. The variance estimator of the trimmed mean is not satisfactory, in particular for the high outlier rate. It seems that important aspects of the variability are not captured. More research is needed to refine this variance estimator and in the mean time only resampling procedures seem to be sufficiently reliable for trimmed means.

10.3 Robust non-parametric estimation of the Quintile Share Ratio

The estimators of the Quintile Share Ratio (QSR) used in the simulations were wPDC-CN from TUW and TQSR and SQSR from FHNW.

Table 10.2: Linearized variance estimators at AAT-SILC with simple random sampling

			OCAR				OAR			
			CCAR		NCAR		CCAR		NCAR	
V	$\epsilon =\rightarrow$	0	0.001	0.01	0.001	0.01	0.001	0.01	0.001	0.01
RHT	relbias	2.0	12.3	71.8	9.8	50.2	10.9	57.5	9.7	54.2
RHT	relmse	5.1	13.4	72.3	11.0	50.7	12.0	58.1	10.9	54.7
TM	relbias	-50.9	68.8	1553.7	52.7	280.4	65.9	1558.9	54.3	342.9
TM	relmse	51.1	69.9	1555.2	53.6	283.2	67.0	1560.2	55.3	345.3
opt α		0	0.005	0.031	0.003	0.026	0.003	0.028	0.003	0.026
TM	relbias	-35.9	68.8	339.5	66.1	108.0	217.1	378.7	67.3	118.2
TM	relmse	36.3	69.9	355.7	67.1	109.1	218.2	388.8	68.4	119.3

© http://ameli.surveystatistics.net/ - 2011

10.3.1 TQSR and SQSR

The robustification by trimming a proportion α_u of the largest observations is called TQSR. TQSR necessarily has a bias, which SQSR tries to compensate for. A detailed description of these two estimators together with an estimator of their variance is described in (Hulliger et al., 2011b, Chapter Robust Quintile Share Ratio).

AMELIA SRS

We start by discussing the simulation runs with a simple random sample of households from the AMELIA population. These are runs **N0010** for SQSR and **N0011** for TQSR. A standard analysis output for the two runs is included in Appendix (Hulliger et al., 2011a).

We first investigate the behaviour of the robust estimators when different tuning constants are chosen. Tuning constants can be chosen by the statistician using these estimators. There is a lot of freedom in this choice but also a lot of uncertainty and insight into the behaviour of the estimators is needed for a well-founded choice. In practice it is clearly not appropriate just to look at one tuning constant. A range of constants must be considered to obtain an overview and make a choice.

Figure 10.1 shows the steadily growing negative bias of TQSR when the trimming proportion is increased. The bias of SQSR becomes slightly negative first and then increases monotonically when the trimming proportion is increased. This shows that robustification induces a bias and that SQSR can compensate a bias in a small range of trimming but then overcompensates heavily. Note that for a trimming proportion of $\alpha = 0$ both TQSR and SQSR correspond to the classical Quintile Share Ratio estimator.

Figure 10.2 shows that the variance of TQSR decreases monotonically with the trimming proportion while for the SQSR the initial decrease changes to a steady increase when larger proportions are used.

Note that the trimming proportions are small in these examples, ranging from $\alpha_u = 0$ to 0.03 only. This must be seen in the light of the sample size, which is approximately 16 000 for the mean AMELIA household size of 2.648. A trimming proportion of 0.03 would already mean to trim away some 480 observations. This would probably be considered as data of low quality.

The relative root mean squared error criterion relrmseT (see Section 6.2) gives a good impression of the joint behaviour of variance and bias. Figure 10.3 shows that the bias is the dominant term and therefore SQSR is about as efficient as the classical non-robust QSR estimator for a moderate range of trimming up to $\alpha = 0.005$ while the bias yields TQSR inefficient already for slight trimming.

This behaviour of TQSR and SQSR changes when contamination is introduced. We look at a situation where a proportion of $\epsilon = 0.001$, i.e. about 16 observations are declared outliers completely at random and where the contamination has a normal distribution with mean 500 000 and standard deviation 20 000. This situation is called OCAR-CCAR since outlyingness and contamination are completely at random (see 4.2.4). This seems quite

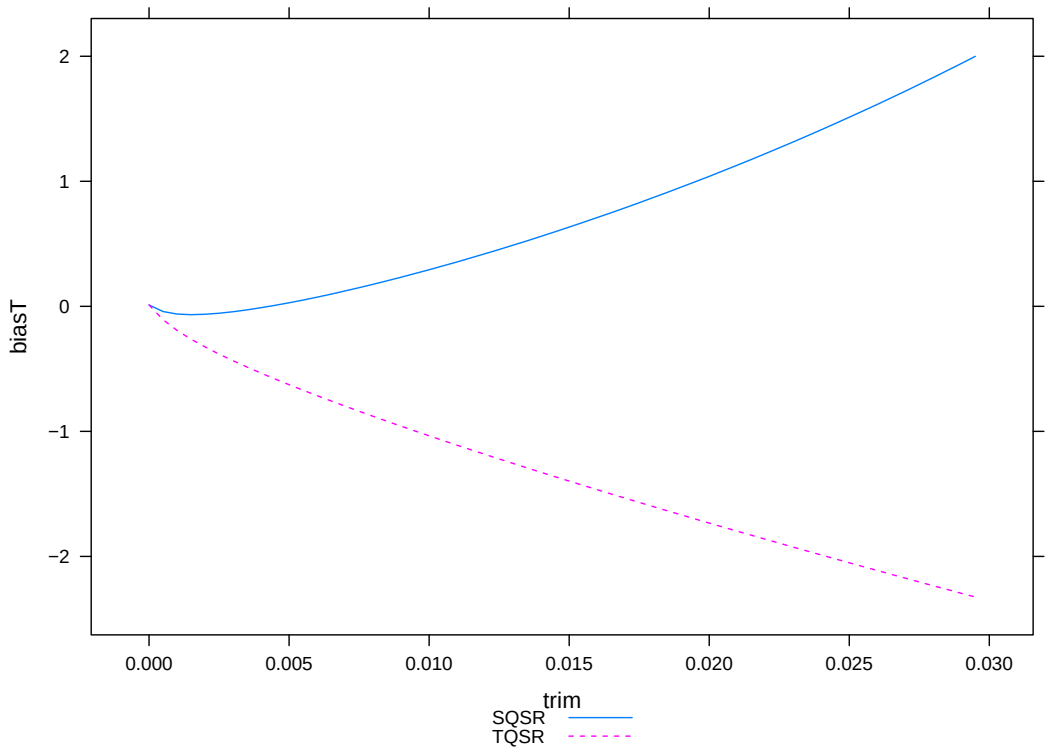


Figure 10.1: Bias of TQSR and SQSR at AMELIA, no contamination, simple random sampling

a realistic situation because even after a careful editing it is possible that 16 observations may not be declared outliers even if they are. Note that we consider the outliers as non-representative and the target characteristic to estimate remains the QSR of the true, uncontaminated data.

Figure 10.4 shows the relative root mean squared error and Figure 10.5 the bias at the contamination rate $\epsilon = 0$, i.e. there is no contamination, $\epsilon = 0.001$ a realistic contamination, and $\epsilon = 0.01$ a heavier contamination. The left most panel of these Figures correspond to Figures 10.3 and 10.1 respectively. The middle panel of Figures 10.4 and 10.5 is the one with a realistic outlier rate $\epsilon = 0.001$. With this light contamination the QSR estimator without robustification has a bias as an estimator for the uncontaminated QSR in the population. The bias of TQSR is minimal at approximately $\alpha = 0.002$ and then becomes more and more negative. The bias of the SQSR is minimal at $\alpha = 0.004$ and is growing when more trimming is applied. In other words, the bias compensation of SQSR is too strong for larger trimming proportions. Due to the dominance of the bias it is also visible in the relative root mean squared error that there is a relatively sharp minimum for TQSR (relrmseT=3.6125) at $\alpha = 0.002$ while the minimum is flatter for SQSR and somewhat higher (relrmseT=4.2639) at $\alpha = 0.0045$. Note that the non-robust classical QSR estimator, which corresponds to $\alpha = 0$, has relrmseT=6.1819. In other words the robust estimators TQSR and SQSR outperform by far the classical estimator if the best tuning constant is chosen. In practice the choice of the tuning constant is difficult since we do not know exactly how much contamination we have and what form the tails of the distribution have. Therefore the choice of the tuning constant may be sup-optimal. Since

© http://ameli.surveystatistics.net/ - 2011

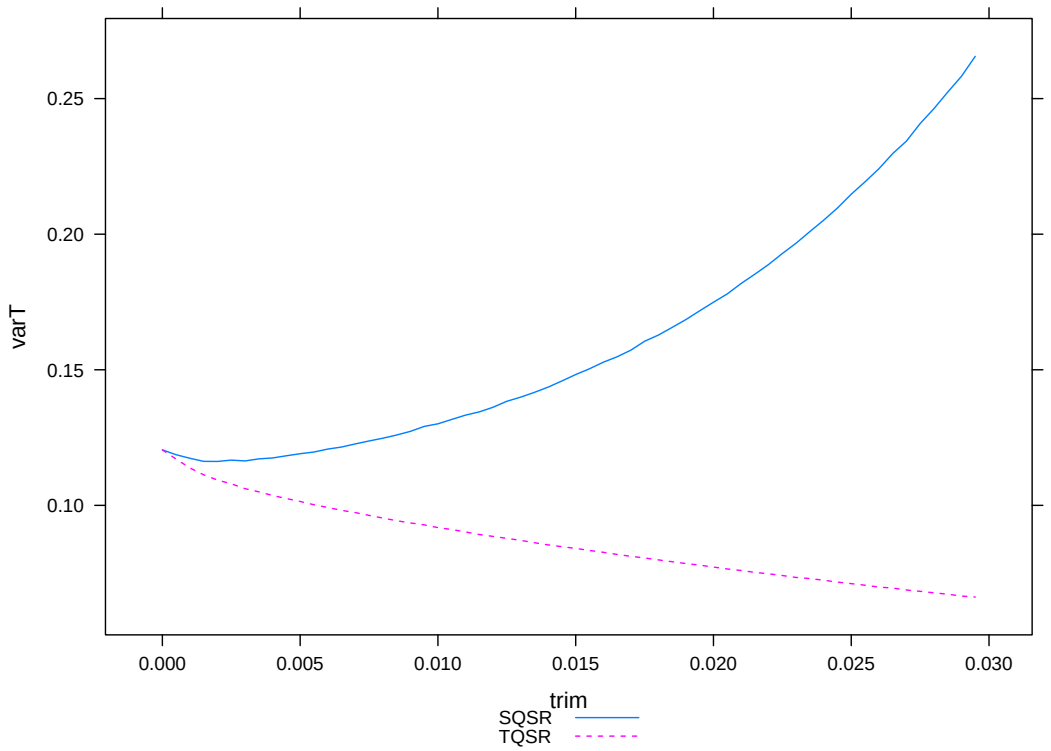


Figure 10.2: Variance of TQSR and SQSR at AMELIA, no contamination, simple random sampling

the minimum is flatter for the SQSR estimator and therefore the choice of the tuning constant is less critical, the SQSR estimator might be the preferable to the TQSR.

The situation is analogue but much more extreme for an outlier rate of $\epsilon = 0.01$. Here the TQSR outperforms SQSR for a large range of tuning constants and would be the preferred robust estimator. Of course the TQSR could be still outperformed by a bias compensation where an optimal trimming of the upper end of the first quintile is picked. This lower trimming proportion α_l would be chosen lower than the default choice of the SQSR which for strongly asymmetric tails is roughly double the trimming proportion α_u in the upper quintile. For practical purposes therefore an alternative to SQSR would be the so-called BQSR which just sets the lower trimming proportion to the upper trimming proportion, i.e. $\alpha_l = \alpha_u$.

Table 10.3 shows results for the performance of classical QSR, TQSR and SQSR at different contamination scenarios. For each scenario the relbias and relrmse of TQSR and QSR are given for a standard trimming proportion $\alpha_u = 0.005$ and for the optimal trimming proportion. The optimal trimming proportion is the one that yields least relative root mean squared error. Remember that TQSR does not trim in the first quintile, while SQSR uses a bias compensation trimming where the ratio of asymmetries in the fifth and first quintile is taken into account to determine the lower trimming proportion α_l.

The effect of the OAR mechanism on bias and MSE of all estimators is small and we concentrate on the analysis of the OCAR mechanisms. Without contamination the relative bias of SQSR is low for a standard tuning constant $\alpha_u = 0.005$ or for the optimal tuning

Table 10.3: Relative bias and root mean squared error of QSR estimators

			OCAR			OAR	
			CCAR	CCAR	NCAR	CCAR	NCAR
T	$\epsilon =\rightarrow$	0	0.001	0.01	0.001	0.001	0.001
QSR	relbiasT%	0.0	4.9	48.4	2.9	4.9	3.3
QSR	relrmseT%	3.7	6.2	48.6	4.8	6.2	5.0
TQSR5	relbiasT%	-6.5	-4.7	32.6	-5.3	-4.7	-5.2
TQSR5	relrmseT%	7.3	5.8	33.0	6.3	5.8	6.2
opt	α	0	0.002	0.0215	0.001	0.002	0.0015
TQSRopt	relbiasT%	0.0	-0.1	0.0	0.1	-0.1	-0.6
TQSRopt	relrmseT%	3.7	3.6	3.6	3.7	3.6	3.6
SQSR5	relbiasT%	0.2	2.2	43.6	1.5	2.2	1.7
SQSR5	relrmseT%	3.6	4.3	43.9	4.0	4.3	4.0
opt	α	0.0045	0.0045	0.0110	0.0030	0.0040	0.0030
SQSRopt	relbiasT%	0.0	2.1	42.4	1.1	2.2	1.3
SQSRopt	relrmseT%	3.6	4.3	42.8	3.8	4.3	3.9

Notes: Standard tuning constant $\alpha_u = 0.005$ and optimal tuning constant.

© http://ameli.surveystatistics.net/ - 2011

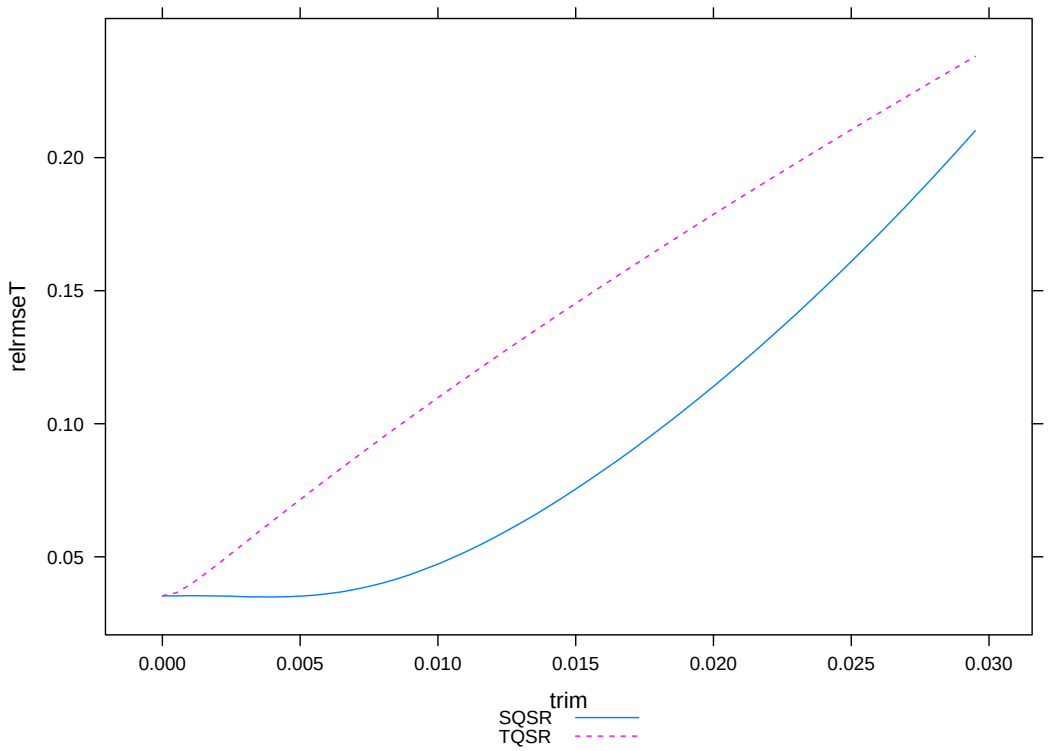


Figure 10.3: Relative root mean squared error of TQSR and SQSR at AMELIA, no contamination, simple random sampling

constant. The standard tuning constant yields a negative bias of 6.5%, which is substantial considering the small amount of trimming. The efficiency is good for the standard and optimally tuned SQSR and for the optimally tuned TQSR. Of course the relatively large bias of the standard TQSR yields a relatively large relrmse, too.

It may be argued that the standard tuning constant $\alpha_u = 0.005$ lies quite close to the optimal tuning constant $\alpha_u = 0.0045$, which is not realistic. However, if there is no contamination the relrmseT of the SQSR is practically constant and in any case much smaller than for the TQSR in a wide range, as can be seen in Figure 10.3.

In a completely random outlying and contamination scenario the non-robust classical QSR has a bias of 4.9% when the outlier rate is $\epsilon = 0.001$ and a bias of 48.4% when the outlier rate is $\epsilon = 0.01$. Thus the impact and biasing effect of outliers is heavy. Trimming by $\alpha_u = 0.005$ yields a negative bias -4.7% of TQSR for the mild outlier rate and a positive bias of 32% for $\epsilon = 0.01$. Obviously this standard tuning constant is far away from the optimal tuning constant $\alpha_u = 0.002$ at $\epsilon = 0.001$ and $\alpha_u = 0.0215$ at $\epsilon = 0.01$, which yields a very low bias and consequently relrmse. Thus it seems that trimming roughly double the outlier rate would yield a good TQSR. The bias compensation of SQSR seems to be about right with $\alpha_u = 0.005$ for the mild outlier rate $\epsilon = 0.001$ because the bias is moderate (2.1%) and the relrmse is optimal at $\alpha_u = 0.0045$. The efficiency loss compared with the optimal TQSR is substantial but the gain compared with the non-robust QSR is larger.

The outlier rate $\epsilon = 0.01$ poses difficulties to the SQSR estimator because the bias com-

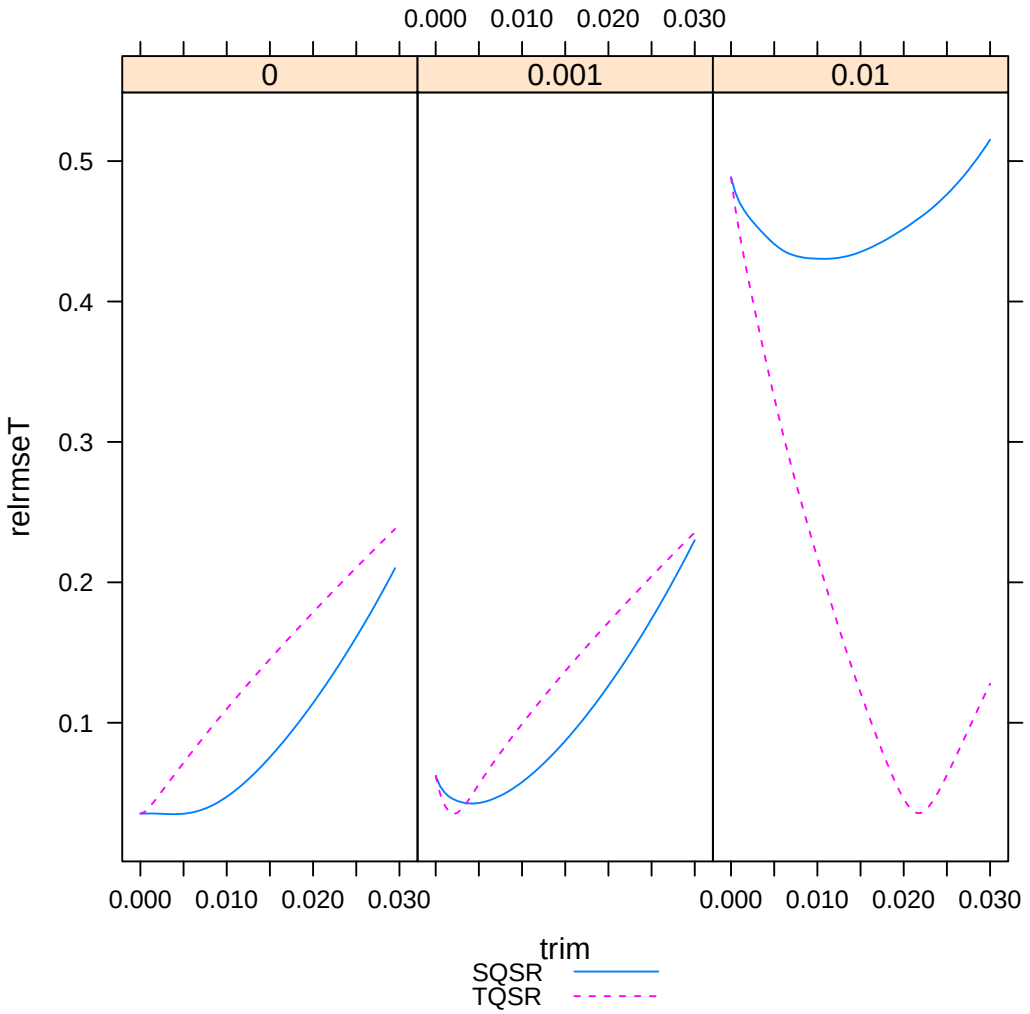


Figure 10.4: Relative root mean squared error of TQSR and SQSR at AMELIA, simple random sampling, OCAR-CCAR with $\epsilon = 0, 0.001, 0.01$

pensation is too strong. The bias with trimming $\alpha_u = 0.005$ is close to the bias when minimising relrmse and only moderately lower than for the non-robust QSR. TQSR is better than SQSR here but only if a good tuning constant is chosen.

For an OCAR-NCAR (and OAR-NCAR) contamination with outlier rate $\epsilon = 0.001$ we get a similar picture as with OCAR-CCAR except that now SQSR with standard and optimal tuning constant favor even better.

Variance estimation of the QSR, TQSR and SQSR was evaluated with the relative bias of the variance estimator and with the relative root mean squared error of the variance estimator. Variance estimation of the QSR-estimators is difficult because of the quantiles involved and because of the non-linearity of the estimators. Table 10.4 shows the relative bias and relative root mean squared error for the AMELIA population when simple random sampling (design 1.2) is used. The upper trimming proportion is $\alpha_u = 0.005$. When there is no contamination, i.e. at $\epsilon = 0$ the variance estimator overestimates the Monte Carlo variance by 25% to 32%, which is moderate. The variance of the variance estimator

© http://ameli.surveystatistics.net/ - 2011

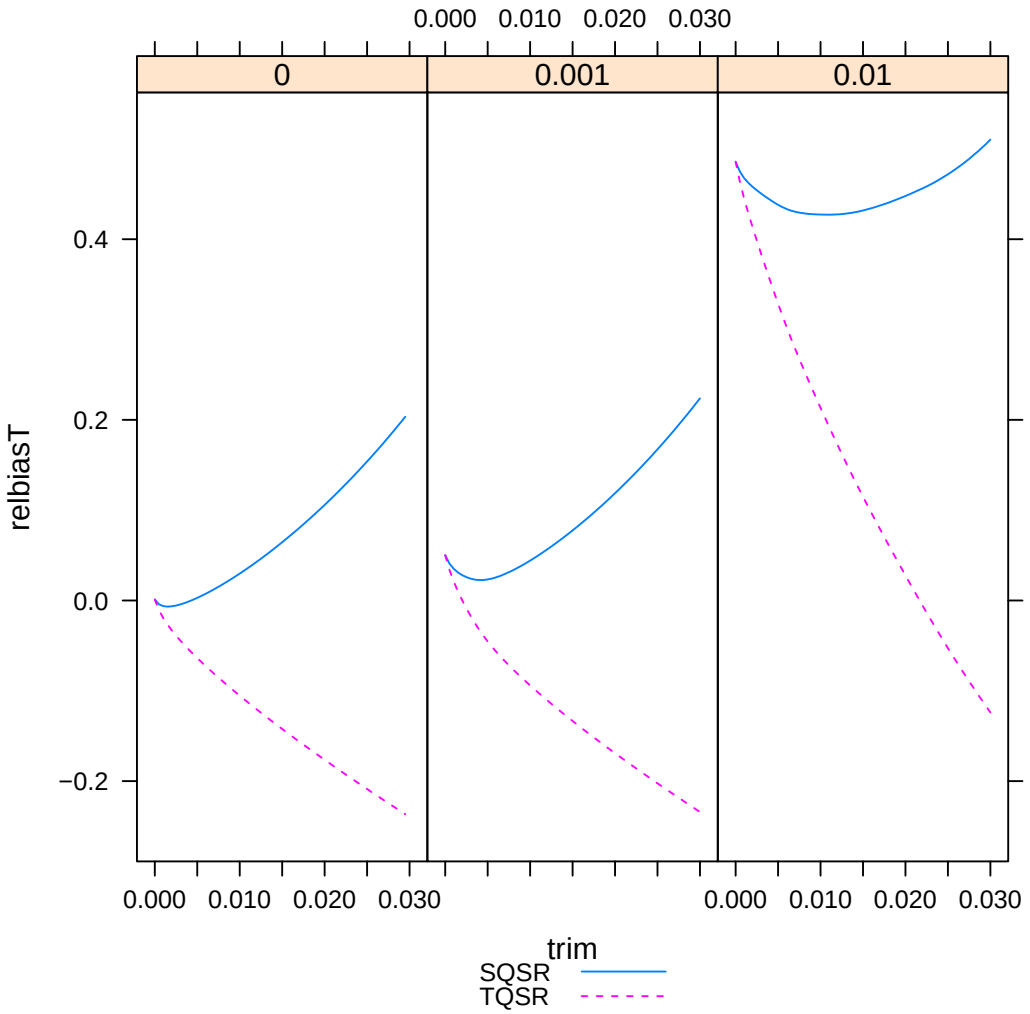


Figure 10.5: Bias of TQSR and SQSR at AMELIA, simple random sampling, OCAR-CCAR with $\epsilon = 0, 0.001, 0.01$

does not seem to be a big problem compared to this bias, since most of the relative root mean squared error is due to the bias. The bias is larger when contamination occurs. The bias at $\epsilon = 0.001$ is about 30% to 40% and at $\epsilon = 0.01$ the bias is about 90%.

The variance estimators are thus not satisfactory and it is difficult to judge how much overestimation will result in a practical situation. The bias seems to increase when the outlier rate increases. Confidence intervals based on these variance estimators will be conservative. As an alternative, though maybe costly, resampling variance estimators might be used.

The choice of the tuning constant obviously depends on an educated guess of how much contamination the sample has. If a minimal protection against extreme disposable incomes should be achieved then α must be larger than 0. For these large samples trimming must be very light in order not to introduce to much bias. A trimming proportion of $\alpha = 0.005$ seems to be a reasonable start for SILC data.

Table 10.4: Relative bias and root mean squared error of variance estimators

ϵ	T	relbiasV%	relrmseV%
0	QSR	25.2	30.2
	TQSR	26.3	30.9
	SQSR	31.9	36.2
0.001	QSR	40.3	44.1
	TQSR	29.6	33.9
	SQSR	34.9	38.9
0.01	QSR	90.7	93.0
	TQSR	92.2	94.5
	SQSR	98.0	100.3

Notes: Universe AMELIA, simple random sampling, trimming $\alpha_u = 0.005$.

© http://ameli.surveystatistics.net/ - 2011

AMELIA with different designs

Stratification with probability proportional to size sampling of households (i.e. design 1.5a in Table 2.1) yields the results in Table 10.5. Comparing the results with simple random sampling (Table 10.3) they look very similar and thus stratification and pps-sampling has a minor effect on the results.

Table 10.5: Relbias and relrmse of QSR estimators at stratified pps sampling

		OCAR			OAR		
		CCAR	CCAR	NCAR	CCAR	NCAR	NCAR
R	$\epsilon =\rightarrow$	0.001	0.01	0.001	0.001	0.001	0.01
QSR	relbiasT	5.2	48.6	3.2	5.2	3.5	33.5
QSR	relrmseT	**6.5**	**48.9**	**5.0**	**6.5**	**5.2**	**33.9**
TQSR5	relbiasT	-4.5	-4.5	32.8	-4.4	-4.9	16.8
TQSR5	relrmseT	**5.6**	**5.6**	**33.2**	**5.6**	**6.0**	**17.3**
opt. α		0.0020	0.0210	0.0015	0.0020	0.0015	0.0145
TQSRopt	relbiasT	0.1	0.1	-0.6	0.1	-0.4	0.1
TQSRopt	relrmseT	**3.6**	**3.5**	**3.7**	**3.6**	**3.6**	**3.7**
SQSR5	relbiasT	2.4	43.8	3.5	2.5	1.9	25.9
SQSR5	relrmseT	**4.5**	**44.1**	**5.2**	**4.5**	**4.2**	**26.3**
opt. α	α	0.0040	0.0110	0.0025	0.0040	0.0025	0.0095
SQSRopt	relbiasT	2.3	42.5	1.3	2.4	1.4	24.9
SQSRopt	relrmseT	**4.4**	**42.9**	**3.9**	**4.4**	**4.0**	**25.3**

In order to compare the effect of the universe the results for the universe AAT-SILC with a simple random sample design are presented in Table 10.6. It seems that the standard trimming chosen for SQSR with the AMELIA population, $\alpha_u = 0.005$ performs well also for AAT-SILC. In other words the optimal α_u for SQSR is close to 0.005 also for AAT-SILC. The optimal trimming for TQSR, which does not use a bias compensation, is slightly closer to 0.005 than for AMELIA and this tends to reduce the bias of TQSR with $\alpha_u = 0.005$. The impact of contamination on the non-robust classical QSR is heavier for the AAT-SILC population, introducing more bias. This is not the case for the robust estimators, where bias is similar to the AMELIA universe. The figures of relative MSE are slightly lower for AAT-SILC but in general follow a similar pattern as for AMELIA.

Table 10.7 shows the results of the simulations for the two universes at stratified random sampling with simple random sampling within the strata (design 1.4a). The contamination mechanism considered is OCAR-CCAR. The features discussed above are similar in both universes but the effect is often quantitatively larger in AMELIA than in AAT-SILC when the outlier rate is 0 or 0.001. At $\epsilon = 0.01$ the contamination may have a large effect and we can see that the choice of the tuning constant $\alpha_u = 0.005$ is far from optimal. The effect of the sub-optimal choice is heavy for TQSR, where the optimal tuning would yield a very good estimator in terms of bias and variance. At the AAT-SILC universe SQSR shows a considerable improvement between the standard choice $\alpha_u = 0.005$ and the optimal tuning constant: Relative RMSE drops from 75.80 to 41.15. At the AMELIA population the improvement of the optimal tuning constant is small.

Table 10.6: Relbias and relrmse of QSR estimators with universe AAT-SILC and simple random sampling

		OCAR				OAR	
		CCAR	NCAR		CCAR	NCAR	
T	$\epsilon =\rightarrow$	0	0.001	0.001		0.001	0.001
QSR	relbiasT	0.0	9.6	3.9		9.4	4.2
QSR	relrmseT	1.9	9.8	4.4		9.6	4.7
TQSR5	relbiasT	-6.1	-3.8	-4.3		-3.8	-4.2
TQSR5	relrmseT	6.4	4.2	4.6		4.2	4.5
opt	α	0.0000	0.0030	0.0020		0.0030	0.0020
TQSRopt	relbiasT	0.0	-0.4	-0.4		-0.5	-0.2
TQSRopt	relrmseT	1.9	1.9	1.9		1.9	1.9
SQSR5	relbiasT	-0.5	2.1	1.6		2.0	1.7
SQSR5	relrmseT	2.0	2.9	2.6		2.9	2.6
opt	α	0.0000	0.0060	0.0045		0.0055	0.0045
SQSRopt	relbiasT	0.0	2.1	1.6		2.0	1.7
SQSRopt	relrmseT	1.9	2.9	2.5		2.9	2.6

Table 10.7: QSR-estimators at stratified random sampling with both universes

		AAT-SILC	AMELIA	AAT-SILC	AMELIA	AAT-SILC	AMELIA
T	ϵ	0	0	0.001	0.001	0.01	0.01
QSR	relbiasT	-0.04	0.11	9.50	5.02	90.60	48.58
QSR	relrmseT	1.89	3.53	9.71	6.22	90.65	48.84
TQSR5	relbiasT	-6.18	-6.38	-3.84	-4.54	64.31	32.86
TQSR5	relrmseT	6.43	7.15	4.25	5.62	64.38	33.17
opt	α	0.000	0.000	0.003	0.002	0.025	0.0215
TQSRopt	relbiasT	-0.04	0.11	-0.46	0.03	-0.26	0.28
TQSRopt	relrmseT	1.89	3.53	1.94	3.50	2.13	3.58
SQSR5	relbiasT	-0.50	0.29	2.05	2.33	75.72	43.81
SQSR5	relrmseT	2.06	3.52	2.91	4.29	75.80	44.09
opt	α	0.0000	0.0045	0.0055	0.0040	0.0295	0.011
SQSRopt	relbiasT	-0.04	0.08	2.01	2.26	40.67	42.72
SQSRopt	relrmseT	1.89	3.50	2.89	4.25	41.15	43.04

Note: Contamination is OCAR-CCAR

© http://ameli.surveystatistics.net/ - 2011

10.4 Robust semiparametric estimation

Since EU-SILC data may contain nonrepresentative outliers in the upper tail of the income distribution, robust Pareto tail modeling for reducing the influence of outliers is investigated. These methods are described in detail in Hulliger et al. (2011b). In order to fit a Pareto distribution to the upper tail of the data, a threshold needs to be selected first. Holzer (2009) concluded that graphical methods perform best in practical situations in the case of EU-SILC. For details on these graphical tools, the reader is referred to Hulliger et al. (2011b). However, graphical methods have the disadvantage that the threshold cannot be determined exactly. Therefore the main aim of the simulation studies for semiparametric estimation was to evaluate how the methods behave for different choices of the threshold. Serveral methods for estimating the shape of the Pareto distribution are thereby investigated: Hill, wHill, ISE, wISE, PDC and wPDC (see Hulliger et al., 2011b, and references therein).

Basically, three general approaches for semiparametric estimation based on Pareto tail modeling have been developed:

Replacement of the tail (RT): All values above the threshold are replaced by values drawn from the fitted distribution. The order of the original values is preserved.

Replacement of nonrepresentative outliers (RN): Values larger than a certain α-quantile of the fitted distribution (with α close to 1) are declared as nonrepresentative outliers. Only these nonrepresentative outliers are replaced by values drawn from the fitted distribution, thereby preserving the order of the original values. For the results in this report, $\alpha = 0.99$ is used.

Calibration for nonrepresentative outliers (CN): Values larger than a certain α-quantile of the fitted distribution (with α close to 1) are declared as nonrepresentative outliers. Since these are considered to be unique to the population data, the sample weights of the corresponding observations are set to 1 and the weights of the remaining observations are adjusted accordingly by calibration. For the results in this report, α is set to 0.99 and the variable giving the NUTS2 regions is used as auxiliary information for calibration.

The basis for the simulation studies included in this report is the synthetic AAT-SILC data set. In these examples, the samples are drawn with stratified Midzuno sampling in order to investigate the differences between the weighted and unweighted estimators for fitting the Pareto distribution (for the estimation of the indicators, the sample weights are of course always taken into account). See Part 4 for a description of the simulation scenarios. It should also be noted that only results for the quintile share ratio (QSR) are shown here.

However, the results included in this report were selected representatively. A complete overview of all simulation results for the semiparametric methods can be found in the appendix (Hulliger et al., 2011a, Ch.F).

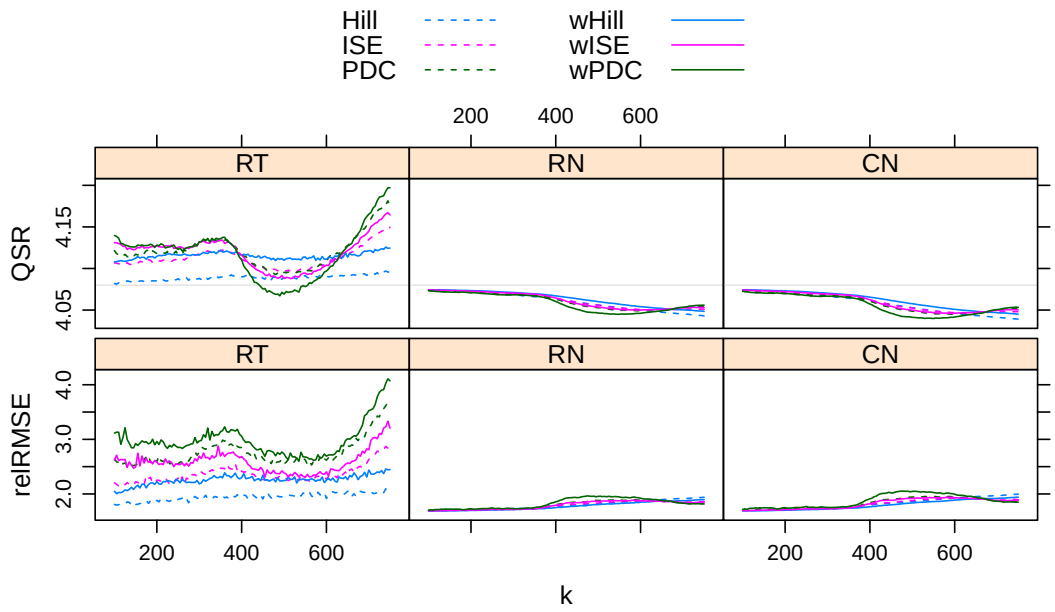


Figure 10.6: Average results (*top*) and RMSE (*bottom*) for the QSR using the three approaches for Pareto tail modeling with k varying between 100 and 750. Scenario: AAT-SILC, stratified Midzuno sampling, and no contamintation.

10.4.1 No contamination

This section presents simulation results for the case where no contamination is added to the data.

Figure 10.6 shows the average results (*top*) and RMSE (*bottom*) for the semiparametric QSR estimates with the number of households k used for tail modeling varying between 100 and 750. The different estimators of the shape of the Pareto distribution are thereby compared for the three approaches for tail modeling. In addition, the grey reference line represents the true population value of the QSR. Clearly, the RT approach introduces a considerable amount of additional uncertainty and produces rather unstable results. The curves for RN and CN approaches, on the other hand, are much smoother and reflect the better performance. In fact, these two approaches lead to very similar results. This is not surprising, because in the case without contamination, they should ideally not flag any observations as outliers, resulting in standard estimation of the QSR. Consequently, the curves for the different estimators of the shape of the Pareto distribution are almost superposed up to $k \approx 400$ and are very close to the true population value. Nevertheless, the quality of the estimates deteriorates a bit for larger values of k, which is visible in the curves by sudden kinks away from the reference line. This means that if too many observations are used for tail modeling, false positives are detected and replaced or downweighted, respectively, which in turn causes a negative bias.

To further investigate the performance of the robust PDC estimator and its weighted counterpart wPDC, Figure 10.7 uses boxplots to compare these estimators for the three approaches for tail modeling with the standard estimation of the QSR and a parametric method based on fitting a GB2 distribution to the data. For the PDC and wPDC estim-

© http://ameli.surveystatistics.net/ - 2011

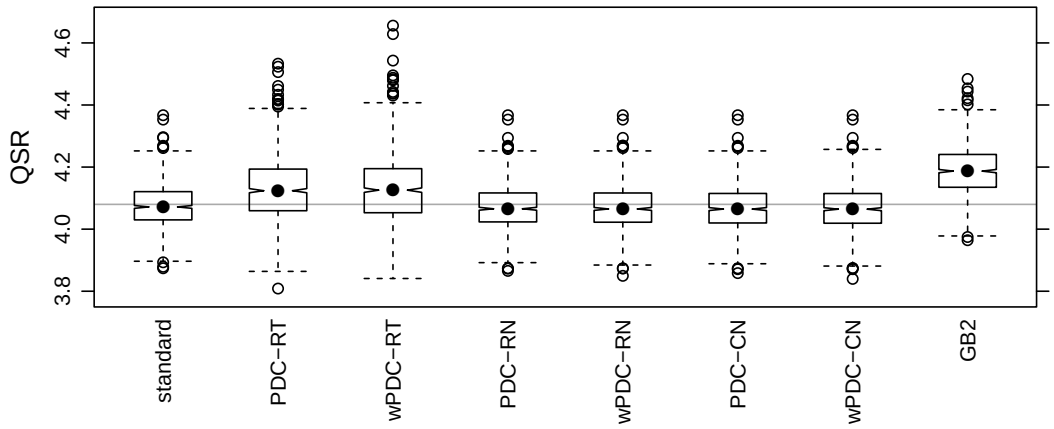


Figure 10.7: Box plots of the simulation results for the QSR using standard estimation, the three semiparametric approaches for Pareto tail modeling with the PDC and wPDC estimators and $k = 300$, as well as parametric estimation using a GB2 distribution with the profile log-likelihood approach. Scenario: AAT-SILC, stratified Midzuno sampling, and no contamintation.

ators, $k = 300$ is selected representatively. This plot again illustrates the similarity of the semiparametric estimates with the RN and CN approaches and the standard estimation. Furthermore, the additional variability with the RT approach is clearly visible, as well as a small bias. An even larger bias is obtained with the parametric estimate based on fitting a GB2 distribution with the profile log-likelihood method. However, Graf et al. (2011) have in the mean time developed a simple adjustment of the parametric estimator that should produce more accurate results.

To complement the graphical evaluation methods, Table 10.8 contains all evaluation criteria for univariate point estimation as described in Section 6.2: average, variance, bias, relative bias (relBias), median, median absolute deviation (MAD), median error (MedE), root mean squared error (RMSE), relative root mean squared error (relRMSE), median absolute error (MedAE), and maximum absolute relative error (MaxARE). In particular the relBias and the relRMSE are of interest. For the semiparametric methods, results for $k = 300, 500, 700$ are shown. These choices of k are selected representatively to compare the results across different contamination settings (cf. the following sections). In any case, Table 10.8 also reflects the similar results for the estimators of the shape of the Pareto distribution in the case of the RN and CN approaches and the excellent performance of these estimates.

With the results presented in this section, it should be noted that there is not much difference between weighted and unweighted estimators of the shape of the Pareto distribution for the RN and CN approaches. The reason is that these procedures remain quite stable for slight misspecifications of the Pareto model, which is also why there is a wide range for the coice of k that leads to excellent results.

© http://amelia.surveystatistics.net/ - 2011

Table 10.8: Evaluation of the different estimation methods for the QSR at AAT-SILC, stratified Midzuno sampling, and no contamination.

Method	Type	k	Average	Variance	Bias	relBias	Median	MAD	MedE	RMSE	relRMSE	MedAE	MaxARE
standard			4.07	0.00	−0.01	−0.13	4.07	0.07	−0.01	0.07	1.68	0.07	0.07
wHill	RT	300	4.12	0.01	0.04	0.89	4.11	0.08	0.03	0.09	2.18	0.09	0.09
wHill	RT	500	4.11	0.01	0.03	0.74	4.10	0.09	0.02	0.09	2.25	0.09	0.09
wHill	RT	700	4.12	0.01	0.04	1.02	4.12	0.09	0.04	0.10	2.44	0.09	0.11
wHill	RN	300	4.07	0.00	−0.01	−0.21	4.07	0.07	−0.01	0.07	1.71	0.07	0.07
wHill	RN	500	4.06	0.01	−0.02	−0.48	4.06	0.07	−0.02	0.07	1.81	0.08	0.07
wHill	RN	700	4.05	0.01	−0.03	−0.73	4.05	0.07	−0.03	0.08	1.88	0.08	0.07
wHill	CN	300	4.07	0.00	−0.01	−0.24	4.07	0.07	−0.01	0.07	1.72	0.07	0.07
wHill	CN	500	4.06	0.01	−0.02	−0.55	4.05	0.07	−0.03	0.07	1.83	0.08	0.07
wHill	CN	700	4.05	0.01	−0.03	−0.81	4.04	0.07	−0.03	0.08	1.92	0.08	0.07
wISE	RT	300	4.13	0.01	0.05	1.28	4.13	0.10	0.05	0.11	2.69	0.10	0.10
wISE	RT	500	4.09	0.01	0.01	0.20	4.08	0.09	0.00	0.10	2.39	0.09	0.11
wISE	RT	700	4.14	0.01	0.06	1.59	4.14	0.09	0.06	0.11	2.81	0.11	0.09
wISE	RN	300	4.07	0.00	−0.01	−0.23	4.07	0.07	−0.01	0.07	1.72	0.07	0.07
wISE	RN	500	4.05	0.01	−0.03	−0.64	4.05	0.07	−0.03	0.08	1.86	0.08	0.07
wISE	RN	700	4.05	0.00	−0.03	−0.67	4.05	0.07	−0.03	0.08	1.85	0.08	0.07
wISE	CN	300	4.07	0.00	−0.01	−0.26	4.07	0.07	−0.01	0.07	1.73	0.07	0.07
wISE	CN	500	4.05	0.01	−0.03	−0.74	4.05	0.07	−0.03	0.08	1.92	0.08	0.07
wISE	CN	700	4.05	0.00	−0.03	−0.74	4.05	0.07	−0.03	0.08	1.88	0.08	0.07
wPDC	RT	300	4.13	0.01	0.05	1.28	4.13	0.11	0.05	0.12	2.97	0.11	0.14
wPDC	RT	500	4.07	0.01	−0.01	−0.26	4.06	0.11	−0.02	0.11	2.73	0.11	0.11
wPDC	RT	700	4.17	0.01	0.09	2.09	4.16	0.11	0.08	0.14	3.53	0.14	0.13
wPDC	RN	300	4.07	0.00	−0.01	−0.28	4.07	0.07	−0.01	0.07	1.73	0.07	0.07
wPDC	RN	500	4.05	0.01	−0.03	−0.82	4.05	0.07	−0.03	0.08	1.96	0.08	0.07
wPDC	RN	700	4.05	0.00	−0.03	−0.65	4.05	0.07	−0.03	0.07	1.83	0.07	0.06
wPDC	CN	300	4.07	0.00	−0.01	−0.32	4.06	0.07	−0.02	0.07	1.75	0.07	0.07
wPDC	CN	500	4.04	0.01	−0.04	−0.95	4.04	0.07	−0.04	0.08	2.05	0.09	0.08
wPDC	CN	700	4.05	0.00	−0.03	−0.72	4.05	0.07	−0.03	0.08	1.87	0.08	0.06
GB2			4.19	0.01	0.11	2.64	4.19	0.08	0.11	0.13	3.24	0.16	0.10

10.4.2 OCAR-CCAR, contamination level $\epsilon = 0.001$

In this section, simulation results for the contamination level $\varepsilon = 0.001$ are discussed. The outliers are thereby generated with an OCAR-CCAR mechanism based on a normal distribution $\mathcal{N}(\mu, \sigma)$ with $\mu = 500\,000$ and $\sigma = 20\,000$. With an average sample size of more than 18 000 persons corresponding to (exactly) 6 000 sampled households, this results in about 19 contaminated household per sample, which can be considered realistic.

Figure 10.8 displays the average results (*top*) and RMSE (*bottom*) for the semiparametric QSR estimates. Since outliers are added to the samples, the number of households that are needed for tail modeling increases compared to the situation without contamination. Consequently, only the results for k between 200 and 750 are included in the plot because of instabilities of the RT approach for too small values of k. First of all, significant differences in the results occur only for the RT approach, as it is more sensitive towards misspecifications of the Pareto model due to the considerable number of observations that are drawn from the distribution. However, the results for the RT approach are not satisfactory. Concerning the RN and CN approaches, the PDC/wPDC estimator still shows a slight dent in the curves at $k \approx 400$, which is again an indication that some false positives are detected for larger values of k. While the RN approach has a positive bias before that dent, the CN approach gives accurate results between $k \approx 250$ and $k \approx 400$. It is also interesting to see that the influence of the outliers on the Hill/wHill and ISE/wISE estimators decreases steadily for the RN and CN approaches as k increases. While the ISE/wISE estimator behaves quite similarly to the PDC/wPDC estimator but only with a slight dent at $k \approx 400$, the Hill/wHill estimator leads to inaccurate results if k is not large enough. Nevertheless, it is surprising that it is at all possible to reduce the influence of the outliers on the QSR with this non-robust estimator for the shape of the Pareto distribution.

In Figure 10.9, box plots are used to compare the PDC and wPDC estimators for the three approaches for tail modeling with the standard estimation of the QSR and the parametric method based on a GB2 distribution. As in the case without contamination, $k = 300$ is selected for the PDC and wPDC estimators. Clearly, the standard estimation is already influenced by the small amount of contamination and shows a significant bias. Also for the GB2 approach, the bias is significantly larger than in the non-contaminated case. In this example, the RN and CN approaches differ as far as bias is concerned. While there is a positive bias for the RN approach, the CN approach does not suffer from this problem. However, the two approaches are similar in terms of variability. In any case, the RT approach shows both a slightly larger bias (although it is a significant improvement to the standard estimation) and higher variability.

For completeness, Table 10.9 gives a numeric overview of the simulation results with the different evaluation criteria. It allows to study the aforementioned differences between the investigated approaches in more detail.

10.4.3 OCAR-CCAR, contamination level $\varepsilon = 0.01$

While the contamination level was quite realistic in the previous section, this section is focued on a larger, rather unrealistic contamination level $\varepsilon = 0.01$. Depending on the

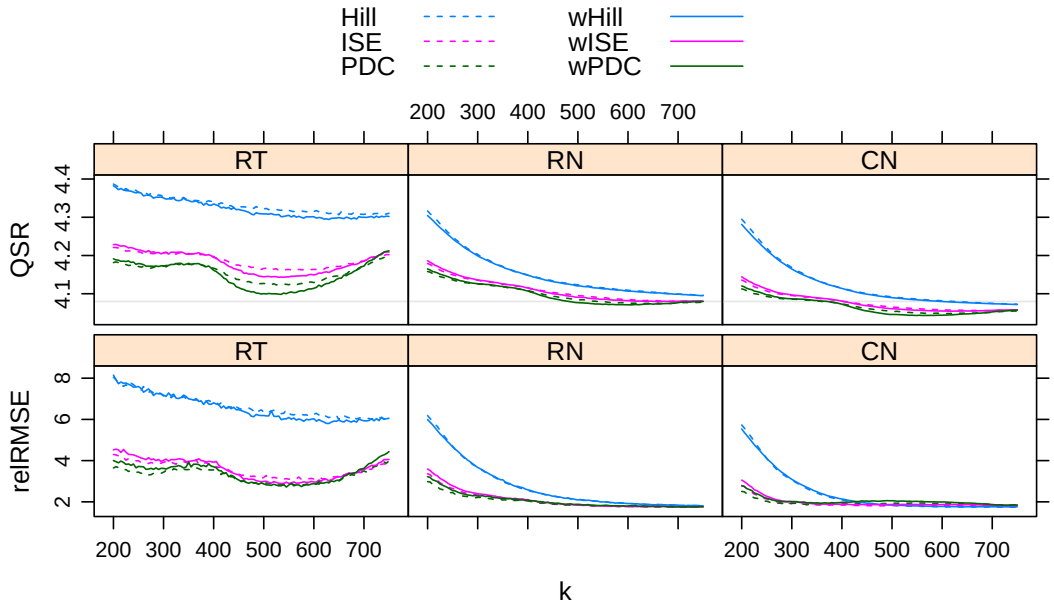


Figure 10.8: Average results (*top*) and RMSE (*bottom*) for the QSR using the three approaches for Pareto tail modeling with k varying between 200 and 750. Scenario: AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.001$.

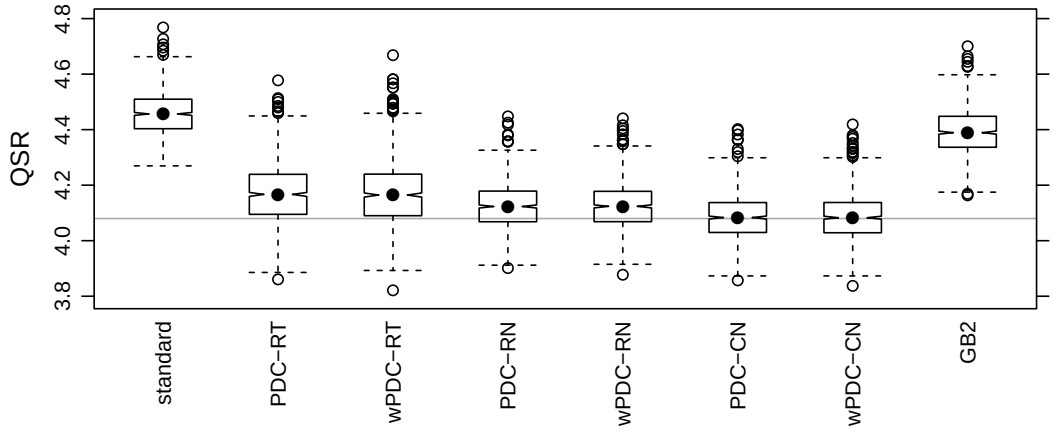


Figure 10.9: Box plots of the simulation results for the QSR using standard estimation, the three semiparametric approaches for Pareto tail modeling with the PDC and wPDC estimators and $k = 300$, as well as parametric estimation using a GB2 distribution with the profile log-likelihood approach. Scenario: AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.001$.

© http://ameli.surveystatistics.net/ - 2011

Table 10.9: Evaluation of the different estimation methods for the QSR at AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.001$.

Method	Type	k	Average	Variance	Bias	relBias	Median	MAD	MedE	RMSE	relRMSE	MedAE	MaxARE
standard			4.46	0.01	0.38	9.29	4.46	0.08	0.38	0.39	9.49	0.56	0.17
wHill	RT	300	4.35	0.01	0.27	6.63	4.34	0.11	0.26	0.29	7.18	0.39	0.19
wHill	RT	500	4.31	0.01	0.23	5.61	4.30	0.11	0.22	0.25	6.19	0.33	0.16
wHill	RT	700	4.30	0.01	0.22	5.40	4.30	0.10	0.22	0.24	5.93	0.32	0.15
wHill	RN	300	4.20	0.01	0.12	2.92	4.20	0.09	0.12	0.15	3.67	0.17	0.11
wHill	RN	500	4.12	0.01	0.04	0.99	4.12	0.07	0.04	0.09	2.09	0.09	0.09
wHill	RN	700	4.10	0.01	0.02	0.46	4.10	0.08	0.02	0.07	1.84	0.08	0.08
wHill	CN	300	4.16	0.01	0.08	2.08	4.16	0.10	0.08	0.13	3.09	0.13	0.11
wHill	CN	500	4.09	0.01	0.01	0.24	4.08	0.07	0.00	0.08	1.84	0.07	0.08
wHill	CN	700	4.07	0.01	−0.01	−0.14	4.07	0.07	−0.01	0.07	1.76	0.07	0.07
wISE	RT	300	4.21	0.01	0.13	3.11	4.20	0.10	0.12	0.16	4.01	0.18	0.15
wISE	RT	500	4.14	0.01	0.06	1.57	4.14	0.11	0.06	0.12	2.99	0.12	0.13
wISE	RT	700	4.19	0.01	0.11	2.58	4.19	0.10	0.11	0.14	3.55	0.16	0.13
wISE	RN	300	4.14	0.01	0.06	1.37	4.13	0.08	0.05	0.10	2.39	0.10	0.08
wISE	RN	500	4.09	0.01	0.01	0.27	4.09	0.07	0.01	0.08	1.84	0.07	0.08
wISE	RN	700	4.08	0.01	0.00	0.01	4.08	0.07	0.00	0.07	1.76	0.07	0.07
wISE	CN	300	4.10	0.01	0.02	0.42	4.09	0.08	0.01	0.08	2.01	0.08	0.08
wISE	CN	500	4.06	0.01	−0.02	−0.47	4.06	0.07	−0.02	0.08	1.87	0.08	0.07
wISE	CN	700	4.06	0.01	−0.02	−0.57	4.06	0.07	−0.02	0.07	1.84	0.08	0.07
wPDC	RT	300	4.17	0.01	0.09	2.28	4.16	0.11	0.08	0.15	3.62	0.14	0.14
wPDC	RT	500	4.10	0.01	0.02	0.50	4.09	0.12	0.01	0.12	2.87	0.11	0.11
wPDC	RT	700	4.17	0.01	0.09	2.27	4.17	0.11	0.09	0.15	3.62	0.15	0.13
wPDC	RN	300	4.13	0.01	0.05	1.13	4.12	0.08	0.04	0.09	2.28	0.09	0.09
wPDC	RN	500	4.08	0.01	−0.00	−0.08	4.07	0.07	−0.01	0.08	1.88	0.07	0.08
wPDC	RN	700	4.08	0.01	−0.00	−0.08	4.08	0.07	−0.00	0.07	1.76	0.07	0.08
wPDC	CN	300	4.09	0.01	0.01	0.17	4.08	0.08	0.00	0.08	2.00	0.08	0.08
wPDC	CN	500	4.05	0.01	−0.03	−0.84	4.04	0.07	−0.04	0.08	2.05	0.09	0.08
wPDC	CN	700	4.05	0.01	−0.03	−0.67	4.05	0.07	−0.03	0.08	1.87	0.08	0.06
GB2			4.39	0.01	0.31	7.68	4.39	0.08	0.31	0.32	7.93	0.46	0.15

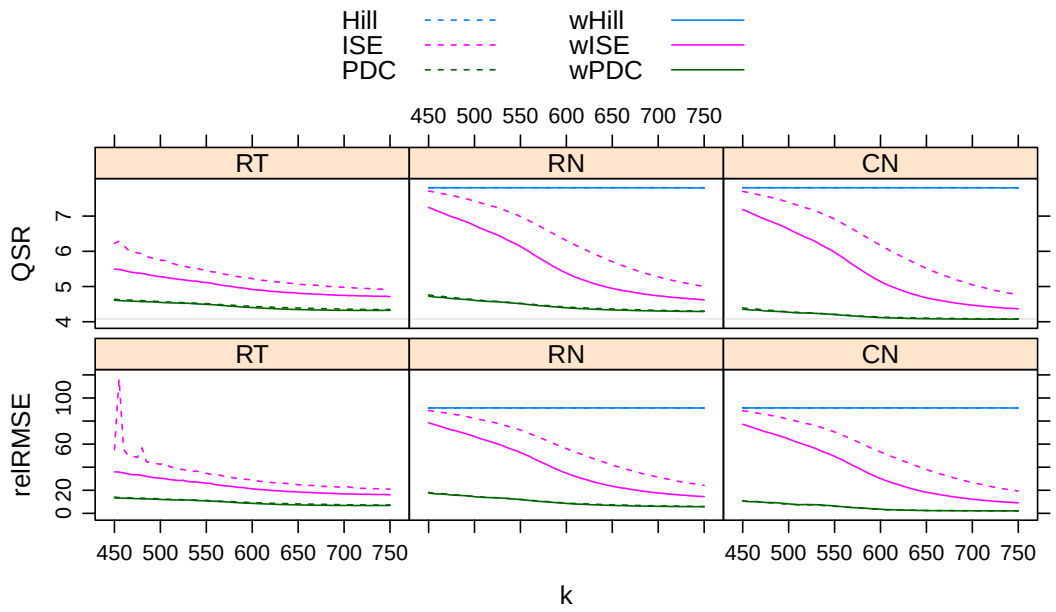


Figure 10.10: Average results (*top*) and RMSE (*bottom*) for the QSR using the three approaches for Pareto tail modeling with k varying between 450 and 750. Scenario: AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.01$.

number of persons in the samples of 6 000 households, this amounts to about 185 to 190 contaminated households. As before, the outliers are generated with an OCAR-CCAR mechanism based on a normal distribution $\mathcal{N}(\mu, \sigma)$ with $\mu = 500\,000$ and $\sigma = 20\,000$.

In Figure 10.10, the average results (*top*) and RMSE (*bottom*) for the semiparametric QSR estimates are shown. Due to the large number of outliers in the upper tail of the distribution, the range of k is limited to 450 to 750, otherwise the plot would be unreadable due to instabilities for the RT approach. In fact, the Hill/wHill estimates for the RT approach had to be omitted completely from the plot because of their instability. Most notably about the plots are the large differences between the weighted and unweighted ISE estimates. In this case, the weighted version performs much better, even though not satisfactory. Nonetheless, the influence of the outliers on the ISE/wISE estimator decreases dramatically with increasing k. The Hill/wHill estimates, on the other hand, show a consistent high influence. Even the PDC/wPDC estimates show a significant bias for lower values of k, but the results are excellent for values of k larger than ≈ 600.

Figure 10.11 contains box plots for the PDC and wPDC estimators for the three approaches for tail modeling with, as well as the standard estimation of the QSR and the parametric method based on a GB2 distribution. In this scenario, $k = 700$ is selected for the semiparametric estimators. Clearly, the standard estimation and the GB2 approach are corrupted by the outliers. Also the RT and RN approaches show a significant positive bias. However, the CN approach leads to excellent results.

The superior performance of the PDC-CN/wPDC-CN estimators—provided k is chosen sufficiently large—is also clearly documented in Table 10.10. It is also worth noting the instabilities for the RT approach with the Hill/wHill and ISE/wISE estimators for $k = 300$.

© http://ameli.surveystatistics.net/ - 2011

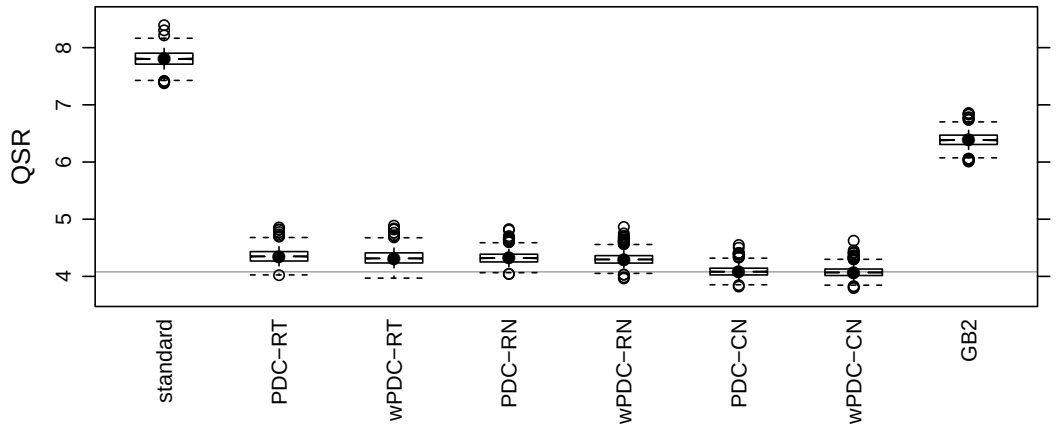


Figure 10.11: Box plots of the simulation results for the QSR using standard estimation, the three semiparametric approaches for Pareto tail modeling with the PDC and wPDC estimators and $k = 700$, as well as parametric estimation using a GB2 distribution with the profile log-likelihood approach. Scenario: AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.01$.

© http://amelia.surveystatistics.net/ - 2011

Table 10.10: Evaluation of the different estimation methods for the QSR at AAT-SILC, stratified Midzuno sampling, and OCAR-CCAR with $\varepsilon = 0.01$.

Method	Type	k	Average	Variance	Bias	relBias	Median	MAD	MedE	RMSE	relRMSE	MedAE	MaxARE
standard			7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wHill	RT	300	31.43	28386.54	27.35	670.43	14.75	5.13	10.67	170.61	4181.65	15.83	1198.82
wHill	RT	500	10.20	34.25	6.12	149.97	9.01	1.37	4.93	8.46	207.48	7.31	25.79
wHill	RT	700	7.75	2.33	3.67	89.87	7.43	0.67	3.35	3.97	97.35	4.96	6.29
wHill	RN	300	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wHill	RN	500	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wHill	RN	700	7.80	0.02	3.72	91.30	7.80	0.14	3.72	3.73	91.36	5.52	1.06
wHill	CN	300	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wHill	CN	500	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wHill	CN	700	7.80	0.02	3.72	91.30	7.80	0.14	3.72	3.73	91.36	5.52	1.06
wISE	RT	300	19.50	1149.06	15.42	378.02	11.99	4.57	7.91	37.23	912.43	11.73	116.51
wISE	RT	500	5.28	0.10	1.20	29.40	5.25	0.30	1.17	1.24	30.41	1.73	1.15
wISE	RT	700	4.75	0.03	0.67	16.42	4.74	0.15	0.66	0.69	16.94	0.98	0.34
wISE	RN	300	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wISE	RN	500	6.73	0.34	2.66	65.08	6.73	0.62	2.65	2.72	66.63	3.92	1.04
wISE	RN	700	4.73	0.09	0.65	16.04	4.68	0.25	0.60	0.72	17.62	0.89	0.54
wISE	CN	300	7.81	0.02	3.73	91.31	7.80	0.14	3.72	3.73	91.37	5.52	1.06
wISE	CN	500	6.63	0.41	2.55	62.49	6.62	0.67	2.54	2.63	64.45	3.77	1.04
wISE	CN	700	4.47	0.10	0.39	9.54	4.40	0.25	0.32	0.50	12.23	0.48	0.51
wPDC	RT	300	5.39	1.98	1.31	32.06	5.06	0.21	0.98	1.92	47.05	1.45	6.71
wPDC	RT	500	4.55	0.02	0.47	11.48	4.53	0.14	0.45	0.49	12.07	0.67	0.34
wPDC	RT	700	4.33	0.02	0.25	6.03	4.31	0.13	0.23	0.28	6.86	0.35	0.20
wPDC	RN	300	6.63	0.51	2.55	62.45	6.36	0.60	2.28	2.65	64.86	3.38	1.04
wPDC	RN	500	4.61	0.07	0.53	12.90	4.54	0.17	0.46	0.59	14.42	0.68	0.64
wPDC	RN	700	4.30	0.01	0.22	5.47	4.30	0.10	0.22	0.25	6.06	0.32	0.19
wPDC	CN	300	6.44	0.68	2.37	57.97	6.14	0.71	2.06	2.51	61.40	3.05	1.04
wPDC	CN	500	4.27	0.08	0.19	4.65	4.19	0.15	0.11	0.34	8.21	0.17	0.62
wPDC	CN	700	4.07	0.01	−0.01	−0.13	4.07	0.08	−0.01	0.09	2.21	0.08	0.13
GB2			6.39	0.02	2.31	56.56	6.38	0.12	2.30	2.31	56.65	3.41	0.68

10.4.4 Conclusions

Clearly, the RT approach introduces too much additional uncertainty and is not recommendable. For RN and CN, in most cases there is not much difference between using the weighted and unweighted estimators for the shape of the Pareto distribution, since the procedures are quite stable in case of slight misspecifications in the Pareto model. Nevertheless, it is advised to use the weighted versions to avoid any unnecessary misspecifications. Furthermore, the RN and CN approaches have very similar behavior in a realistic contamination setting and give excellent results in this case. Both approaches show wide ranges of the number of households used for tail modelling that lead to stable results. Selecting a suitable threshold for tail modelling should therefore not be too difficult in practice. The CN approach is favorable, though, as it does not require random draws from the Pareto model and performs better in the case of heavier contamination. Concerning the estimators of the shape parameter of the Pareto distribution, the PDC/wPDC estimator performs best in the presence of outliers and also performs as well as the other estimators in the case of uncontaminated data. For semiparametric estimation, it is thus recommended to use the wPDC-CN estimator.

10.5 Multivariate outlier detection and imputation

10.5.1 Introduction

For the multivariate outlier detection and imputation methods (MODI), we pursue two evaluation strategies. The first evaluation strategy is concerned with numerical criteria of the outlier-detection and imputation performance. In particular, we report the *average proportion of false negatives* (AVEPFN) and *average proportion of false positives* (AVEPFP) for different parameterizations of outlier-detection methods. The former is a measure of the relative number of undetected outliers, and AVEPFP is the relative number of observations falsely declared as outliers. In addition, we report the total number of declared outliers. The second strategy is to analyze both the effect of multivariate outliers and the effect of MODI methods on a set of Laeken indicators. We consider the following (representative) set of Laeken indicators

- sample mean (Hajek estimator),
- at-risk-of-poverty rate (ARPR),
- relative-risk-poverty gap (RMPG),
- quintile share ratio (QSR),
- Gini coefficient.

This set comprises the primary poverty and income-inequality measures (and the mean as a benchmark indicator). All indicators are computed with the equivalized disposable income.

Population and Sampling Design

We restrict attention to the stratified (NUTS2 regions) simple cluster sampling design (with proportional allocation) based on the AAT-SILC population. The results of all other setups can be obtained from the simulation-run reports. The basic problems of robustness depend only moderately on the sample design. For example in a OAR-CAR contamination where the same variables determining the outlier mechanism are also involved in the sampling designs additional effects may be observed.

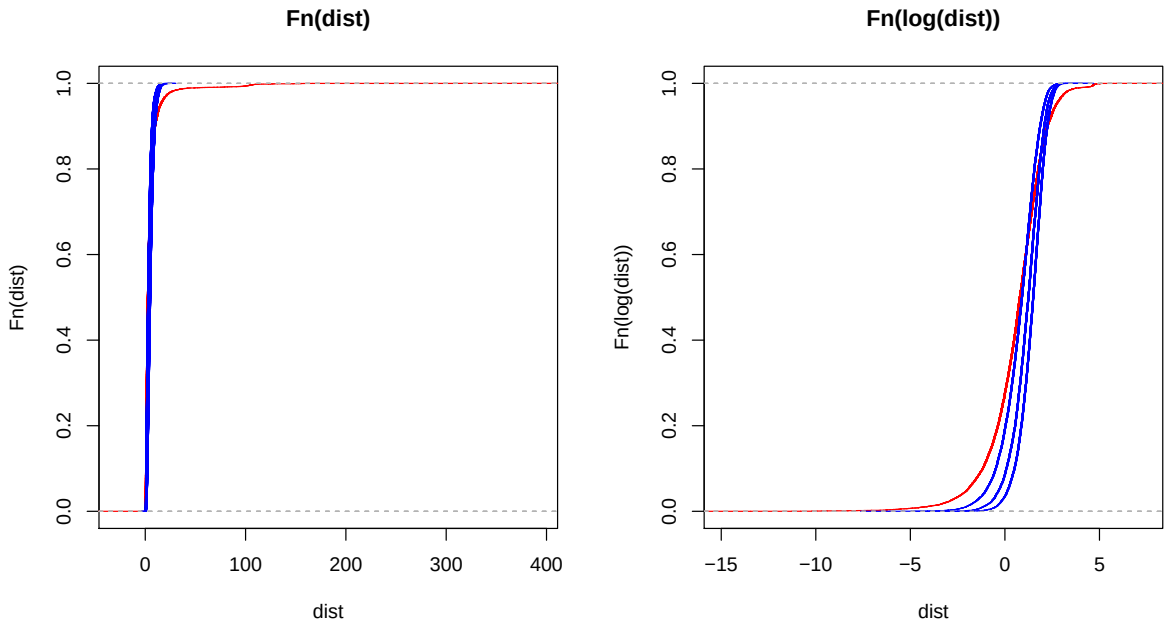


Figure 10.12: Empirical distribution function of the Mahalanobis distances (in original and log scale) for BACON-EEM (red color) and a corresponding χ^2-distribution with 3, 4, and 5 degrees of freedom (blue color).

Variables, Outliers and Missing Values

The (18 household- and 14 individual-level) income components have been aggregated. We use the following four aggregated components: `workinc`, `capinc`, `transh`, and `transp`; see Hulliger et al. (2011b) for more details. The outliers have been generated by means of an OCAR-NCAR mechanism. That is, a proportion ε of the observations is shrinked by$\lambda = 0.2$ and has been displaced to the 95% quantile of the data projected onto the the eigenvector corresponding to the smallest eigenvalue of the original covariance matrix. We study two contamination scenarios: 1% and 5% outliers (this corresponds to an expected number of 127.9 and 641.8 outliers, respectively). Note that the outliers in the components are smeared over the household due to the calculation of the disposable income, which is in fact a redistributed household income. Therefore the effect of outliers in the components on the disposable income is less severe than if the disposable income is contaminated directly as has been the case in the univariate contamination scenarios.

© http://ameli.surveystatistics.net/ - 2011

As missingness mechanism we use missing completely at random (MCAR) with a proportion of 2% in each of the (aggregated) income components. This choice results in a probability of $p = 1 - (1 - 0.02^4)^n = 0.0026$ that at least one observation exhibits missing values in all its components.

All structural zeros, i.e. all zero values in any of the income components, are set to a missing value prior to outlier detection. When it comes to imputation (for outliers and missing observations) those structural zeros are set back to zero.

Simulation

All methods are evaluated on grounds of a design-based simulation study. The numerical criteria are computed (and compared) with respect to the true population charateristics.

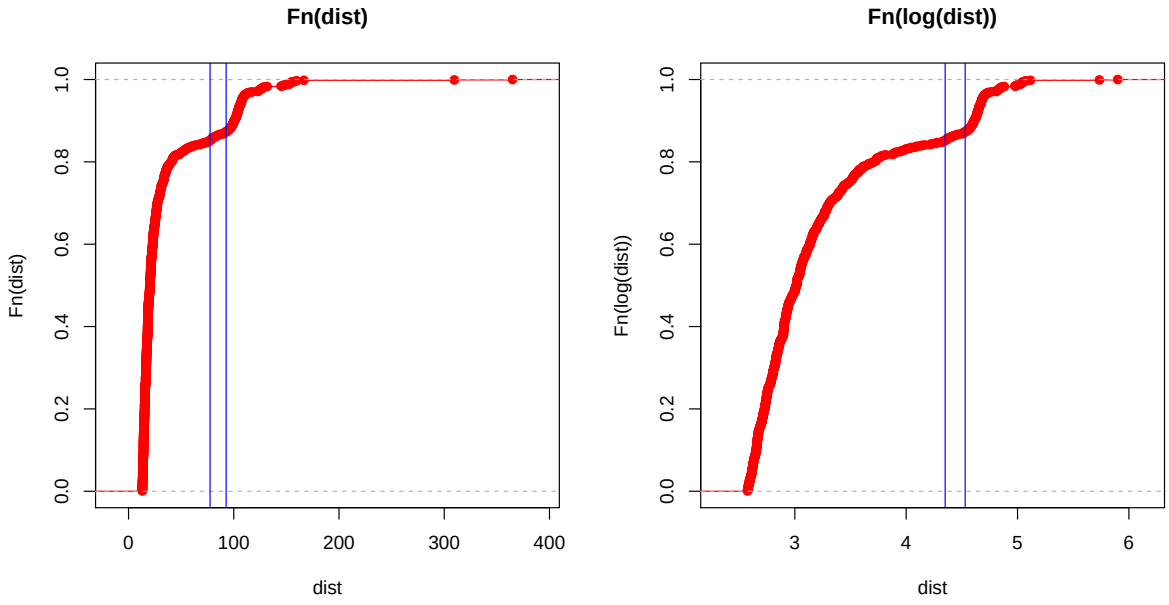
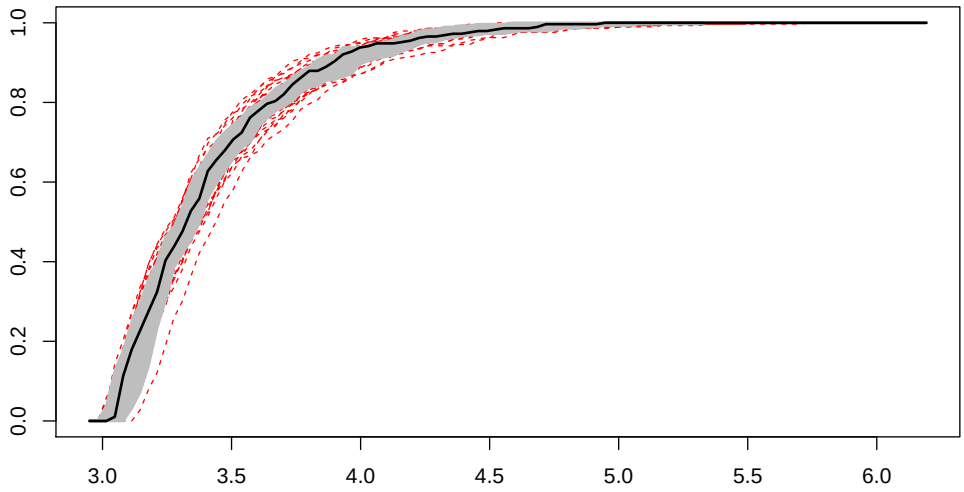


Figure 10.13: Upper tail of the empirical distribution function of the Mahalanobis distances (in original and logscale; see Figure 10.12) for BACON-EEM. The blue vertical lines indicate some (reasonable) cutoff points (i.e., in log-scale 4.34 and 4.53)

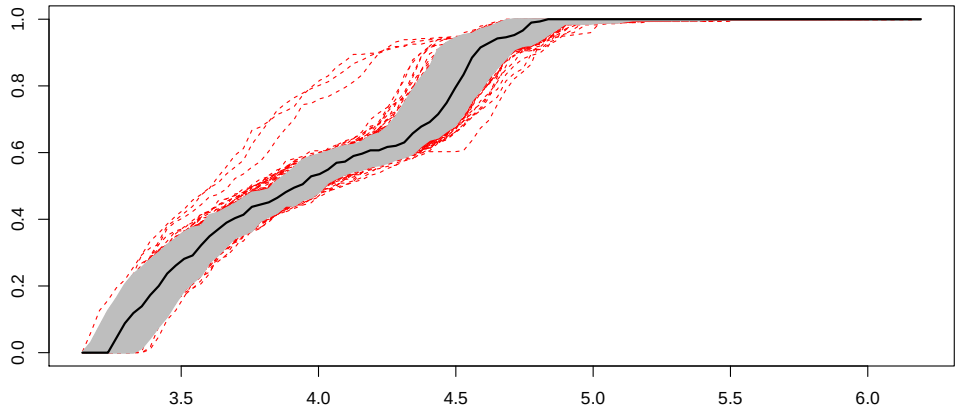
Outlier Detection

In the case of p-variate normally (MVN) distributed data (or in general for elliptically contoured (EC) data), outlier-detection methods that employ identification rules based on the robustly estimated squared Mahalanobis distance (MD) are well known; see e.g., Maronna and Zamar (2002). If the data were MVN, one may choose a threshold of the MD and subsequently declare all points as outliers with MD larger than that particular threshold. The threshold is usually chosen as the α quantile, say, of the χ^2_p distribution, referring to the fact that (asymptotically) $MD_i \sim \chi^2_p$. In the matter at hand, these rules

are of very limited (if any) value because the data on income components are far from being MVN (even after appropriate transformation).



(a) Un-contaminated data



(b) 1% contaminated data

Figure 10.14: Functional boxplot of the upper tail of the ECDF of the squared Mahalanobis distances (for contaminated and uncontaminated data). Each line represents the upper tail of the ECDF of the MD for an application of BACON-EEM on a Monte Carlo sample (here subsample of 500 ECDFs among 1000; MD in log scale; upper tail: 300 largest MD values). The black curve represents the median curve; outlier curves are colored red.

By way of illustration, we depict in Figure 10.12 the empirical cumulative distribution function (ECDF) of the MD for BACON-EEM. It is evident that the ECDF of the MD clearly deviates from a theoretical χ^2 distribution (with d.f. 3,4, or 5) in the upper tail (the same holds for the ECDF in log-scale). As a result, choosing the MD cutoff on grounds of the $\chi^2_p(\alpha)$ leads to thresholds that are far too large. In addition, also the Wilson-Hilferty transform, $(MD_i/p)^{1/3} \sim \mathcal{N}(1 - 2/(9p,\ 2/(9p))$ (cf. Little and Smith, 1987, 61), gives results that are in no way better. Neither transform brings the empirical MD sufficiently close to a theoretical distribution so that a probabilistic argument could automatically be used for outlier detection. Consequently, the analyst has to choose the cutoff point in a case-by-case manner. Notably this means that one has to study the distribution of the

© http://ameli.surveystatistics.net/ - 2011

MD. In Figure 10.13, we depict the upper tail of the ECDF of the MD in ordinary and log scale. As is apparent from the display, the ECDF does not approach the horizontal limit at 1.0 in a smooth manner as the CDF of a corresponding χ^2 distributed r.v. would. The ECDF features a clearly visible bulge. For the ECDF in log-scale (RHS panel in Figure 10.13), we therefore propose to choose the cutoff point in the region (more precisely at the right boundary of the region) where the ECDF still behaves like the CDF of the χ^2 r.v. In the example depicted in Figure 10.13, the chosen threshold would coincide with one of the blue vertical lines – in terms of numbers (in log-scale) 4.35 or 5.43 (note that the curvature of the ECDF before and after the first cutoff is slightly different). As a result, we identify those observations as outliers with a squared Mahalanobis distance larger than the chosen cutoff point.

For outlier-detection methods appealing to the concept of MVN (or EC) data, the choice of tuning constant (denoted α) is based on the aforementioned relation between MD and a theoretical distribution, e.g., χ^2. Referring to the (above) α-quantile argument for outlier detection with the income data at hand, will be uninformative or misleading. These problems pertain to BACON-EEM, GIMCD, and TRC (but not to the Epidemic Algorithm). For these methods, we therefore propose to study the ECDF of the MD. Given the ECDF, one choses an appropriate cutoff based on changes of the curvature the ECDF and the number of declared outliers corresponding to potential cutoffs. For the simulation, studying the ECDF of the MD for each sample separately is impracticable. We therefore rely on the following heuristic. Figure 10.14 shows functional boxplots (Sun and Genton, 2011) of the upper tail of the ECDF of the squared Mahalanobis distances (in log scale and w.r.t. the location estimate of BACON-EEM) for 500 randomly drawn samples (out of 1000; this subsampling results in a slight underestimation of the dispersion of the functions; upper tail:= largest 300 observations). Subfigure a) and b) depict functional boxplots of the uncontaminated and contaminated data, respectively. It is apparent that for uncontaminated data, the ECDF curves are reasoably well behaved (i.e., their behavior is similar to the CDF of a χ^2 r.v. For contaminated data, on the other hand, the median curve (black) for the 500 samples features the bulge at approximately 4.35 (in log scale), too. The heuristic consists of choosing cutoff=4.35 (and the corresponding number of outliers) for all samples. Clearly, this heuristic does not give an optimal cutoff for samples with an outlying MD-ECDF curve (red lines in Figure 10.14). Subsequently the methods are tuned (by means of α) such that they give the right number of declared outliers. Insofar, the methods could be tuned correctly when used for data that conform to the assumed model, i.e., MVN or EC data. When the data do not comply with the underlying model (or only partially), like for income components, we do not modify the methods, but use different heuristics to choose the tuning constants.

10.5.2 BACON-EEM

Suppose the data on the income components have been transformed appropriately for BACON-EEM. In Table 10.12, we report the detection performance of BACON-EEM for three scenarios: uncontaminated data and contaminated data with either 1% or 5% outliers (corresponding to an expected number of 129.7 and 641.8 outliers, respectively). For each scenario, the number of declared outliers (DecOut) and the average proportion (average of the 1000 Monte Carlo replicates) of false negatives (avepfn) and false positives

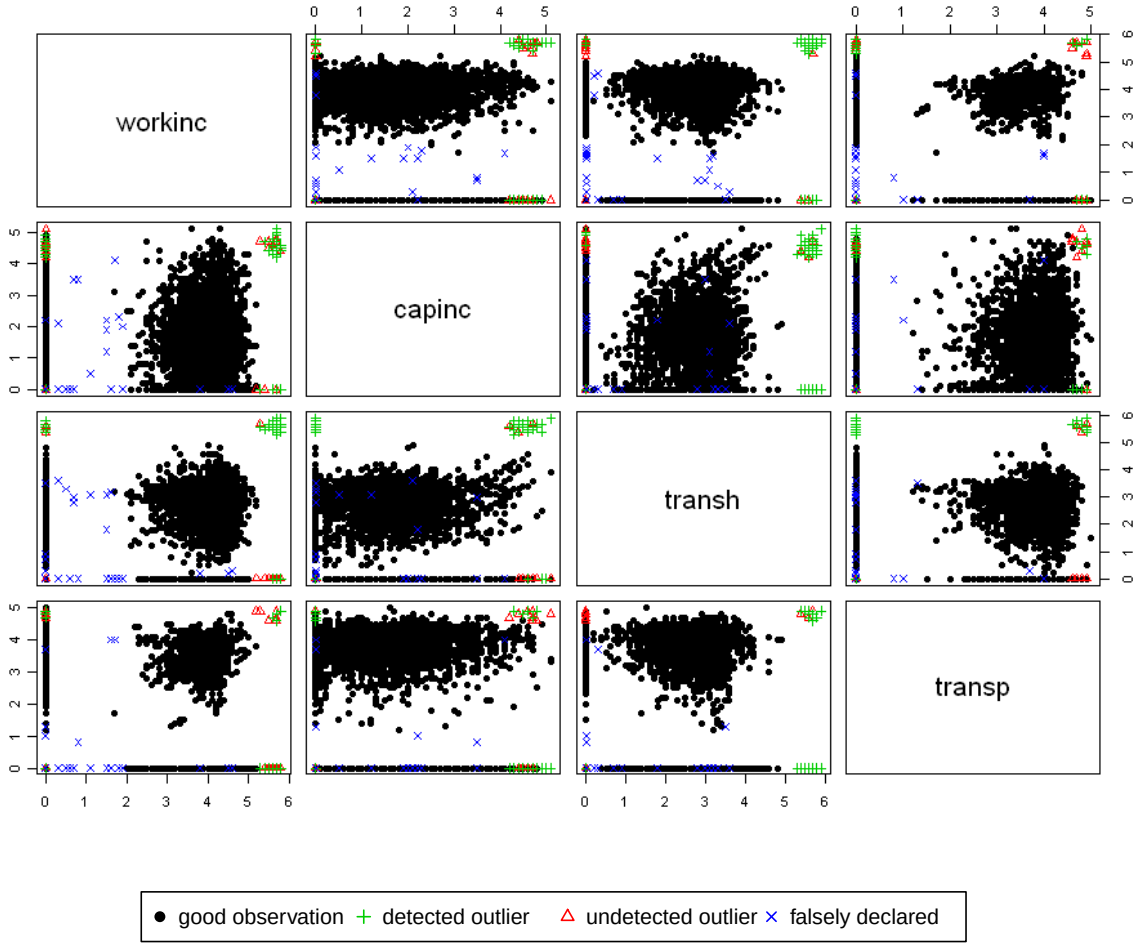


Figure 10.15: Detection performance of BACON-EEM for a single sample.

(avepfp) are shown, corresponding to a particular parametrization (α) of BACON-EEM. It is evident that the smaller α, the smaller is the declared number of outliers. However, the relation between α and DecOut does not relate to the standard α-quantile (of a χ^2 distribution) argument (see discussion above). Moreover, and irrespective of the amount of contamination, the larger DecOut, the larger is also avepfp (the relative number of good observations that have been declared outliers). This shows that BACON-EEM detected representative outliers (of the fairly scattered distribution of good observations). Therefore, if α would be chosen to large (1 in the limiting case), a large share of points (all points) would be declared "representative" outlier. The relationship between α and avepfn (undetected outliers), on the other hand, is reversed, in that larger values of α are associated with smaller numbers of avepfn. Clearly, if α has been chosen such that the declared number of outliers is far below the number of contaminated observations, avepfn is relatively high because BACON-EEM tends not to identify all outliers. Note that in the case of uncontaminated data, avepfn can not be computed because there are no outliers to be detected at all. To summarize, avepfn and avepfp react in opposite direction when moving from a particular choice of α to another. As a result, there is a trade-off between avepfp and avepfn.

The choice of appropriate tuning constants with data whose distribution is neither MVN

© http://ameli.surveystatistics.net/ - 2011

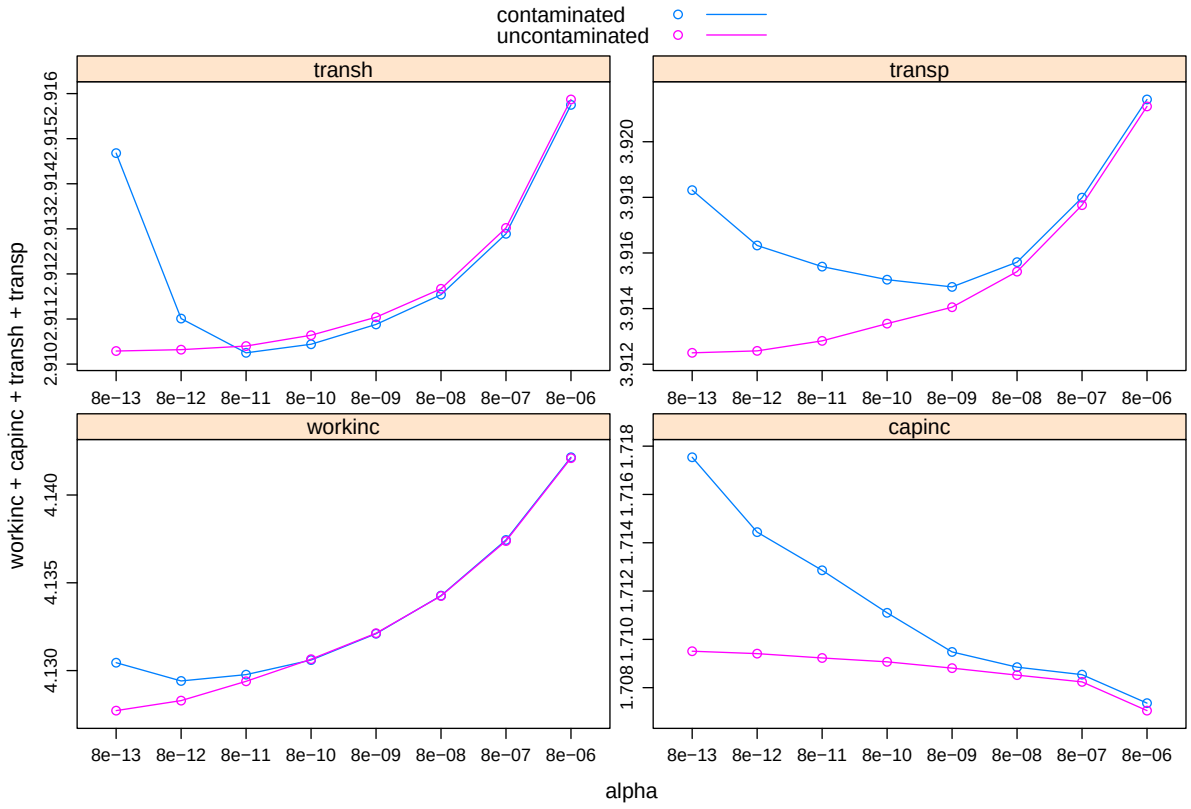


Figure 10.16: Location estimates for BACON-EEM and several tuning constants: contaminated vs. un-contaminated data. Note that the means are computed in log-scale and having set all zeros to NA.

nor EC is very involved. This especially means that we cannot rely on some simple decision rules or heuristics (cf. discussion above). Further evidence concerning this issue is presented in Figure 10.5.2. The plot is a visual display of the location estimates of the four variables in question (in the transformed space) versus α. In addition, we computed the location estimates for both contaminated (1%) and uncontaminated data. Notably, the plots indicate how the location estimates change when α is varied. Moreover, at the particular α where the two lines (representing estimates of the un- and contaminated data) coincide for the location estimates of a variable, one may say the α –kind of – "sweeps off" the effect of contamination for the variable under consideration. The plots clearly show that almost all variables have their specific α value where the lines overlap (viz., e.g., for `workinc` 8E-10 and for `transp` 8E-7). Accordingly, this fact is another indication that no single optimal choice of α exists.

In agreement with the discussion on the tuning constant (above), we propose to choose α in correspondence with the cuttoff on the squared Mahalanobis distance such that the ECDF of MD behaves in the right tail like the CDF of a χ^2 r.v. In the simulation study, the number of observations with MD larger than the cutoff roughly corresponds to the number of generated outliers. Consequently, in case of 1% contamination we would choose α equal to 8E-11, for 5% contamination, we choose 8E-09 (highlighted numbers in Table 10.12). For $\varepsilon = 0.01$ we would have about 0.12% good observations declared as outliers

and about 12.6% of the outliers not detected. However, this fixed choice does not fully reflect the uncertainty of the choice of the tuning constant because even if a bulge is visible in the ECDF and thus a reasonable region for the cutoff can be determined the exact choice cannot be ensured in practice.

Table 10.11: Summary statistics of the number of declared outliers based on the cutoff heuristic (cutoff κ in log scale).

κ	$Q(\#[MD > \kappa]; 0.25)$	$Q(\#[MD > \kappa]; 0.50)$	$Q(\#[MD > \kappa]; 0.75)$	$avg(\#[MD > \kappa])$
3.0	100.0	108.5	114.0	106.1
3.5	91.0	100.0	108.0	97.5
4.0	75.0	90.0	100.0	86.2
4.5	55.7	74.0	90.0	71.6

Notes: MD: squared Mahalanobis distance; $Q(\cdot; p)$ denotes the pth quantile; $\#[< condition >]$ denotes the number of observations that obey the condition.

Table 10.12: Detection performance of the BACON-EEM (for un-contaminated data and 1% and 5% contamination)

	un-contaminated		1% contamination			5% contamination		
α	DecOut	avepfp	DecOut	avepfp	avepfn	DecOut	avepfp	avepfn
8.0E-06	127.7	0.0091	253.4	0.0091	0.008	758.2	0.0091	0.0085
8.0E-07	76.0	0.0054	202.1	0.0054	0.009	708.1	0.0054	0.0096
8.0E-08	48.4	0.0035	174.3	0.0034	0.012	677.5	0.0034	0.0161
8.0E-09	33.4	0.0024	157.1	0.0023	0.03	632.1	0.0023	0.0631
8.0E-10	24.0	0.0017	141.7	0.0016	0.076	551.7	0.0015	0.1719
8.0E-11	17.4	0.0012	128.6	0.0012	0.126	324.2	0.0008	0.5117
8.0E-12	12.4	0.0009	118.3	0.0008	0.167	80.5	0.0004	0.8827
8.0E-13	9.5	0.0007	*104.9*	*0.0006*	*0.25*	10.3	0.0002	0.9890

Notes: DecOut: average number of declared outliers; avepfn/avepfp: average proportion of false negatives/-positives; *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* BACON-EEM with $v = 2$, $C0 = 600$, `em.steps.start=10`, and `em.steps.loops=5`; structural zeros are set to NA.

We next turn to studying the effect of contamination and the effect of MODI methods on a set of Laeken indicators. To start with, we report the average (over the 1000 Monte Carlo experiments) relative bias (in %) of the MODI-methods for the poverty measures (ARPR, RMPG), the income-inequality measures (QSR, Gini), and the mean (benchmark) in the case of uncontaminated data (Table 10.13, panel a) that is associated with a parametrization (α) of BACON-EEM. In addition, we distinguish two imputation methods, ZTOaM and ZGWI. The prefix "Z" of TOaM and WGI indicates that the structural zeros (which have been set to NA prior to outlier detection) have been set back to zero before imputation. Therefore, the imputation methods have only to impute for declared outliers and incomplete observations. The results indicate that for uncontaminated data, the relative bias produced by BACON-EEM and the subsequent imputation method is negligible, whatsoever α or imputation method has been chosen (e.g., for α=8E-13, relbias is at most -0.27% for RMPG).

© http://ameli.surveystatistics.net/ - 2011

Contaminated Data

The presence of contamination alters the picture completely. In Table 10.13, we report the relative bias (relbias) that is associated with a specific parametrization of BACON-EEM followed by either ZTOaM or ZWGI for the scenario with 1% contaminated observations. The findings are as follows.

- First, relbias induced by contamination only (i.e., neither outlier-detection nor imputation; denoted by BACON-EEM with α=0; Table 10.13, panel b)) is different for the poverty and income inequality measures (and the mean). Both ARPR and RMPG are relatively robust with regard to contamination (relbias of 1.8% and -0.1%). The inequality measures, QSR and Gini, are seriously affected since relbias is 251.4% and 122.0%, respectively. The mean is also affected from contamination (relbias 92.7%). Therefore, applying BACON-EEM with α=8E-13 (which is by no way optimal) instead of α=0, yields estimates of the inequality measures that are 8-9 times better in terms of relbias. Recall that in the case of no contamination, BACON-EEM with α=8E-13 produced an absolute relative bias of at most 0.2%. Thus, we argue that if the sample data are supposed to be slightly contaminated, one may gain a lot (in terms of relative bias and relative MSE (relmse); see below) in processing the data by BACON-EEM with a relatively low α (even if the choice of α is not optimal). If the choice of α is poor, we have to apprehend only a relatively small loss in terms of relbias.

- Yet we argued that the loss in relbias is relatively low. The same is true for relrmse since it is strongly dominated by the bias part (the expected sample size of 16000 renders variance considerations negligible; for results on mse/relmse see appendix).

- Besides the issue of how to tune BACON-EEM, some emphasis must be put on choosing the imputation method. For a given α, we report values of relbias for ZTOaM and ZWGI (Table 10.13, panel b; for rmse, see appendix). In contrast to the results for uncontaminated data (ibid., panel a), it is evident that the imputation methods differ in terms of relbias in the case of contaminated data. Although the difference in relbias between ZTOaM and ZWGI depends on α, it is in general higher for ZWGI (for the inequality measures). This is evidence that ZWGI tends to preserve the direction of the contaminated observations too much when imputing. In other words, ZWGI does not down-weight the outliers sufficiently strong so that it ends up with a higher bias than ZTOaM. By way of example, we depict the difference in the point estimates of QSR for the imputation methods in Figure 10.5.2. Note that for α=8E-8 and 8E-9, the bias produced by ZTOaM is negligible (if present at all), whereas for ZWGI, the median point estimate is clearly biased (e.g., for α=8E-11 and ZWGI relbias is 38.4%). In line with these findings, we conclude that the ZTOaM tends to give estimates with a lower relative bias. In other words the strategy to set outliers to missing is better than to winsorized them.

- In the case of 5% contaminated observations (Table 10.13, panel c), the gain in relbias and relmse is accentuated. For the income inequality measures, relbias takes extreme values if the α is too low (viz. relbias=1154.9% for α=8E-13 and ZTOaM). In contrast to the 1% contamination, 5% outlying observations also affect

Table 10.13: Relative bias (in %) of a set of Laeken indicators. *Un-contaminated data*, and *1% and 5% contaminated data* processed by BACON-EEM and two different imputation methods (ZTOaM and ZWGI).

	ZTOaM					ZWGI				
α	mean	ARPR	RMPG	QSR	Gini	mean	ARPR	RMPG	QSR	Gini
Panel a): no contamination										
8.0E-06	1.63	0.04	-1.21	0.65	0.73	-0.33	-0.02	0.27	-0.78	-0.71
8.0E-07	1.12	-0.03	-0.83	0.53	0.60	-0.15	0.10	0.19	-0.31	-0.30
8.0E-08	0.76	0.00	-0.61	0.32	0.39	-0.10	0.15	0.08	-0.20	-0.21
8.0E-09	0.54	0.01	-0.44	0.22	0.28	-0.07	0.16	0.04	-0.14	-0.15
8.0E-10	0.40	0.00	-0.40	0.14	0.19	-0.05	0.15	-0.01	-0.10	-0.12
8.0E-11	0.30	0.01	-0.35	0.07	0.12	-0.03	0.14	-0.04	-0.08	-0.09
8.0E-12	0.23	0.03	-0.32	0.04	0.09	-0.01	0.13	-0.07	-0.06	-0.06
8.0E-13	0.19	0.04	-0.27	0.03	0.07	0.00	0.13	-0.09	-0.04	-0.04
Panel b): 1% contamination										
8.0E-06	3.19	0.76	-1.16	2.20	1.98	3.46	0.86	0.03	6.62	5.66
8.0E-07	2.74	0.76	-0.90	2.23	1.98	4.83	1.04	0.00	10.18	8.73
8.0E-08	2.40	0.78	-0.75	2.14	1.87	6.76	1.27	-0.07	15.22	12.99
8.0E-09	2.41	0.79	-0.62	2.69	2.36	9.29	1.50	-0.04	21.93	18.48
8.0E-10	3.22	0.83	-0.53	5.18	4.63	12.27	1.66	-0.07	29.96	24.78
8.0E-11	4.63	0.86	-0.49	9.23	8.30	15.37	1.75	-0.07	38.34	31.06
8.0E-12	6.55	0.88	-0.42	14.63	12.83	18.49	1.83	-0.10	46.79	37.04
8.0E-13	12.35	1.00	-0.36	30.65	24.21	23.88	1.87	-0.09	61.59	46.16
0	92.79	1.89	-0.10	251.43	122.03	–	–	–	–	–
Panel c): 5% contamination										
8.0E-06	9.2	3.7	0.5	7.9	6.3	18.4	5.0	1.2	34.0	25.5
8.0E-07	8.8	3.7	0.5	8.2	6.5	24.7	5.5	1.6	49.4	35.6
8.0E-08	8.9	3.9	0.6	9.1	7.2	34.5	6.4	2.1	74.0	49.8
8.0E-09	12.8	4.1	0.6	20.1	15.9	50.2	7.5	2.8	114.2	69.5
8.0E-10	36.0	4.7	0.9	81.1	41.7	80.9	8.6	3.1	194.2	93.8
8.0E-11	174.2	6.8	1.9	443.5	105.0	207.9	9.6	3.4	528.5	137.0
8.0E-12	373.6	9.2	3.1	966.5	168.1	384.1	10.0	3.5	994.1	176.5
8.0E-13	445.2	9.9	3.5	1154.9	185.5	446.4	10.0	3.5	1158.7	186.4
0	xxx	xxx	xxx	xxx	xxx	–	–	–	–	–

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* BACON-EEM with $v = 2$, $C0 = 600$, `em.steps.start=10`, and `em.steps.loops=5`; structural zeros are set to NA.

the estimates of the (robust) poverty measures. It is evident that even a moderate robustification yields far better estimates in terms of bias and relmse.

© http://ameli.surveystatistics.net/ - 2011

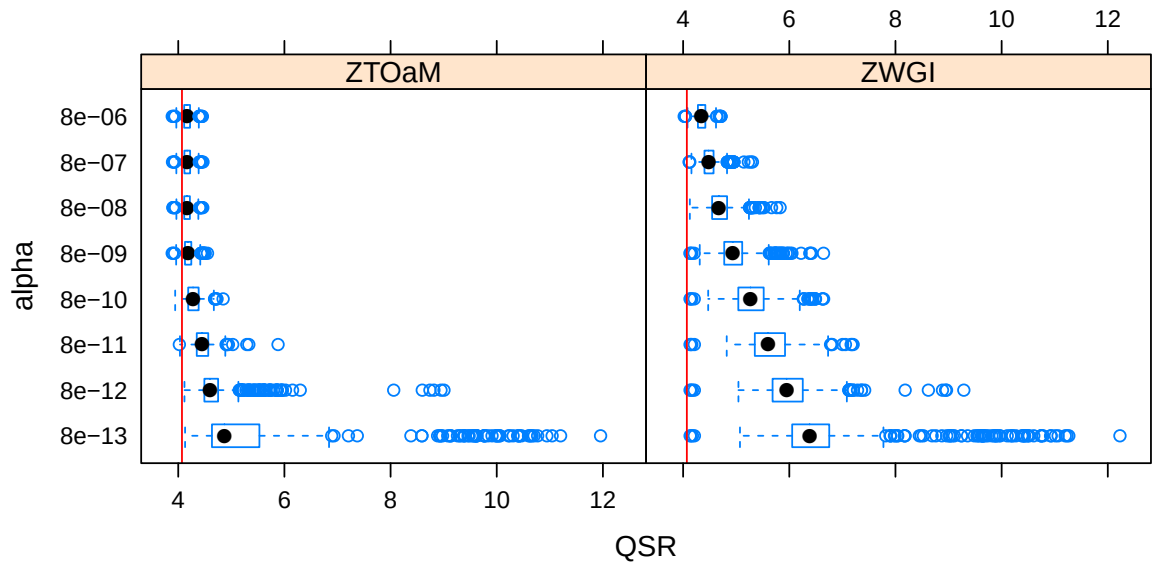


Figure 10.17: Differences in point estimates of QSR (1000 Monte Carlo replications) for two different imputation methods: ZTOaM and ZWGI. Data: 1% contaminated; Method: BACON-EEM; Red line indicates the population value.

Contaminated Data and Missing Values

Besides the effect of contamination on the estimates, we are interested how the estimators behave when the contaminated data feature also missing values. In Table 10.14, we report the numerical criteria of the detection performance (we also report the number of BACON iterations, the final set of good observations, and the MD cutoff associated with α). When comparing the detection performance of BACON-EEM with and without missing values (Tables 10.14 and 10.12) we recognize hardly any difference, except that in the latter case the number of undetected outliers tends to be slightly higher. Overall, the detection performance of BACON-EEM with incomplete data (for the relatively small amount of missing values) is as good as for complete data.

We then turn to studying the effect of contamination and the effect of MODI methods on a set of Laeken indicators in the presence of missing values. In this case the imputation methods, ZTOaM and ZWGI, have to impute for the missing observations and the declared outliers. The results are shown in Table 10.15. In addition, we show boxplots of the point estimates of QSR for ZTOaM vs. ZWGI and complete vs. incomplete data (Figure 10.5.2; the plots for complete data are the same as in Figure 10.5.2 and have been repeated for ease of comparability). For the income-inequality measures, the findings are as follows.

- The results of ZWGI and ZTOaM are very similar in the case of incomplete data. This is in contrast to complete data.
- Irrespective of the imputation method and the choice of α (among the possible values), the relative bias is slightly larger than in the case of no missing values.

For the poverty measures:

- The estimates of the poverty indicators are remarkably biased (Figure 10.5.2).
- Moreover, the relative bias tends to be independent of the choice of α. This is evidence that estimated model (particularly the location) does not perfectly conform with the underlying data. This in turn is no surprise as for income data, an EC model can at most serve as a working model. Fortunately, the relative bias is relatively small (4.1%-5.1%).

Table 10.14: Detection performance of the BACON-EEM (for 1% contamination; 2% component-wise missing values)

α	DecOut	avepfp	avepfn	final_set	niter	cutpoint
8.00E-06	257.7	0.0095	0.0115	13613.3	9.5	29.0
8.00E-07	204.8	0.0056	0.0119	13666.2	8.0	34.0
8.00E-08	175.8	0.0036	0.0157	13695.1	7.4	39.0
8.00E-09	158.0	0.0025	0.0330	13712.9	7.0	43.9
8.00E-10	142.2	0.0018	0.0799	13728.7	6.8	48.8
8.00E-11	128.7	0.0013	0.1314	13742.2	6.6	53.7
8.00E-12	118.1	0.0009	0.1735	13752.9	6.3	58.6
8.00E-13	104.3	0.0007	0.2585	13766.7	6.9	63.4

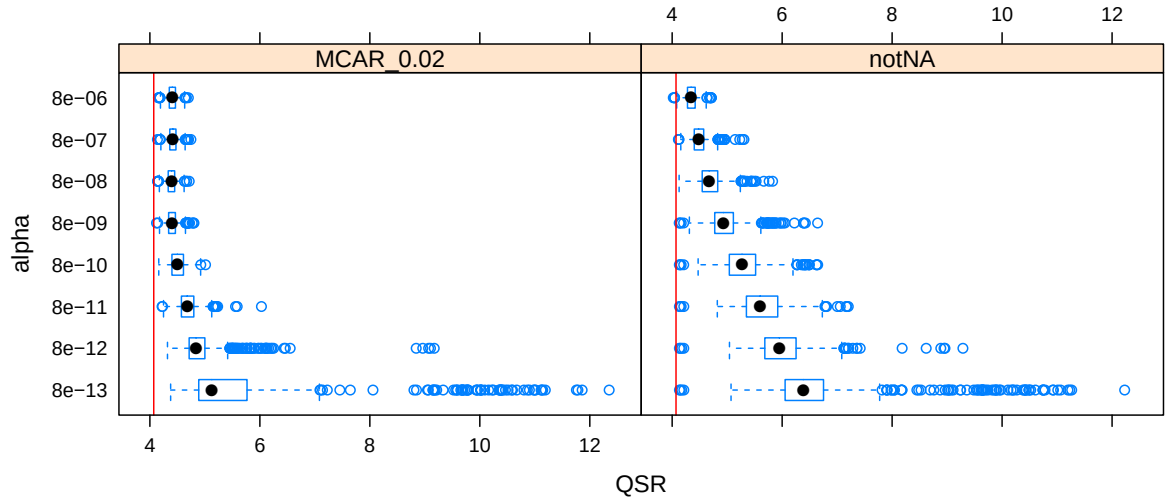
Notes: DecOut: average number of declared outliers; avepfn/avepfp: average proportion of false negatives/-positives; niter: number of BACON iterations; cutpoint: squared MD threshold *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* BACON-EEM with $v = 2$, $C0 = 600$, `em.steps.start=10`, and `em.steps.loops=5`; structural zeros are set to NA.

Table 10.15: Relative bias (in %) of a set of Laeken indicators. *Contaminated data* with *2% component-wise missing observations* processed by BACON-EEM and two different imputation methods (ZTOaM and ZWGI).

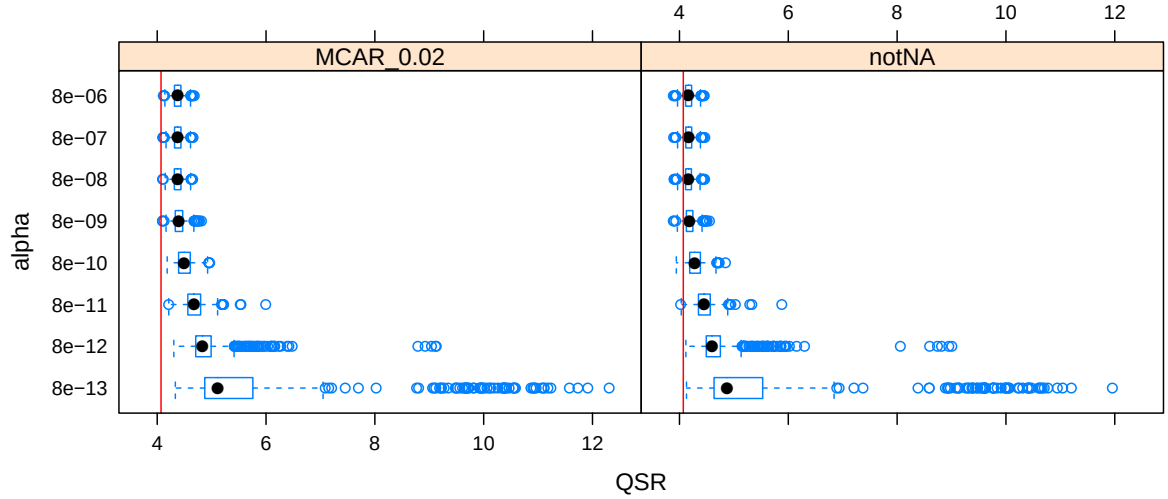
	ZTOaM					ZWGI				
α	mean	ARPR	RMPG	QSR	Gini	mean	ARPR	RMPG	QSR	Gini
8.0E-06	4.48	4.90	4.10	7.31	4.90	4.71	4.85	4.23	8.07	5.51
8.0E-07	4.02	4.88	4.34	7.30	4.87	4.26	4.86	4.37	8.28	5.59
8.0E-08	3.73	4.88	4.53	7.29	4.85	3.85	4.85	4.51	7.71	5.15
8.0E-09	3.75	4.82	4.73	7.86	5.33	3.77	4.81	4.62	7.99	5.41
8.0E-10	4.54	4.87	4.69	10.41	7.55	4.51	4.91	4.73	10.52	7.61
8.0E-11	5.92	5.00	4.86	14.66	11.19	5.94	5.01	4.72	14.85	11.32
8.0E-12	7.74	5.08	4.99	20.12	15.56	7.77	5.10	4.90	20.41	15.75
8.0E-13	13.42	5.25	5.24	36.70	26.71	13.50	5.24	5.15	37.06	26.96

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* BACON-EEM with $v = 2$, $C0 = 600$, `em.steps.start=10`, and `em.steps.loops=5`; structural zeros are set to NA.

© http://ameli.surveystatistics.net/ - 2011



(a) Imputation: ZWGI



(b) Imputaion: ZTOAM

Figure 10.18: Differences in point estimates of QSR (1000 Monte Carlo replications) for two different imputation methods: ZTOaM and ZWGI and complete and incomplete (MCAR_0.02) data. Data: 1% contaminated; Method: BACON-EEM; Red line indicates the population value.

10.5.3 GIMCD

GIMCD is also based on the assumption of MVN (EC) data. Therefore, the difficulties of choosing an appropriate tuning constant (α) pertain here too. As with BACON-EEM the choice of α is based on heuristics adapted from the ECDF of the squared Mahalanobis distances (MD). In Figure 10.20, we show the upper tail of the ECDF of the MD for a subsample of 500 ECDF curves (among the 1000). In contrast to BACON-EEM, the median MD-ECDF curve for GIMCD does not feature a clearly visible bulge. Nonetheless, the slope of the curve changes at approximately $log(MD)$=3.75 its behavior. This point is associated with an average of 114 declared outliers (Table 10.16). It should be stressed

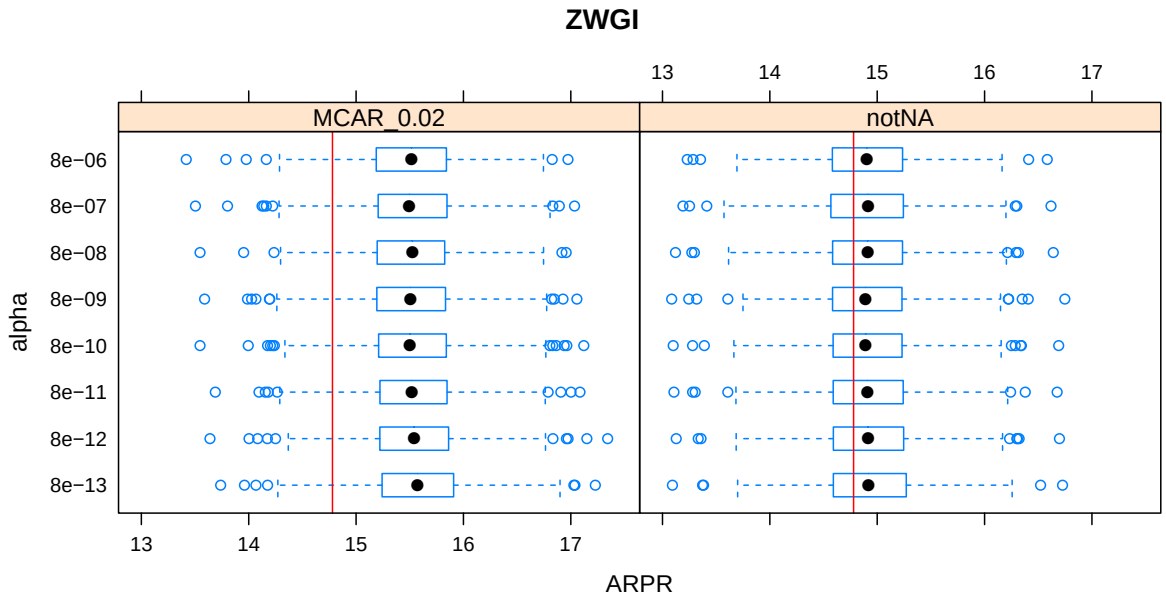


Figure 10.19: Differences in point estimates of ARPR (1000 Monte Carlo replications) between complete and incomplete (MCAR_0.02) data for ZTOaM. Data: 1% contaminated; Method: BACON-EEM; Red line indicates the population value.

that this heuristic should be considered as a rule of thumb (which may far from optimal). Despite the reservations, it may serve as a useful tool.

Table 10.16: Detection performance of GIMCD: summary statistics of the number of declared outliers based on the cutoff heuristic (cutoff κ in log scale).

κ	$Q(\#[MD > \kappa]; 0.25)$	$Q(\#[MD > \kappa]; 0.50)$	$Q(\#[MD > \kappa]; 0.75)$	$avg(\#[MD > \kappa])$
3.25	219.0	229.0	241.0	230.0
3.50	150.0	157.0	163.0	156.7
3.75	108.0	114.0	120.0	114.0
4.00	69.0	79.0	87.0	77.1
4.25	22.0	33.0	41.0	31.2

Notes: MD: squared Mahalanobis distance; $Q(\cdot; p)$ denotes the pth quantile; $\#[< condition >]$ denotes the number of observations that obey the condition.

Contaminated Data

Thoughout the discussion, we will adhere to the above heuristic and the therby obtained number of 114 declared outliers (cf. Table 10.16). The detection performance of GIMCD (for a series of tuning constants, α) in the case of uncontaminated and contaminated data ($\varepsilon = 0.01$) is reported in Table 10.17. From the tabulated values, we deduce that GIMCD with α=8E-7 declared a number of 130 outliers. Note that the α values have been chosen on a equally spaced grid (for ease of simplicity). Accordingly, the values of DecOut, the number of declared outliers associated with the respective α, are on a grid,

© http://ameli.surveystatistics.net/ - 2011

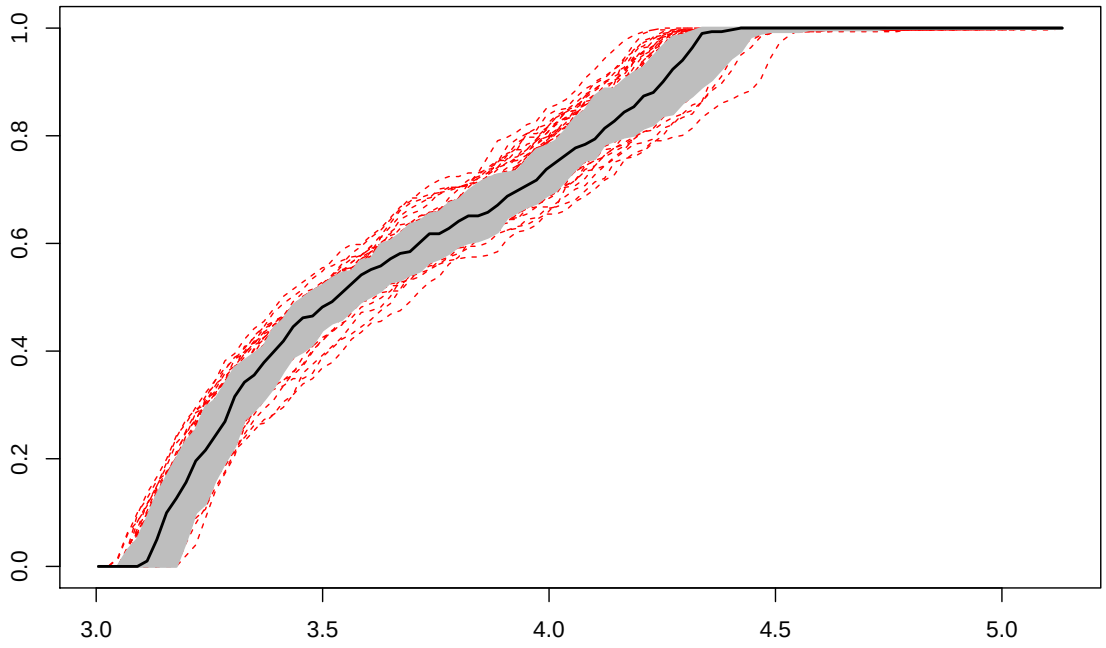


Figure 10.20: Functional boxplot of the upper tail of the ECDF of the squared Mahalanobis distances. Each line represents the upper tail of the ECDF of the MD for an application of GIMCD on a Monte Carlo sample (here subsample of 500 ECDFs among 1000; MD in log scale; upper tail: 300 largest MD values). The black curve represents the median curve; outlier curves are colored red.

too. As a result, there are two choices of alpha, i.e., 8E-8 and 8E-7, that accord with the number of 114 declared outliers (which is due to the heuristical argument). In the presence of (anticipated) contamination, it is safe to choose α=8E-7 because it declares a larger number of outliers – i.e., a conservative choice. In terms of the avepfp-avefpn, α=8E-7 seems to be are reasonable choice. However, avepfn=0.19 is rather high, insofar that almost 20% of the declared outliers are falsely declared outliers. This is also in sharp contrast to the declaration performance of BACON-EEM with avepfn=0.12 on grounds of DecOut=128.6 (cf. Table 10.12). It seems that GIMCD tends to falsely declare a larger number of observations than BACON-EEM does (e.g., GIMCD with DecOut=157.1 features an avepfn of 0.15 (Table 10.17) whereas BACON-EEM with 157.1 exhibits an avepfn of only 0.03 (Table 10.12); avepfp, on the other hand, is similar for both methods).

In passing we note that

Contaminated Data and Missing Values

Besides the effect of contamination on the estimates, we are interested how the estimators behave when the contaminated data feature also missing values. In Table 10.19 we

Table 10.17: Detection performance of the GIMCD (for un-contaminated data and 1% contamination)

	un-contaminated			1% contamination		
α	DecOut	avepfp	avepfn	DecOut	avepfp	avepfn
8.00E-03	574.72900	0.04103	–	661.279	0.03868	0.02755
8.00E-04	229.62800	0.01639	–	325.09	0.01467	0.05035
8.00E-05	112.17700	0.00801	–	210.11	0.00682	0.09805
8.00E-06	61.06600	0.00436	–	158.70	0.00360	0.15005
8.00E-07	34.97500	0.00250	–	130.83	0.00198	0.19223
8.00E-08	20.77800	0.00148	–	109.32	0.00113	0.26847
8.00E-09	12.89200	0.00092	–	94.02	0.00068	0.33957
8.00E-10	8.09200	0.00058	–	77.03	0.00042	0.44329
8.00E-11	5.10900	0.00036	–	58.88	0.00026	0.56825
8.00E-12	3.26800	0.00023	–	43.66	0.00017	0.67698
8.00E-13	2.23600	0.00016	–	28.74	0.00011	0.78792

Notes: DecOut: average number of declared outliers; avepfn/avepfp: average proportion of false negatives/positives; *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* GIMCD, structural zeros are set to NA.

report the numerical criteria of the detection performance. With regard to the scenario of complete data, GIMCD behaves almost the same with 2% component-wise missing values.

We then turn to studying the effect of contamination and the effect of GIMCD on a set of Laeken indicators in the presence of missing values. In this case the imputation methods, ZTOaM and ZWGI, have to impute for the missing observations and the declared outliers. The results are shown in Table 10.20.

The findings can be summarized as follows.

- The bias of all indicator in the set of Laken indicators is slightly larger than in the case of complete data.
- Income inequality measures: The results of ZWGI and ZTOaM are very similar in the case of incomplete data. This is in contrast to complete data.
- Poverty indicators: The estimates of the poverty indicators are remarkably biased. Moreover, the relative bias tends to be independent of the choice of α. This is evidence that estimated model (particularly the location) does not perfectly conform with the underlying data. This in turn is no surprise as for income data, an EC model can at most serve as a working model. Fortunately, the relative bias is relatively small.
- Overall, the results of GIMCD with incomplete data are very similar to those of the BACON-EEM with the same data.

© http://ameli.surveystatistics.net/ - 2011

Table 10.18: GIMCD: relative bias (in %) of a set of Laeken indicators. *Un-contaminated data*, and *1%* and *5% contaminated data* processed by GIMCD and two different imputation methods (ZTOaM and ZWGI).

	ZTOaM					ZWGI				
α	mean	ARPR	RMPG	QSR	Gini	mean	ARPR	RMPG	QSR	Gini
Panel a): no contamination										
8.0E-13	0.05	0.11	-0.13	-0.03	-0.02	0.01	0.13	-0.10	-0.03	-0.04
8.0E-12	0.06	0.12	-0.15	-0.02	-0.01	0.00	0.13	-0.09	-0.04	-0.04
8.0E-11	0.09	0.12	-0.15	0.00	0.00	0.00	0.13	-0.09	-0.04	-0.04
8.0E-10	0.14	0.13	-0.18	0.03	0.04	0.00	0.13	-0.09	-0.04	-0.04
8.0E-09	0.20	0.12	-0.21	0.08	0.08	0.00	0.14	-0.09	-0.04	-0.05
8.0E-08	0.32	0.12	-0.24	0.18	0.18	-0.01	0.13	-0.08	-0.05	-0.06
8.0E-07	0.51	0.15	-0.31	0.39	0.38	-0.03	0.11	-0.05	-0.08	-0.08
8.0E-06	0.81	0.19	-0.41	0.67	0.62	-0.07	0.06	-0.03	-0.17	-0.15
8.0E-05	1.34	0.34	-0.63	1.09	0.99	-0.19	-0.08	-0.03	-0.44	-0.37
8.0E-04	2.50	0.63	-0.90	2.12	1.87	-0.51	-0.43	-0.07	-1.33	-1.11
8.0E-03	5.41	1.27	-1.18	4.46	3.83	-1.71	-2.02	-1.04	-4.98	-4.09
Panel b): 1% contamination										
8.0E-13	57.73	1.61	-0.17	155.23	93.33	67.32	1.90	-0.10	181.24	102.35
8.0E-12	44.51	1.47	-0.20	118.97	78.26	56.16	1.90	-0.10	150.49	91.03
8.0E-11	34.11	1.37	-0.27	90.43	64.18	46.17	1.89	-0.10	122.97	79.46
8.0E-10	24.36	1.24	-0.27	63.65	48.81	35.00	1.85	-0.11	92.25	64.59
8.0E-09	17.52	1.11	-0.30	44.77	36.45	25.52	1.66	-0.12	66.31	49.96
8.0E-08	12.52	1.09	-0.39	30.90	26.25	18.63	1.48	-0.15	47.51	37.82
8.0E-07	7.55	1.01	-0.42	17.01	15.03	11.91	1.24	-0.21	29.25	24.51
8.0E-06	5.73	1.02	-0.52	11.53	10.31	8.37	1.07	-0.25	19.77	16.96
8.0E-05	4.46	1.13	-0.59	7.13	6.31	5.29	0.90	-0.23	11.54	10.01
8.0E-04	4.60	1.35	-0.74	5.48	4.72	2.91	0.53	-0.27	5.26	4.58
8.0E-03	7.60	2.13	-0.75	8.02	6.74	0.50	-1.08	-1.13	-1.45	-0.99

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* GIMCD; structural zeros are set to NA.

10.5.4 Epidemic Algorithm

In contrast to BACON-EEM and GIMCD the Epidemic Algorithm (EA) is not based on a distributional assumption (or specific data model) such as multivariate normality. Therefore, the data on the income components do not necessarily have to be transformed prior to the outlier-detection phase. In addition, the algorithm can cope with the structural zeros.

Table 10.19: Detection performance of GIMCD: *1% contaminated data; 2% component-wise missing values*

α	DecOut	GenOut	avepfp	avepfn
8.00E-13	24.943	127.978	0.00011	0.81738
8.00E-12	39.622	127.978	0.00017	0.70843
8.00E-11	54.222	127.978	0.00026	0.60419
8.00E-10	72.036	127.978	0.00041	0.48183
8.00E-09	90.216	127.978	0.00067	0.3681
8.00E-08	105.634	127.978	0.00107	0.29111
8.00E-07	127.029	127.978	0.00187	0.21049
8.00E-06	154.582	127.978	0.00342	0.16351

Data: AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* GIMCD; structural zeros are set to NA.

Table 10.20: GIMCD: relative bias (in %) of a set of Laeken indicators: *1% contaminated data; 2% component-wise missing values*, processed by GIMCD and two different imputation methods (ZTOaM and ZWGI).

	ZTOaM					ZWGI				
α	mean	ARPR	RMPG	QSR	Gini	mean	ARPR	RMPG	QSR	Gini
8.0E-13	61.16	5.58	4.85	170.28	96.95	65.33	5.71	4.99	181.92	100.71
8.0E-12	47.68	5.41	4.84	132.45	82.26	51.05	5.60	4.94	141.88	85.84
8.0E-11	37.28	5.36	4.78	103.32	68.87	39.18	5.44	4.87	108.58	71.18
8.0E-10	27.38	5.18	4.84	75.54	54.04	27.98	5.26	4.85	77.29	54.93
8.0E-09	19.65	5.06	4.91	53.86	40.74	19.86	5.07	4.85	54.55	41.16
8.0E-08	14.32	4.95	4.82	38.72	30.27	14.43	4.94	5.00	39.14	30.55
8.0E-07	8.99	4.92	4.79	23.50	18.63	9.01	4.91	4.79	23.62	18.73
8.0E-06	6.89	4.96	4.65	17.05	13.33	6.87	4.81	4.62	17.00	13.32

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* GIMCD; structural zeros are set to NA.

EA Detection

The data on the four income components, `workinc`, `capinc`, `transp`, and `transh` are directly processed by EA. All computations have been done with the following specifications of EA: `EAdet(data, weights, reach = "max", transmission.function = "root", power = ncol(data), distance.type = "euclidean", global.distances = F, maxl = 5, prob.quantile = 0.9, random.start = F, threshold = F, deterministic = TRUE)`; all other tuning constants are set to their default values. Two issues are important to note. First, the root transmission function leads to a relatively good spreaded/dispersed distribution of infection times, which is an important characteristic in order to study the outlier-declaration behavior. If the infection times of the majority of the data were very similar (if not identical), it would be very difficult to distinguish potential outliers from ordinary observations. Second, `power` was set to number of variables—here, four. Though, setting power to three gives very similar results.

© http://ameli.surveystatistics.net/ - 2011

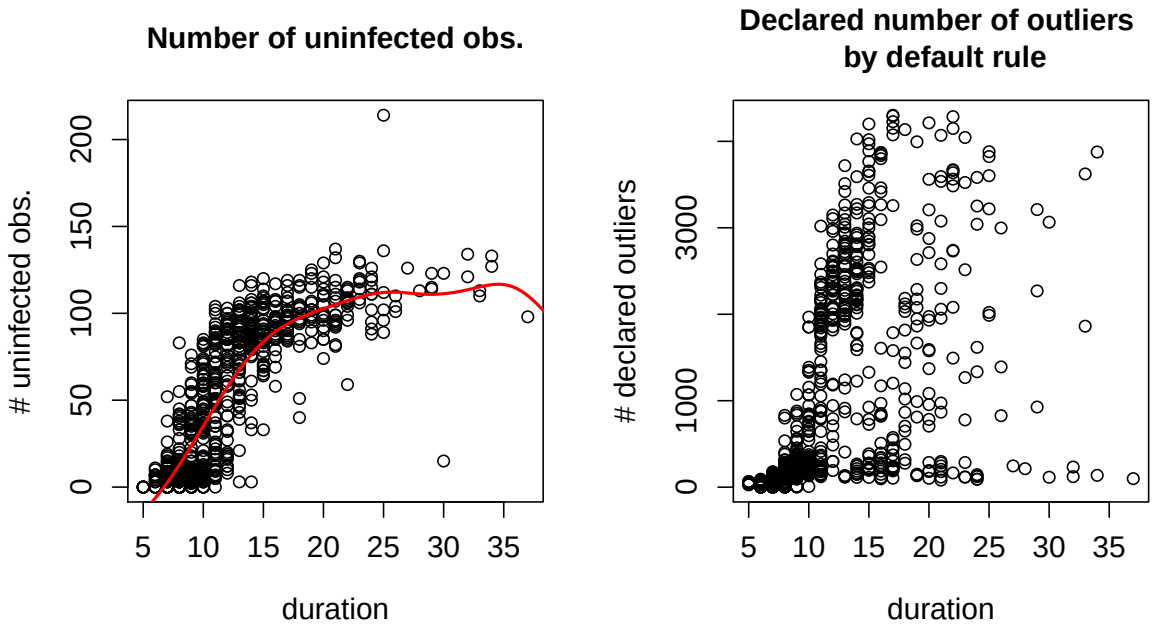


Figure 10.21: Detection performance of EA: number of uninfected observations and number of declared outliers (using the default rule) versus duration. (red line shows a local polynomial fit; bandwidth selection: smooth cross-validation criterion).

In general, EA cannot be used in an out-of-the-box, i.e., rather automatic, manner. It requires the user to study certain diagnostics and to tune the method accordingly. As a result, it is extremely difficult to tune EA in a simulation study such that the results of particular simulation runs can be aggregated to yield a consistent overview. Notably, we encounter rather high variability among the numerical criteria of the detection performance (Figure 10.21). In view of the high degree of variation among the simulation runs, we decided to study the detection and imputation properties separately.

The findings of the EA detection step can be summarized as follows.

- The default outlier-detection rule of EA does not work for the income-component data (Figure 10.21). This method declares between 0 and more than 4000 outliers (observe that we generated on average 127.9 outliers, with 1.19 std. dev.). Moreover, the number of declared outliers tends to be almost independent from the duration of the epidemic.
- Declaring the uninfected observations as outliers, on the other hand, seems to be a sensible outlier-declaration rule (Figure 10.21), given that the duration of the epidemic was sufficiently long (e.g., duration larger than 15 or 20).
- The duration of the epidemic is crucial in order to declare outliers properly. In Figures 10.22 and 10.23 we show the average proportion of false negative (avepfn) and false positives (avepfp) – i.e., undetected outliers and falsely declared outliers – versus duration and the number of declared outliers, respectively. It is evident

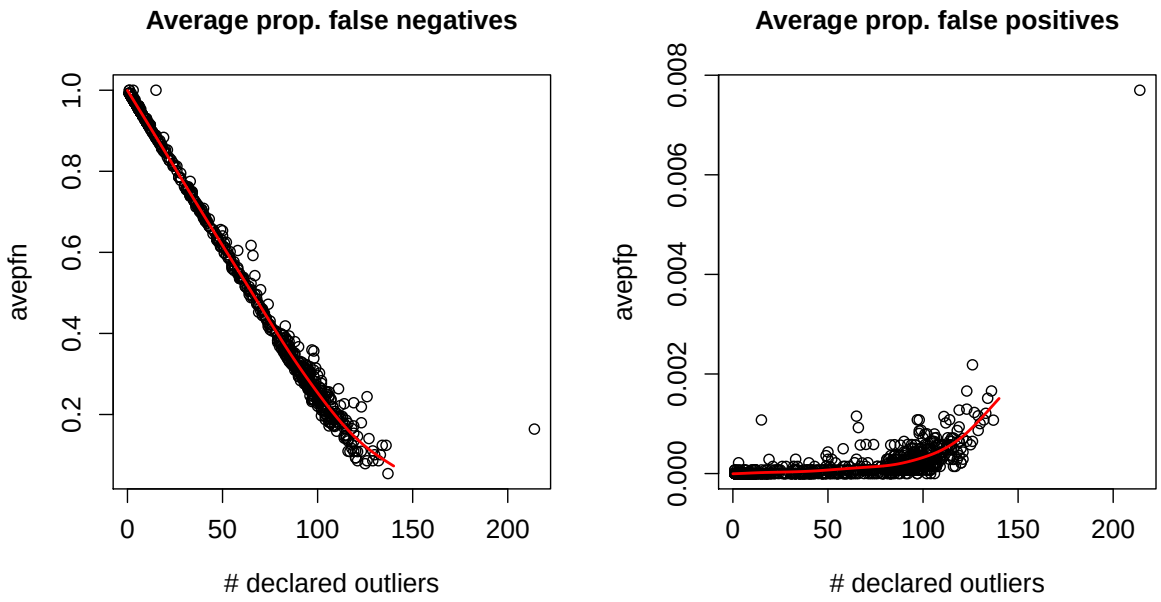


Figure 10.22: Detection performance of EA: avepfp and avepfn versus declared outliers. (red line shows a local polynomial fit; bandwidth selection: smooth cross-validation criterion; fit is restricted to a range of 0–150 declared outliers.)

that the larger the number of declared outliers, the smaller the average proportion of undetected outliers (avepfn; and vice versa). In contrast, the larger the average number of declared outliers, the larger the number of falsely declared outliers (avepfp). Thus, there is a trade-off. The situation is clarified, when we consider avepfn and avepfp contrasted with the duration of the epidemic (Figure 10.23): Up to a duration of approx. 22, avepfn decreases considerably, whereas the increase of avepfp is moderate. As a result, EA works best for a duration between 20 and 25.

Reverse EA Imputation

In view of the high degree of variation among the simulation runs, we decided to study the detection and imputation properties separately. We therefore modified the EA-detection step so that it produces a fixed number of declared outliers. In this respect, the vector of infection times for all observations was sorted in ascending order, and the largest g observations are then declared outliers (where g can take the (arbitrarily chosen) value 100, 130, 140, 160 or 200). Essentially, this modification impedes that the detection-related variability carries over to the final estimates of the Laeken indicators. Consequently, the Monte Carlo distribution of the estimates is determined by the imputation step.

Given the fixed number of declared outliers, g, the Reverse Epidemic Algorithm (REA) is used to impute for the declared outliers. `EAimp(data, weights, outind=EAdet.i$outind, duration = EAdet.r$duration, maxl = 5, kdon = 1`; all other arguments are set to

© http://ameli.surveystatistics.net/ - 2011

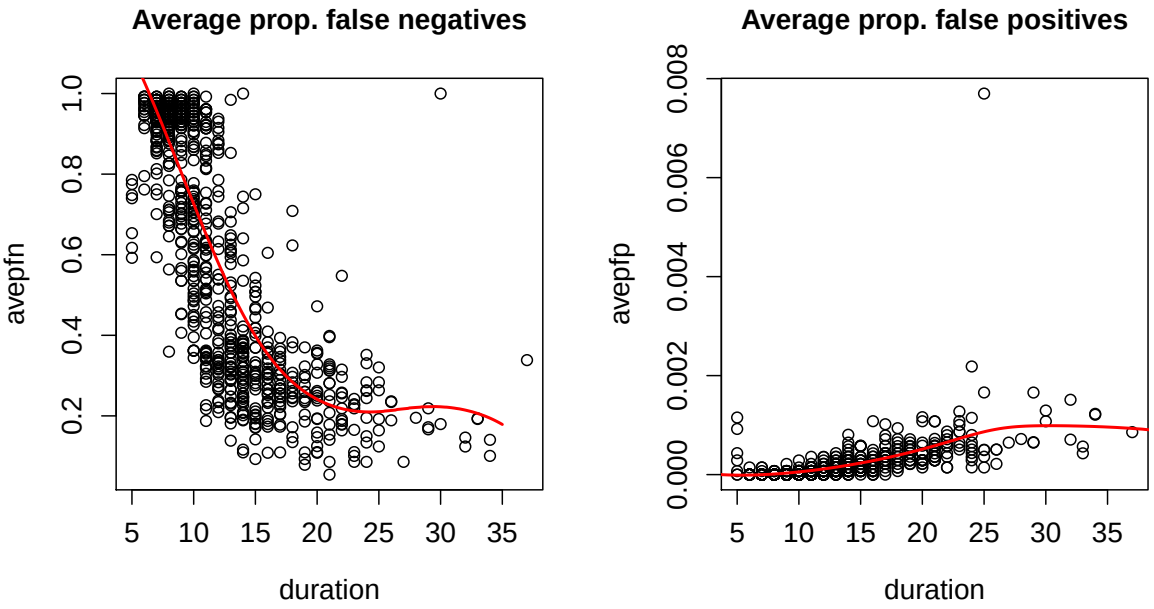


Figure 10.23: Detection performance of EA: avepfp and avepfn versus duration. (red line shows a local polynomial fit; bandwidth selection: smooth cross-validation criterion).

their default value. Observe that the duration of the reverse epidemic is the same as the one of the detection phase.

The relative bias (in %) of the set of Laeken indicators, computed on grounds of the data processed by REA imputation, are reported in Table 10.21. Note that the root mean squared error is of similar magnitude that the bias, because the former is dominated by the bias term. For this simulation, an average number of 127.9 outliers (std. dev. 1.19) has been generated. The results are qualitatively similar to those of BACON-EEM and GIMCD, insofar that the estimates of the poverty indicators have almost zero bias, and the income inequality measures are severely biased. In contrast to BACON-EEM and GIMCD, the bias of the income inequality measures is extremely high, even in the case of 200 declared outliers (DecOut). Though, it might be possible that the bias can be reduced when increasing the DecOut. On the other hand, it would be questionable how to motivate the choice of such a large number of declared outliers. Overall, the tuning of REA is extremely difficult.

Table 10.21: Reverse Epidemic Algorithm: relative bias (in %) of a set of Laeken indicators: *1% contaminated data*.

DecOut	Mean		QSR		Gini		ARPR		RMPG	
	avg	med	avg	med	avg	med	avg	med	avg	med
100	52.2	50.4	141.9	136.7	88.6	80.4	1.5	1.5	0.05	0.20
130	30.9	37.7	84.2	102.0	60.8	63.1	1.3	1.4	0.01	0.25
140	27.3	35.8	72.4	97.0	55.0	60.5	1.3	1.4	0.03	0.28
160	22.6	32.7	60.8	88.4	47.2	56.1	1.3	1.4	0.04	0.30
200	19.1	28.5	50.8	77.1	40.5	50.1	1.3	1.4	0.07	0.38

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* EA detection with a fixed number of outliers, DecOut, and subsequent imputation by Reverse Epidemic Algorithm.

© http://ameli.surveystatistics.net/ - 2011

Table 10.22: BACON-EEM: Root MSE (in %) of a set of Laeken indicators. *Uncontaminated data*, and *1%* and *5% contaminated data* processed by BACON-EEM and two different imputation methods (ZTOaM and ZWGI).

	ZTOaM					ZWGI				
α	mean	ARPR	RMPG	QSR	Gini	mean	ARPR	RMPG	QSR	Gini
Panel a): no contamination										
8E-06	368.2	0.47	1.01	0.09	0.43	167.71	0.48	1.06	0.08	0.40
8E-07	278.7	0.48	1.01	0.08	0.41	155.97	0.48	1.07	0.08	0.36
8E-08	222.1	0.47	1.03	0.08	0.39	154.47	0.48	1.06	0.08	0.35
8E-09	192.9	0.47	1.03	0.08	0.37	153.70	0.48	1.06	0.08	0.35
8E-10	176.5	0.47	1.04	0.08	0.37	153.32	0.48	1.06	0.08	0.35
8E-11	166.9	0.47	1.04	0.08	0.35	153.16	0.48	1.06	0.08	0.35
8E-12	162.0	0.47	1.05	0.08	0.35	153.27	0.48	1.06	0.08	0.35
8E-13	159.0	0.47	1.05	0.08	0.35	152.87	0.48	1.06	0.08	0.35
Panel b): 1% contamination										
8E-06	664.8	0.49	0.98	0.12	0.66	727.98	0.50	1.00	0.29	1.62
8E-07	576.6	0.49	0.97	0.12	0.66	1011.86	0.50	1.00	0.44	2.47
8E-08	511.6	0.49	0.98	0.12	0.63	1419.52	0.52	0.99	0.66	3.68
8E-09	517.3	0.50	0.99	0.14	0.78	1946.49	0.53	0.99	0.94	5.21
8E-10	693.0	0.50	0.98	0.25	1.45	2562.83	0.54	0.98	1.28	6.93
8E-11	979.2	0.50	0.99	0.41	2.42	3185.44	0.55	0.98	1.62	8.60
8E-12	1516.4	0.50	0.99	0.73	3.99	3821.74	0.56	0.98	1.97	10.21
8E-13	3358.8	0.51	0.99	1.78	8.30	5127.88	0.56	0.98	2.71	12.97
Panel c): 5% contamination										
8E-06	1876.2	0.74	0.94	0.34	1.75	3761.09	0.90	1.02	1.41	6.97
8E-07	1796.0	0.75	0.94	0.35	1.81	5078.47	0.96	1.02	2.08	9.80
8E-08	1820.6	0.76	0.96	0.41	2.12	7149.38	1.08	1.07	3.13	13.76
8E-09	2846.6	0.78	0.99	1.03	5.34	10372.07	1.23	1.13	4.81	19.10
8E-10	14113.6	0.87	1.01	7.21	14.29	19306.83	1.38	1.16	9.66	25.99
8E-11	49452.9	1.20	1.10	25.87	33.28	51121.59	1.51	1.19	26.69	38.41
8E-12	81496.6	1.49	1.17	42.83	46.93	81882.37	1.57	1.20	43.06	48.14
8E-13	92269.5	1.56	1.19	48.55	50.34	92300.21	1.57	1.20	48.61	50.47

Notes: ARPR: at-risk-of-poverty rate; RMPG: relative median poverty gap; QSR: quintile share ratio; Gini: Gini coefficient. *Data:* AAT-SILC; stratified simple cluster sample, 6000 households; 1000 Monte Carlo replications; *Method:* BACON-EEM with $v = 2$, $C0 = 600$, `em.steps.start=10`, and `em.steps.loops=5`; structural zeros are set to NA.

10.6 Recommendations

Means: Even if no outliers are present in the data the very skew distribution of income favours a very light robustification over the non-robust classical estimators. A robustified Horvitz-Thompson (RHT) estimator with a tuning constnat of $k = 6$ seems to be a good candidate with low bias if the data contains no outliers and with at least a minimal protection against some rare outliers. A trimmed mean (TM) with light trimming of 0.5% of large observations is similar to a robustified Horvitz-Thompson estimator. The RHT has the advantage that it does not downweight any observation in case of very well behaved data, while the TM always downweights the specified proportion of the data.

Quintile Share Ratio: The quintile share ratio should always be estimated with a robust estimator. The non-parametric SQSR estimator with a very slight trimming above, say some 0.5% and a bias compensation in the lower quintile of similar magnitude or roughly double the upper trimming proportion, seems to be versatile, robust and sufficiently efficient over a range of mild contamination rates. If contamination is larger then the choice of the trimming proportion becomes more difficult. In any case, before fixing a trimming proportion, several choices should be evaluated.

If the tail of the income distribution can be approximated with a Pareto distribution, which is the case for the AMELIA and AAT-SILC simulation universes, a semi-parametric robustification is promising. Replacement of non-representative outliers (RN) with an additional calibration are the best versions. The choice of the tuning constant seems to be less critical than for the non-parametric estimators, however at the price of a more complex procedure.

Multivariate outliers: The multivariate non-elliptical distribution of the income components makes it very difficult to detect and impute multivariate outliers. Nevertheless this is necessary when the structure of income must be investigated more closely. The pre- and post-treatment of the data is crucial for the methods to work. In particular the components must be aggregated or segmented such that the detection and imputation can be carried out in four, five, maybe up to eight or ten dimensions but not for all original variables together. Setting the zero values to missing is a possible way of treatment if the subsequent algorithms can cope with many missing values. The BACON-EEM algorithm for outlier detection is remarkably stable. A subsequent imputation with the same multivariate model as underlying the outlier detection proved to be feasible and with good results. Non-parametric methods like the Epidemic Algorithm proved to be complex in their handling and are rarely better than the BACON-EEM with Gaussian imputation.

The default choice of tuning constants of the methods studied often gives poor results. This is mainly a problem for simulations, where no visual inspection of the distribution of the Mahalanobis distances or of infection times is possible. In an application several tuning constants would be tested and visual inspection of distribution plots would be used to decide on the cut-point for outlyingness.

Variance estimation: Univariate robust estimators allow for a decent variance estimator. However, with complex designs the variance estimators may overestimate the true variance rather heavily.

© http://ameli.surveystatistics.net/ - 2011

Variance estimation for data which has undergone multivariate outlier detection and imputation as well as subsequent reaggregation into disposable income followed by classical estimators of Laeken indicators might be possible with resampling techniques. However their calculation is very complex and the costs seem prohibitive for the moment at least for routine application. It is nevertheless recommended to investigate with simulation the impact on the variance of estimators of multivariate outlier detection and imputation, and in fact of any editing and imputation of income components.

Bibliography

Graf, M., **Nedyalkova, D.**, **Münnich, R.**, **Seger, J.** and **Zins, S.** (**2011**): *Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion*. Research Project Report WP2 – D2.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Holzer, J. (**2009**): Robust methods for the estimation of selected Laeken indicators. Diploma thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

Hulliger, B., **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**a): *Report on the simulation results: Appendix*. Research Project Report WP7 – D7.1 - Appendix, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Hulliger, B., **Alfons, A.**, **Filzmoser, P.**, **Meraner, A.**, **Schoch, T.** and **Templ, M.** (**2011**b): *Robust Methodology for Laeken Indicators*. Research Project Report WP4 – D4.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Little, R. J. A. and **Smith, P. J.** (**1987**): *Editing and Imputation for Quantitative Survey Data*. Journal of the American Statistical Association, 82 (397), pp. 58–68.

Maronna, R. and **Zamar, R.** (**2002**): *Robust estimates of location and dispersion for high-dimensional datasets*. Technometrics, 44, pp. 307–317.