EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

# Deliverables 10.1 and 10.2

# Policy Recommendations and Methodological Report

Version: 2011

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

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 deliverables 10.1 and 10.2

**Chapter 1:** Beat Hulliger, University of Applied Sciences Northwestern Switzerland;
Jan-Philipp Kolb and Ralf Münnich, University of Trier.

**Chapter 2:** Beat Hulliger, University of Applied Sciences Northwestern Switzerland;
Jan-Philipp Kolb and Ralf Münnich, University of Trier.

**Chapter 3:** Monique Graf, Swiss Ferderal Statistical Office;
Beat Hulliger, University of Applied Sciences Northwestern Switzerland;
Risto Lehtonen, University of Helsinki;
Jan-Philipp Kolb and Ralf Münnich, and Stefan Zins, University of Trier;
Matthias Templ, Vienna University of Technology.

# Main responsibility

The AMELI Team

# Evaluators

**Internal expert:** General Assembly

# Aim and objectives of deliverable 10.1 and 10.2

The main focus of this report is to summarize the results of the AMELI project. Further, it is the target to formulate the results of all work packages in a way which is understandable for a general audience. The deliverables 10.1 and 10.2 have been combined in a single deliverable to ensure a maximum of clearness and functionality.

This final report consists of two parts. The first part draws together policy-relevant results from all nine research WPs. The body for this part is the analysis of the results of WP 7 and their political impacts. Therein recommendations for future policy use are made.

The second part focuses on technical and methodological issues. Detailed overviews of the project methodology and significant results from the project are provided in this part. Deliverable Templ et al. (2011a) will comprise the computer programs and selected source codes in R language (R Development Core Team, 2011). The codes are split between several R-packages and specific code published in the annex of deliverable Hulliger et al. (2011a).

The work package at hand was only possible due to an intensive collaboration with all members of the AMELI project.

# Contents

# Chapter 1

# Introduction

A full benchmarking system is important to monitor policy performance and the impact of European strategies on progress. This holds for the Lisbon strategy as well as for the Europe 2020 strategy. For this reason, the European Commission has engaged in selecting, collecting and analysing a set of indicators that are published each year. The Stockholm European Council has further emphasised the need for effective, timely and reliable statistics and indicators. A main challenge is to develop indicators for important characteristics and key drivers. An utmost important and challenging area to be measured is social cohesion. Based on a clear definition of social cohesion, a universally-accepted high-quality and robust statistic to adequately measure social cohesion is required. Further, tools for measuring temporal developments and regional breakdowns to sub-populations of relevance are of great importance. In order to measure social cohesion with Laeken indicators adequately while regarding national characteristics and practical peculiarities from the newly created EU-SILC, an improved methodology was elaborated within AMELI. This will ensure that future political decision in the area of quality of life can be based on more adequate and high-quality data and a proper understanding of the Laeken indicators by the users.

A large simulation study based on EU-SILC data was necessary to allow for the simultaneous elaboration of the methodology focusing on practical issues aiming at support for policy.

Chapter 2 gives clear policy recommendations. In the subsequent chapter the methodology used reach in these recommendations is provided. To approach the investigation of the indicators for social exclusion and poverty it is necessary to develop a clear definition of the indicators to be evaluated. This was done in work package 1. Thus, the current status of research is documented in one document, whereas specific issues are taken up again in the several work packages. Chapter 4 is an overview of all R packages developt during the project, most which are available under public domain http://cran.r-project.org/. There are also references to additional code which has not yet been integrated into a R package.

The deliverable at hand summarizes the outcome of the project. Policy recommendations are made, based on the insights gained from the research on the highly sophisticated methodology which is underlying the indicators. The insights result from the various simulations carried out within the AMELI project.

# Chapter 2

# Policy recommendations

The purpose of social indicators is to measure progress of the society and thus to give reliable information about the impact of policy and therefore increase the steering-capacity of governments and politicians. Social indicators have reached a degree of maturity which allows them to give the necessary support for decisions in policy and for organisations and enterprises which are stakeholders in our modern society. Poverty and social exclusion are highly multidimensional and complex phenomena. Various different means can be applied to measure these phenomena. It has to be clarified whether to measure income poverty or social deprivation, whether to measure in absolute or relative terms and in which depth information is necessary for realizing this. The basis of social indicators are surveys of high quality. One quality aspect is that these surveys use the same concepts and operate the same definitions across countries. Harmonization is of great importance for the applied concepts as well as for the underlying data bases. The SILC surveys in Europe are an example of a coordinated high quality information gathering instrument. Thus progress was made on this field but one should always keep in mind that still different exceptional events can occur for some countries or regions at one point in time. Therefore it was important to evaluate these exceptional events within a simulation study.

The indicators based on EU-SILC which are computed to evaluate poverty and social exclusion differ in nature. More subjective information like the health perception is incorporated as well as more objective attributes, like personal income. Often it is a matter of definition how to assess the soft information. But even for the hard facts measurement errors and a large variability can occur. Furthermore, the variety within Europe should be taken into account, when establishing social indicators. And part of this establishment of social indicators is a concept and concrete measures to judge their quality.

In this chapter the impact of various stages of the establishment on the social indicators and the implication for policy use of these indicators is described.

In the following, two areas of policy will be covered:

- Policy of users of social indicators
- Policy of producers of social indicators.

This was necessary to distinguish between the inferences of the AMELI project and their different impact in the two areas.

## 2.1 Recommendations for Using Social Indicators

Social indicators are highly complex measures which try to convey in a very concentrated form important messages about the state and development of the society across Europe. Stating with this ambitious goal it is clear that any social indicator should be, in principle, accompanied by a long list of caveats. Here are the recommendations that the AMELI project came up with:

> **Recommendation 1:** The variability of the social indicators should always be kept in mind and taken into account when using them for policy making.

It is difficult to come up with good standard errors for social indicators but the research of AMELI showed that useful approximations are feasible. Even a rough indication of a standard error is better than forgetting that a social indicator inherently has a variability built in. Much work on the communication of the accuracy and how to take it into account is necessary. In any case, the users of social indicators should make more use of provided accuracy indications and should require more indications on accuracy.

> **Recommendation 2:** The impact of data collection and data preparation on social indicators is heavy. Therefore it is difficult to separate the analysis of the indicators from the data preparation. When analysing social indicators, the data collection and data preparation must be taken into account.

This means that the analysis of indicators must necessarily be based on the knowledge of data collection and data preparation steps. A more seamless collaboration is necessary between data producers and data users to ensure that all necessary knowledge is used at the analysis stage.

> **Recommendation 3:** Social indicators on the one hand must track the main trends of the society. These main track indicators should be robust against large deviations. On the other hand the degree of variability and extremeness in a society should also be monitored.

Robust methods ensure that the main trends can be seen without too much disturbance from large deviations and at the same time they serve as a baseline to see these large deviations. The AMELI project showed that more robust procedures can be used for social indicators.

> **Recommendation 4:** Social indicators on a global level must be complemented by social indicators on regional levels and for important subgroups of the society. The methods developed under the common denomination of small area estimation are now sufficiently mature to be applied.

The AMELI project showed that small area estimation for social indicators is possible and useful. They need expertise for their application and communication about the results gained from small area estimation is complex. However, the rich picture provided by them makes it worthwhile to include results from small area estimation into regular reporting on the social reality alongside the global indicators.

> **Recommendation 5:** The definition, establishment, validation and further development of social indicators is transdisciplinary effort. More collaboration and integration of the disciplines of sociology, economics and statistics is necessary to ensure progress.

The complex collection, preparation and analysis process for social indicators is not yet fully integrated and a close collaboration and major effort of subject matter scientists, statisticians and users is necessary to further develop the social indicators of the European Union.

## 2.2 Recommendations for Producing Social Indicators

The following recommendations are formulated in WP7 of the AMELI project and thus are based on the extensive simulations documented in deliverable D7.1 of the AMELI project.

### 2.2.1 Parametric Estimation of Income Distributions and Derived Indicators

Two parametric models have been used for parametric estimation, the generalized beta distribution of the second kind (GB2) and the two component Dagum distribution (TCD).

#### Parametric Estimation Using the GB2 Distribution

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 linearisation: The sandwich variance estimator using the linearisation 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. (2011b, 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 ($p_3 = 0.2$).

#### Parametric Estimation Using the TCD distribution

All in all, the indirect estimation with a two component Dagum distribution (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 linearisation 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 (at risk of poverty rate), while the direct way of estimation seems to be beneficial for the QSR (quintile share ratio). The results for the Gini coefficient are rather ambiguous, but tend to favour 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.

### 2.2.2 Small Area Estimation

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.

### 2.2.3 Variance estimation

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 linearisation and re-sampling methods have been studied.

In general we endorse the usage of linearisation over that of re-sampling, because linearization is less computationally intensive and usually more efficient. Further, the linearisation 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 re-sampling methods that are often not applicable to complex sampling designs.

However, because the results of the simulation study on the performance of the linearisation methods are mixed, we can not recommend linearisation without reservation. The simulation has shown that variance estimates based on the linearisation 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 re sampling 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 re sampling methods we refer also to Table 3.1 in Bruch et al. (2011), which contains an overview of the different re sampling methods with regard to there applicability to some characteristics of (complex) sample surveys.

### 2.2.4 Robustness

Estimation of 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 constant 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.

Estimation of 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 is the best version out of three alternatives that have been investigated. 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 outlier detection and imputation: 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.

Variance estimation for data which has undergone multivariate outlier detection and imputation as well as subsequent re-aggregation into disposable income followed by classical estimators of Laeken indicators might be possible with re-sampling 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.

# Chapter 3

# Methodological report

In the following the methodological work processed within the AMELI project will be summarized by approximately one page per work package. It is the target to give an overview of the methodological work processed. Further details can be looked up in the particular deliverables.

## 3.1 Laeken indicators (WP1)

Deliverable 1.1 (Graf et al., 2011a) provides an overview of different methodological issues concerning the Laeken indicators, a set of indicators monitoring the multidimensional phenomena of poverty and social exclusion. The topics we treat include the definition of social cohesion, the rationale for the Laeken indicators and the need for a harmonized definition of income.

The elaboration of a clear definition of social cohesion has been an important basis for the AMELI research: the study includes the relation of Laeken indicators to social cohesion and gives a large set of bibliographical references on the subject. Follows a description of the context in which the European Union Statistics on Income and Living Conditions (EU-SILC) database has been constituted, this database being the main source for the estimation of the Laeken indicators. We describe the former international databases that are related to EU-SILC, current results on Laeken indicators and provide a review of European research projects having a potential impact on AMELI. The results from the studies already done for the Luxembourg Income survey, the ECHP, and EU-SILC have been a starting point for the work in this work package. The JRC-OECD studies on composite indicators regarding the Laeken indicators were also reviewed.

References on income distributions have been scrutinized from the point of view of statistical estimation, specifically for the estimation of poverty and inequality indices. Another step was the review of the methods for precision assessment compiled within FP5 projects with special focus on Laeken-Indicators.

Finally five thematic summaries sketch the methodology that will be developed in the AMELI project for the context of Laeken indicators: state-of-the-art in small area estimation, in parametric income distributions, on variance estimation, on robustness methods, and on visualization.

## 3.2 Estimation (WP2)

### 3.2.1 Aim and Objectives

The objective of work package 2 is to investigate different aspects of the estimation methodology for indicators. The emphasis is on indicators on poverty and social exclusion (monetary Laeken indicators), because they are among the most difficult to estimate. Nevertheless the findings could be applied to other areas of indicator estimation as well. The results are divided into two parts: in deliverable 2.1 (Graf et al., 2011b), we conduct investigations on the use of parametric income distributions for the indirect estimation of the poverty and inequality indicators; in deliverable 2.2 on the other hand (Lehtonen et al., 2011), different estimation methods of these indicators for population subgroups or domains and small areas are compared.

### 3.2.2 Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion

Parametric income distributions have long been used for modelling income. Both modelling of the whole income range or of the tails of the distribution have been investigated in the literature. Here we concentrate on the modelling of the entire distribution. The advantage is that there exist explicit formulas for the poverty and inequality measures as functions of the parameters of the theoretical income distribution. From the literature review (Graf et al., 2011b, Chapter 1), it is found that a certain four-parameter distribution, the generalized beta distribution of the second kind (GB2) outperforms other distributions for income modelling. Moreover, generalizations to five-parameter distributions seem not to improve the fit in general. This is why the main emphasis in this deliverable is on the use of the GB2.

#### GB2 distribution for the estimation of monetary indicators

The expressions of the indicators under the GB2 assumption can be found in Graf et al. (2011b, Chapter 2). We investigated different estimation procedures of the GB2 parameters and distribution (Graf et al., 2011b, Chapter 3) and their variance under a complex sampling design.

First we consider the context (met with the SILC data) where we have access to the micro-data, that is to the individual incomes. We use the pseudo log-likelihood: the population log-likelihood is approximated by the extrapolated sum of the scores, i.e. by its Horvitz-Thompson estimate.

Once the parameter estimates have been obtained, the indicators estimates are derived by plugging the parameter estimates into the functional expression for the indicators. We provide the design-based variance and covariance matrix of the GB2 parameters and of the derived indicators by linearisation.

Suppose we do not have the income micro data at disposal, but that the indicators, fitted on empirical data, are publicly available. This is the situation of an external user of the Eurostat website. The indicators have been produced without any reference to a theoretical income distribution. It is then possible to go the other way round, that is to reconstruct the whole income distribution, knowing only the values of the empirical indicators and assuming that the theoretical GB2 distribution models the empirical distribution to an acceptable precision. This approach has been applied to EU-SILC data with success. This means that the set of indicators contains enough information to permit the reconstruction of the empirical distribution generally to an acceptable precision.

#### GB2 as a compound distribution

A theoretical distribution is a compound distribution if its density can be expressed as an integral over the distribution of a random parameter. The GB2 possesses this property, the random parameter being the scale parameter. More precisely the GB2 density is the expectation of gamma densities with random scales. The expectation is taken with respect to the distribution of the scale. An interpretation is that the income distribution is the result of a mixture of inhomogeneous groups. We show in Graf et al. (2011b, Chapter 5) that there are two ways of expressing the GB2 as a compound distribution that put the emphasis respectively on the right tail or on the left tail, the latter being better when the interest is in poverty. The scale range can be partitioned into a number of intervals and by integrating over these intervals, we obtain the GB2 density in the form of a mixture of component densities. Keeping the components fixed and leaving their weights in the mixture vary, we can optimize the fit for subgroups.

#### Dagum mixtures

The estimators have been compared under the income distributions can suffer instability in the estimated parameters. Goodness-of-fit tests showed the characterization properties and served as study of the sensitivity of Laeken indicators to parametric assumptions. Mixture distributions have been used to account for population subgroups.

### 3.2.3 Small Area Estimation of Indicators on Poverty and Social Exclusion

Indicators on poverty and social exclusion selected for consideration are at-risk-of poverty rate, quintile share ratio (S20/S80 ratio), relative median at-risk-of poverty gap and the Gini coefficient. The indicators are typically non-linear and are based on non-smooth functions such as medians and quintiles, which makes the estimation a non-trivial task. This holds especially for the estimation of the indicators for population subgroups or domains and small areas. Therefore, extended methodology for small area statistics was considered necessary for the selected Laeken indicators.

The aim was to assess the relative merits and practical applicability of various design-based and model-based estimation procedures developed for the selected Laeken indicators. Statistical properties (design bias and accuracy) were investigated by extensive simulation experiments using real register and survey data. To cover a broad variety of typical practical situations, methods were investigated under equal and unequal probability sampling as well as under various outlier contamination schemes. Thus, we were able to evaluate the sensitivity of methods to varying design weights and robustness of methods to outliers. Access to unit-level population data was assumed in many cases being increasingly realistic assumption in statistical infrastructures of the EU countries. We also developed frequency-calibrated prediction estimators when aggregate population data only were assumed available. Technical description of methods and numerical results are in Lehtonen et al. (2011).

### Estimation of Poverty Rate

In poverty rate estimation with design-based methods we used generalized regression (GREG) and model calibration type estimators. In model-based estimation we used synthetic and empirical best prediction (EBP) type estimators. Logistic fixed-effects and mixed models were used because the underlying study variable was binary. The model specification yielded in indirect type design-based and model-based estimators. Methodological background can be found in Lehtonen et al. (2003, 2005) and Lehtonen and Veijanen (2009). In general, domain sample size is the most important factor affecting accuracy of estimation for a domain. Typically, accuracy improved when domain sample size increased. This holds for both estimation approaches. In model-based estimators, domain sample size had less effect on bias. Incorporation of powerful explanatory variables into model tended to decrease bias of model-based estimators and improve accuracy. This was more apparent for model-based estimators than model-assisted estimators.

Design-based estimators including direct (default) estimator and model-assisted logistic generalized regression estimators and model-calibration estimators appeared nearly design unbiased as expected, independent of the model choice. Model-based synthetic and empirical best prediction estimators were design biased especially in domains with small sample size. Bias in model-based estimators was connected to the model choice: bias tended to decline when improving the capacity of the model to capture the possible heterogeneity of the data. With respect to accuracy, design-based model-assisted estimators and model-based estimators outperformed the direct estimators. In model-assisted estimators, there were no clear differences between generalized regression estimators based either on logistic model with domain-specific fixed intercepts or model with random intercepts. A logistic mixed model however might be favoured because of parsimony; the drawback is that estimation of logistic mixed model is more challenging than that of logistic fixed-effects model. Domain size correction did not substantially improve the accuracy of generalized regression estimators. Generalized regression estimators yielded the best results, although differences with respect to model calibration estimators were small. Model calibration however offers more flexibility in defining the levels of calibration, which property might be useful in practice.

Model-based estimators indicated better accuracy than model-assisted estimators, when the same underlying model was specified for both estimator types. This holds for domains with small sample size in particular. Differences in accuracy declined with increasing domain sample size. In design-based estimation, information on sampling design is routinely incorporated into estimation procedure. Accounting for sampling design complexities also appeared important for model-based estimation under unequal probability sampling. Bias tended to decline and accuracy improved when incorporating sampling design information into the estimation procedure. This was achieved either by including design variables into the mixed model or by incorporating weight variable into the estimation of the model. However, sampling design did not seem to affect model-based estimators as much as we expected.

In general, results were not necessarily improved when adding domain-specific terms to the model. We obtained better estimates by including terms of higher aggregation level, such as random intercepts associated with NUTS3 levels when domains were defined at NUTS4, for example. Poverty rate estimators were quite immune to outlier contamination.

In poverty rate estimation, model-based empirical best prediction type estimators might offer the best option because of good accuracy, unless it is important to avoid design bias. This is an indication of trade-off between bias and accuracy. If bias should be avoided then logistic generalized regression estimators or model-calibration estimators offer realistic alternatives. Logistic mixed models are at least theoretically preferable to fixed effects models as they describe differences between domains parsimoniously.

### Estimation of Poverty Gap, Gini Coefficient and Quintile Share Ratio

Estimation for poverty gap, Gini coefficient and quintile share is more demanding than for poverty rate, because these indicators cannot be expressed as functions of indicator variables. We thus modelled the (log-transformed) response variable (equivalized household disposable income) by linear mixed models with domain-specific (or higher-level) random intercepts. Ordinary predictors involved predictions plugged into the default formula in place of original observations. These predictors were substantially biased: in our simulations poverty gap estimates and Gini coefficient estimates were too small and quintile share estimates were too large. Due to the large bias, accuracy of ordinary predictors was often worse than that of the corresponding direct design-based estimator.

Because a synthetic estimator based on ordinary predictions did not perform well, we developed expanded prediction type estimators based on transformed predictions. The expanded predictors benefit greatly from the transformation of predictions bringing the distribution of predictions closer to the distribution of observations. Bias decreased substantially and accuracy improved due to the transformation, as compared with the ordinary predictors. Inclusion of design weights in the technique probably reduced design bias in experiments with unequal probability sampling.

The expanded quintile share and Gini predictors were more robust against outlier contamination than the direct design based estimator or the ordinary predictor method. As the expansion incorporates percentiles of observations up to 99th percentile, rare outliers occurring with frequency of 1 percent do not affect the expanded predictor too much. When the proportion of outliers was very large (e.g. 15 percent), the expanded predictor failed but not as badly as the other estimators. The breakdown point of the estimator can probably be adjusted by changing the range of percentage points. In small domains, the expanded predictor usually had better accuracy (although larger bias) than the default estimator. In largest domains, the default estimator may be preferred to the expanded predictor if there are no outliers, but in contaminated data expanded predictors appear to be better than the direct design-based estimator, although poverty gap is an exception.

In poverty gap estimation, only the left tail of the distribution of predictions contributes to estimates. The expansion method does not seem to work as well as in quintile share and Gini coefficient, where most of the predictions are included in the estimators.

The frequency-calibrated estimator was developed for situations where only aggregate population data are available. The method was not usually as accurate as the expanded predictor with unit-level data and same auxiliary variables. This was expected, as the frequency-calibrated predictor has access only to the domain frequencies of classes of auxiliary variables in the population, not to unit-level information. The estimator appears to have similar robustness properties as the expanded predictor. However, for poverty gap the frequency-calibrated method may perform poorly.

A composite estimator was constructed as a linear combination of a direct design-based estimator and the corresponding expanded predictor type estimator. In our experiments, the weight coefficient was estimated separately for minor, medium-sized and major domains. Under no outlier contamination, composite estimators had smaller bias than the expanded predictors, but accuracy was usually not as good as for expanded predictors. In small domains, composite estimators were more accurate than the direct design-based estimators. If contamination yielded bias in the direct estimator, composite estimators consequently suffered from bias. Composite estimators of quintile share or Gini coefficient may not be a good choice if some contamination is suspected. However, we might prefer composite poverty gap estimators over predictors.

We also studied a model-based method based on random imputation. The method resulted in fairly good poverty gap estimates, although there was systematic bias: estimates were too large in small domains and too small in large domains. The assumption of log-normal income distribution was probably not realistic enough in this case, as some of the imputed incomes were unrealistically large, and the calculated Gini coefficients were too large. Better results might be obtained with a more realistic income distribution.

Variance and mean squared error (MSE) estimation has been considered in selected cases only. Pseudo-replication methods such as bootstrap and jackknife provide applicable options for variance and MSE estimation of the alternative estimators of the poverty indicators discussed here.

## 3.3 Variance Estimation (WP3)

### 3.3.1 Aim and Objectives

The work on variance estimation methodology provided, that at first the state-of-the-art in variance estimation methodology with latest developments was to be investigated and adequately summarized. The focus was to be laid on methods which occur to be important for indicator application. In addition was the problem non-response or missing values and there treatment on indicator outcomes as well as evaluation for small groups such as for small area estimation of indicators, to be considered. Thereafter, based on the first results of the simulation study, the state-of-the-art methodology will be examined to enable improvements of the methodology with respect to the applicability of the methods and the efficiency of the output.

### 3.3.2 Implementation of Work Assignment and Findings

These assignments where addressed in Bruch et al. (2011) and Münnich and Zins (2011). The state-of-the-art in variance estimation methodology was mainly captured in Bruch et al. (2011). Bruch et al. (2011) deals in general with variance estimation for complex surveys. The term complex surveys encompass sampling designs including multi-stage, stratified or unequal probability sampling, because most countries involved in EU-SILC employ such sampling designs to obtain the date on which bases indicators are estimated. The work of Bruch et al. (2011) has two main themes, first direct variance estimation is addressed and second re-sampling methods for variance estimation. The methodology on direct variance estimation is mainly concerned with finding simple but still reliable variances estimator in the presence of unequal probability sampling. Especially approximation to the true design variance are considered, that allow for easy to compute variance estimates. The re-sampling methods are treated more generally under the aspect of there suitability in the presence of complex sampling designs. Difference re-sampling techniques (and developments) are presented and their advantages and disadvantages are pointed out. The final chapter of Bruch et al. (2011) describes a Monte Carlo simulation study that evaluates the performance of different variance estimators in the presence of a two-stage sampling design. The aim of the study was to investigate the dependency of the different variance estimation methods on parameters like the size of the primary sampling units (PSUs) or the homogeneity of the elements between and within the PSUs was investigated. The study shows that these parameters have a strong influence on the accuracy of the variance estimation methods. Dependant on the peculiarity of the parameter the variance estimation only applied at the first stage can be sufficient or necessary to consider all stages. Münnich and Zins (2011) completes the methodological review on variance estimation by introducing linearisation methods, which enable estimating the variance of these statistics using standard variance estimation software developed for linear estimators. Thus, chapters 1 and 2 dealing with direct variance estimation (for linear estimators) in Bruch et al. (2011) are complements of each other, together the comprehend the tools of estimating the variance for both complex estimates and sampling designs.

The results of simulation study to examination the state-of-the-art methodology and there discussion can be found in chapter 9 of Hulliger et al. (2011a). Although re-sampling methods have also been evaluated the study focuses more on the usage of linearisation, because for all indicators of poverty and income inequality used in the context of EU-SILC, there linearisation is already available, which allows for the direct usage of standard variance estimation software. Thus there is a greater flexibility with regard to the actual sampling design. However, one of the major findings of the simulation was that when using linearisation techniques in practise the user has still to into account peculiarities of the sample data, that can stem either form the distribution of the reference population, or are introduced by the sampling design. So did linearisation perform very well for the one-stage design but are negatively biased for the two-stage designs. This presumed to caused by the highly skewed distribution of sampling elements in case of the two-stage samples.

In addition to the main task in this work package, Münnich and Zins (2011, chapter 3) is dedicated to variance estimation of a change in indicator values, which is required to test whether a measured change of an indicator was statistical significant or not. The work incorporated the methods employed earlier in the field of variance estimation for non-linear statistics, in particular those of linearisation. This work was motivated by the fact that rather the evolution of indicator values than there value in a certain period is of particular interested for policy making.

The influence of missing values on point and variance estimation are also studied in Alfons et al. (2009).

## 3.4 Robustness (WP4)

### 3.4.1 Objectives

The work package on Robustness (WP4) should study the behaviour of classical non-robust estimators when outliers occur and develop new robust methods for the social inclusion and poverty indicators (Laeken indicators). The effect of outliers on bias and variance of estimators of the indicators should be investigated. Multivariate outlier detection and imputation should be developed and their effect on the indicators should be studied. Research on robustness of small area estimation procedures and on robust small area estimation was targeted, too.

All these methods should be implemented in R and evaluated in the main simulation study of the AMELI project.

### 3.4.2 Implementation of work

The work package summarized the state-of-the-art on robust estimation, outlier and influential value detection and robust imputation for the social inclusion and poverty indicators. The most vulnerable indicator is the Quintile Share Ratio. This is inherently due to its purpose, to measure inequality. Since inequality depends on extreme values of income the Quintile Share Ratio must be sensitive to extreme values. However, this sensitivity introduces rapidly large biases, and worse, a large variability. This large variability is important even for the large samples of the SILC surveys. Any inequality measure must necessarily suffer from this non-robustness, e.g. also the Gini-indicator. However, also during the project, it was recognised that the Quintile Share Ratio has a very valuable advantage, namely its simplicity.

Two lines of research for the robustification of the Quintile Share Ratio were followed, a non-parametric approach by FHNW and a semi-parametric approach by TUW. Both approaches lead to estimators which are robust, unlike the classical Quintile Share Ratio estimator. However, the estimators depend on assumptions and on so-called tuning constants which makes their handling more complex than the classical ones. Nevertheless it is recommended to always use very mildly robustified estimators at least as a comparison measure if not as a full replacement of the classical non-robust estimators.

The development of efficient code in R and the study of the effect of various versions of these estimators, as well as the search for good tuning of them, needed careful and extensive tests. Also variance estimators proved to be challenging due to the highly non-linear form of the Quintile Share Ratio and due to the additional complexity of the robustification. Approximate closed-form variance estimators have nevertheless been developed. This research is documented in AMELI Deliverable D4.2.

The simulation study profited from the set-up developed in the EUREDIT project. However, only in the AMELI project the full complexity of the interaction between missingness and outlyingness could be developed and implemented in the simulations. The concepts, notations and implementation are mostly described in AMELI Deliverable D7.1. This contamination set-up was also used in the other work packages to investigate the effect of contamination, i.e. outliers, for example on small area estimators. Contamination models have also been practically tested and compared in case of synthetic AAT-SILC data in Alfons et al. (2010b).

Together with the contamination set-up also criteria were developed. However, the main criteria are the same as for all methods: bias and variance of the indicator estimators. In addition, the false positive and false negative detection rates were used. The AMELI criteria are described in AMELI Deliverable D7.1.

The work on robust small area estimation proved to be very challenging. In particular, the numerical aspects of the estimation procedures are difficult. Only towards the end of the AMELI project first concepts and estimators could be formulated. However, it was not possible to include these estimators in the main simulation study.

The main income variable of the SILC surveys, equivalized disposable income, depends on many income components on the level of persons and households. Instead of robust estimation starting from the equvalized disposable income, multivariate methods would study the income components together. Multivariate outliers detection in incomplete survey data has been studied in the EUREDIT project. However the application to the SILC data proved to be difficult because of two common phenomena: The distribution is not only skew but semi-continuous at zero because many components are actually zero for a particular person or household. Since most multivariate methods rely on elliptically contoured data new methods had to be developed to cope with this situation. Finally two methods seemed to stand up practical problems: The BACON-EEM algorithm for multivariate outlier detection (Béguin and Hulliger, 2008) combined with a robustified version of multiple regression imputation and the Epidemic Algorithm (Béguin and Hulliger, 2004). Again the tuning of these algorithms is complex, in particular for the Epidemic Algorithm. Nevertheless they permit much more control over the outlier treatment than the manual procedures which are customary at the moment. In any case, the pre- and post-treatment of the data, i.e. the income components, proved to be of crucial importance.

A comparison of robust tail modelling methods has been carried out in the simulation study in Deliverable 6.1 (Alfons et al., 2011a) but also in Alfons et al. (2010d). The full AMELI simulation with AMELIA and AAT-SILC data with non-parametric and semi-parametric robust methods and with multivariate robust methods is documented in AMELI Deliverable D7.1.

The code of all methods and algorithms was developed in R. Some of the code has been released as packages of R on CRAN and more code is documented either in private packages or as functions in D4.2, D6.1, D6.2, D7.1 and D7.1-Appendix.

## 3.5 Data quality (WP5)

There is a large amount of documentation on the EU-SILC Database and one difficulty for the user is to find the right reference for the question at hand. The first goal addressed by Graf et al. (2011c) is thus to help the user to find the necessary information easily. We have tried to give the Internet links for the references we cite, and update them during the course of the project. The documents describing the database are of several kinds: description of the database each year, quality reports and guidelines.

We continue with an analysis of the differences in definition between countries of the main variable of interest in the project: the equivalized disposable income.

Finally we provide recommendations for the simulation setup in Work package 6 on the choice of sampling designs for use in the simulations.

## 3.6 Simulation (WP6)

Monte-Carlo simulation studies have been planned for testing the methods from WPs 2-4 in a close to reality environment based on realistic data sets from selected European countries under different conditions. Robustness of model assisted and model based methods as well as parametric estimation methods in respect to important practical needs and needed relevant assumptions have been topics of these simulation studies.

It is always the optimum to test these methods on real data. However, the lack of real population data and the complexity of the real processes involved in contamination and non-response make completely realistic scenarios virtually impossible. Thus, it was the target to clarify the data needs within the simulation study and to develop a synthetic data set to fulfil these needs. Therefore the work load of work package 6 was divided into two major tasks. In a first step an outline of the simulation process and the application of common simulation guidelines is provided. And in a second step suitable mechanism for generating population data as basis for the simulations are developed (see Alfons et al. 2011b for further details).

Concerning the simulation design, special emphasis was placed on typical data problems such as outliers and non-response. A general question was whether these should be included on the population level or in the samples. While the first approach may be more realistic, it results in an unpredictable number of outliers or missing values in the samples and may thus be suitable if these issues are not the main interest of the simulations.

Nevertheless, to evaluate procedures such as outlier identification methods or imputation methods, maximum control over the amount of outliers or missing values is required. In these situations, the proceeding to include them in the samples was identified as more practical approach.

It is an usual proceeding to test individual methods in smaller preliminary simulation studies. These studies showed that several challenges have to be covered in field of measuring social exclusion and poverty and that the requirements and starting points can be different. Survey data like the SILC data are often subdivided into several strata corresponding to certain geographical units. Different initial situations do exist for the EU-countries. Methods for handling specific problems like the treatment of outliers and non-response are usually applied on smaller data sets whereas the Small Area Estimation is applied on bigger data sets. Therefore it was decided to split simulations into different tasks and to combine them later on within work package 7. Another topic within the deliverable 6.1 are evaluation criteria for different types of simulations.

One problem with generating population data from real life samples is that the structure of the original data needs to be represented in the synthetic population. This must happen under the restriction that statistical non-disclosure must be compromised by all means.

The first step of data generation is thus to set up the household structure using only regional information, the number of persons in the households as well as their gender and age class. This avoids disclosure problems since no other auxiliary information of the individuals in the real life sample is used and all categories after cross tabulation of these variables are far away to be unique in the sample. Further variables of the synthetic population can then be generated with statistical models. It has been demonstrated that this approach succeeds in reflecting the dependencies among variables and the heterogeneities between subgroups. Therefore, the populations created in this manner are perfectly suitable for close to reality simulation. Moreover, within the project it has been proofed that the data fulfils the requirements of confidentiality (see TEMPL and ALFONS, 2010).

The relative weight reflecting these geographical units have been used as inclusion probabilities for the data generating process. Further relevant grouping structure is determined by variables like the number of households in each stratum, gender, and age. All these factors and their relative occurrences have to be determined first from the observed data and from results of WP 5.

The generation of a data set was a necessary basis for the simulations within the AMELI project. For the generation of a synthetic data set different problems had to be addressed. A fine administrative structuring was necessary to realize the complex sampling designs. Furthermore, many income components and their correlation structure had to be reproduced for the synthetic data set.

It is important to keep in mind that many methods are conceivable to produce a synthetic data set and the best method depends always on the starting point and the requirements coming from the simulations. The description of work scheduled a lot of simulations and different data sets have been available. Due to these reasons it was decided to generate two different synthetic data sets, which are the AAT-SILC data set and the AMELIA data set. The methods which is behind the generation of the AAT-SILC data set is published in a R-package (see ALFONS and KRAFT, 2010; ALFONS et al., 2011d).

One has to bear in mind that only those effects can be measured in microsimulation approaches which have been included into the data set. There is always the risk to get in a garbage in garbage out process. Estimation techniques can work perfectly well in a microsimulation model which do not work in reality. The risk of doing so can never be completely excluded. To catch some of these uncertainties a scenario analysis was applied.

The results of these simulation studies are displayed in work package 7 (Hulliger et al., 2011a). Which provides a thorough analysis of the impact of different situations on the indicators on social exclusion and poverty. In work package 10 this output is used to support policy making with best practice recommendations.

## 3.7 Analysis (WP7)

### 3.7.1 Objectives

The objective of Work package 7 Analysis was to collect the simulation results, synthesise them and to derive recommendations focussed on social inclusion and poverty indicators and, if possible, for other similar indicators. In addition the impact on policy measures should be discussed. The work package should draw on and compare with the studies in the work packages on estimation, robustness and variance estimation. Obviously also the simulation work package should input into WP 7.

### 3.7.2 Implementation and main results

The original idea of establishing a common database for all simulation results was abandoned after noting that the form of the results for the different methods was so divers that no common format could cover all needs. The solution chosen was then to have a distributed data base with a common metadata interface which allows a quick overview over the simulation runs. The flow of information on the simulations, including the metadata, was defined and implemented.

A main task of WP7 was also to fix the multitude of simulation set-ups and their parametrisation. The sample designs were fixed based on the results of WP5 and a careful evaluation of the practical needs. The missing value mechanisms for multivariate missingness were investigated on the basis of the original SILC data. The contamination mechanisms had to implement the new concepts of outlying and contaminated completely at random and at random. Also domains were suggested for the analysis of indicators, though they were not used in the simulations. An effort was also made to harmonise notation and formulae for evaluation criteria. The description of the simulation setup can be found in AMELI Deliverable D7.1, Part I. Code snippets for the simulation bed are in the main Deliverable D7.1 and in D7.1-Appendix. It is to be noted that the creation of the universes underlying the simulations was done under WP6 Simulation and is described in AMELI Deliverable D6.1.

Overall 55 simulation runs have been documented according to the metadata-specifications. They cover 154’000 replicates of SILC samples and usually each of the replicates is used to evaluate a whole set of methods with a set of tuning parameters. The simulations results are collected in a large document, AMELI Deliverable D7.1-Appendix. Some of them are based on a template to simplify the orientation but others are custom-made to suit the needs for the analysis.

The analysis of these results was carried out by the partners involved in the simulations and collected into reports. They usually did not comment all results but concentrated on the main issues to draw conclusions of practical relevance. All reports have a recommendation section at there end. The reports are collected in AMELI Deliverable D7.1, Part II.

The recommendations and derived policy issues are collected in the present final report of AMELI.

## 3.8 Visualization (WP8)

The work done in WP 8 was to research and implement visualization tools for visualizing indicators with and without mapping for policy support and to support end-users of the indicator values, and to implement visualization tools for microdata (visualization of missing values and imputed values). In addition, functionality to automatically display simulation results with suitable plots were developed.

For visualization task for microdata, the open-source R-package VIM (Templ et al., 2011d) - available on CRAN - have been developed. It allows to explore the structure of missing values in microdata using newly developed univariate, bivariate, multiple and multivariate plots. The software can be used via a graphical user interface as well. The exploration of missing values gives the data analyst an overview of the structure of the missing values and the data in short time.

New methods to present indicators are developed, such as the checkerplot. Checkerplots allows visualizing indicators for each region in a grid with equal space for each region by not losing the geographical relationships of the regions. Other methods to visualize indicators were modified, like the evaluation plots and the funnel plot. Other developments include mapping features, such as displaying pre-whitening cross correlations between countries in maps.

The automatic plot features for simulation results have been fully integrated in the simulation environment simFrame. Depending on the structure of the simulation result, suitable plot methods are chosen automatically. This feature has impact to the analysis package were simulation results can be easily produced by this feature.

Methods for displaying indicators considered within AMELI have been implemented in R as easy to grasp graphics. Based on the statistical software R, additional statistical summaries can be visualised. Exploratory tools and various visualisation methods have been adapted and implemented in R. The goal of these tools was the detection of the missing data mechanisms, and the identification of outliers and values that are influential to statistical analyses. Possibilities to visualize indicators which are available on geographical units in maps are proposed. The corresponding maps from several countries have been stored as objects in R.

Evaluation plots are described in detail in Hulliger and Lussmann (2008) and (Hulliger et al., 2008) (the references are listed in Deliverable 8.1). The aim was to evaluate an indicator compared with a target path. Evaluation plots are easy-to-understand illustrations of indicators over time for policy makers, which take variability and relevance into account. An implementation and detailed explanation of Evaluation plots are described in Deliverable 8.2 (Templ et al., 2011b).

Another possibility to visualize indicators are sparklines Tufte (2001). Sparklines are specific types of information graphics - usually barplots or polygonal lines - that are designed to be small in size and possess high data density. Sparkfunnels are extensions of sparkelines showing more detailed information. Sparkefunnels are applied to visualize Laeken indicators in graphical tables and in maps.

Indicators are often presented in maps. The map of Europe from Eurostat is given in the longitude and latitude coordinate system. Thus, projection onto other coordinate systems is absolutely necessary because extracting smaller regions such as countries results in extremely biased representations of the map otherwise. The aim was that projections should be made interactively when selecting countries for plotting by a mouse click on the underlying figure.

The aim was to visualize more than one kind of indicator on domains. Poverty risk groups were visualised by starplots and mosaicplots showing their multivariate dependencies.

Whenever missing values are imputed, one must be aware of the missing values mechanism(s). It can be shown that visualization tools provide deeper insights into the distribution of the missing values. Afterwards, a suitable imputation model could be chosen. The developed R functions are available in package VIM. Together with this information, a package vignette (also included in the Appendix) shows the application of the function to Austrian EU-SILC data.

For the simulations within the AMELI project, a framework has been implemented in the R package simFrame (Alfons et al., 2010c). Using lattice graphics, the results from different domains are displayed in separate panels of the plot. The simulation results can be visualised with parallel boxplots, density plots and lines. Depending on the structure of the simulation results, a suitable plot method is chosen automatically by the framework. Examples are included in the Appendix.

## 3.9 Support for policy (WP9)

To elaborate an improved methodology within the AMELI project it is important to get an overview of policy needs. Therefore, it was necessary indicate these policy needs and to transfer them into statistical objectives. Consequentially the main activity in this work package has been divided into several tasks. First of all the state-of-art in research on indicators and social reporting has been reviewed and documented. Further, a document study of relevant official documents was processed. The national strategy reports and the joint report have been in the focus of this study. In a third step an online survey was implemented to contact political intermediates and to gather information about the usage of the indicators. The target of an online survey was it to foster additional ideas and trends on the usage of indicators in the field of measurement of poverty and social exclusion. Often composite indicators are discussed to measure these multidimensional phenomena. Therefore, a sensitivity analysis on composite indicators was performed in a last step. The methodological research on this topic is documented in deliverable 9.1.

Starting point of the work in the analysis of the policy use of the indicators was the fact that the amount of indicators (currently eleven primary, eight secondary, and eleven context indicators) as well as the several changes the portfolio of indicators has undergone so far, complicated the use of this important policy-tool. To uncover the demands of the relevant actors and to evaluate the use of the Laeken Indicators the relevant literature have been analysed.

The Laeken indicators are considered to be a central part of the democratic process. The measurement of progress with consistent information helps to improve the evaluation of the policy process. Further, it is possible to improve the participation of a broader audience. If the indicators should be useful for these tasks different conditions have to be fulfilled. The consistent and harmonized reporting based on a reliable database is indispensable. Predetermined benchmarks could help to raise the level of interest in the evaluation of social processes.

Summarizing the questionnaire results it can be said that most respondents judge the indicators on poverty and social exclusion as a useful instrument without neglecting the problems which do occur in a framework of big diversities between the different European societies and subgroups. New dimensions of poverty have to be covered or fully excluded. New investments in data quality and infrastructure are also necessary for the indicators on poverty and social exclusion (and EU-SILC) to become the number one reference for policy making in the field of poverty and comparisons across countries. Steps have been taken to ramp up the social dimension and the fight against poverty, this development has to be supported and extended. Further research on ideas of how to foster the use of the indicators is indispensable and will be carried out.

To analyse the possible usage of composite indicators a sensitivity analysis on the construction scheme of the composite indicators was performed. Its main goal was to quantify the impact of the different constructions steps on the output. After a short overview about the general issue, two possible examples for composite indicators are illustrated. One is based on the synthetic Amelia data set set, while the other uses the Portfolio of indicators for the monitoring of the European strategy for social protection and social inclusion.

The result on the study about composite indicators is that composite indicators can have a major contribution to the analysis of the multidimensional phenomena of social exclusion and poverty but that a high attention has to be put on the construction scheme.

# Chapter 4

# References and description of R packages

## 4.1 SimFrame

In order to simplify using common guidelines in simulation experiments, a software framework has been developed in the R (R Development Core Team, 2011) package `simFrame` (Alfons et al., 2010c). It allows the use a wide range of simulation designs with a minimal effort of programming. In addition, the object-oriented implementation provides clear interfaces for further extensions.

Simulation studies in research projects such as AMELI require a precise outline. If different partners use, e.g., different contamination or missing data models, the results may be incomparable. A software framework for statistical simulation may thus contribute its share to avoid such problems. For this purpose, the R package `simFrame` (Alfons et al., 2010c) has been developed. The object-oriented implementation with S4 classes and methods (Chambers, 1998, 2008) gives maximum control over input and output and provides clear interfaces for user-defined extensions. Moreover, the framework allows a wide range of simulation designs to be used with only a little programming.

One of the main goals of the AMELI project is to improve the methodology for the indicators on social exclusion and poverty under typical data problems such as outliers and missing data. The package `simFrame` therefore allows to add certain proportions of outliers or non-response. In addition, depending on the structure of the simulation results, an appropriate plot method is selected automatically.

While the `simFrame` paper is published in Alfons et al. (2010c) and the basic structure is also outlined in Alfons et al. (2011a), a vignette shows the application of the package for EU-SILC. This package vignette is included in the Appendix (see also Alfons et al., 2010a).

## 4.2 SimPopulation

One aim of the AMELI project was to investigate robust estimation of the Laeken Indicators. For this purpose, ALFONS et al. (2011d) developed a data generation framework, which is implemented in the R package `simPopulation` (ALFONS and KRAFT, 2010; ALFONS et al., 2011d). Based on Austrian EU-SILC sample data, the synthetic population AAT-SILC was generated with this framework (see ALFONS et al., 2011b). AAT-SILC was designed to resemble a representative country. A further objective was that the population data should not contain any large outliers, as these are included in the samples during the simulations for full control over the amount of outliers (see ALFONS et al., 2011a).

While the `simPopulation` is published in ALFONS et al. (2011d), a package vignette in the appendix shows the application of the package.

## 4.3 VIM

Imputation of item non-responses in complex surveys has an effect on the final estimates of the indicators. One aim of the AMELI project was to develop robust methods for estimation (see HULLIGER et al., 2011b) and to visualize the structure of microdata (see TEMPL et al., 2011b). Package `VIM` (TEMPL et al., 2011d) allows to explore the data with missing values and learn about the structure of the missing values. Visualisation methods such as modified parallel coordinate plots, mosaic plots, scatterplot matrices, etc., have to be modified to deal with missing values and to show their structure.

EM-based regression imputation algorithms are mainly used to impute missing values automatically, i.e. such methods are very helpful in hand of subject matter specialists who are not statistical experts. Since data virtually always comes with outlying observations, robust methods for statistical estimation of missing values should be used here. The implemented algorithm (see TEMPL et al., 2011e) again is able to deal with all data challenges like representative and non-representative outliers and a mixture of different distributed variables, for example.

The free and open-source R package `VIM` provides a graphical user interface for users having no experience with R.

More about `VIM` is shown in TEMPL et al. (2011b) and in the package vignette in the appendix.

## 4.4 Laeken

One aim of the AMELI project was on estimation of social inclusion indicators. The methodology of estimating these indicators is implemented in package `Laeken` (ALFONS et al., 2011c). The package contains a subset of synthetically generated data for the European Union Statistics on Income and Living Conditions (EU-SILC), which is used in the code examples throughout the package. The package has an object-oriented design and different classes and subclasses are introduced.

In addition, robust semi parametric estimation (see Hulliger et al., 2011b) of social exclusion indicators is available. Special emphasis is thereby given to income inequality indicators, as the standard estimates for these indicators are highly influenced by outliers in the upper tail of the income distribution. This influence can be reduced by modelling the upper tail with a Pareto distribution in a robust manner.

Moreover, variance estimation methods are implemented in the package. To be more precise, it describes a general framework for estimating variance and confidence intervals of indicators under complex sampling designs. Currently, the package is focused on bootstrap approaches. While the naive bootstrap does not modify the weights of the bootstrap samples, a calibrated version allows to calibrate each bootstrap sample on auxiliary information before deriving the bootstrap replicate estimate.

The package vignettes are included in the appendix.

## 4.5 GB2

Package GB2 (Graf and Nedyalkova, 2010) implements the methods described in Graf et al. (2011b, Chapters 1-4). For the Generalized Beta distribution of the second kind (GB2) - density, distribution function, quantiles, moments are provided. Functions for the full log-likelihood, the profile log-likelihood and the scores are given. Formulae for various Laeken indicators under the GB2 are implemented. Package GB2 performs pseudo maximum likelihood estimation and non-linear least squares estimation of the model parameters and computes the design based variance of the parameters and the indicators by linearisation. It provides various plots for the visualization and analysis of the results.

## 4.6 Robust non-parametric QSR estimation

The `rqsr` package is an implementation of the robust, non-parametric QSR estimators. Namely, it enables the user to compute `TQSR`, `SQSR`, `BQSR`, and `MQSR`. In addition, it includes a device to compute variance estimates for all variants of the robustified QSR. For more details see Templ et al. (2011a, Appendix).

## 4.7 MODI

The `modi` package contains functions for robust multivariate outlier detection and imputation. The following detection methods are implemented: `BACON-EEM`, `TRC`, `GIMCD`, and `Epidemic Algorithm`. All algorithms can cope with both missing values and complex survey samples. Once the data have been processed by outlier-detection methods, one considers (robustly) imputing for the missing values and the declared outliers. The implemented imputation methods are `Gaussian imputation` (based on robustly estimated location and scatter), `Nearest Neighbour Imputation`, and `Reverse Epidemic Algorithm`. For more details see TEMPL et al. (2011a, Appendix).

## 4.8 rsae: Robust Small Area Estimation

The `rsae` package offers a general framework to robustly estimate area-level- and unit-level small area estimation (SAE) models. Once a particular model has been set up, it can be fitted by various robust methods (and also maximum likelihood). The `rsae` consists of two fitting modes: *default mode* and *safe mode*. The latter involves a high-breakdown-point regression estimator initialization and uses several numerical checks whether the iteration-specific estimates behave well. Currently, only Huber-type $M$-estimation is implemented. This method has assured super-linear convergence (given that (1) the amount of contamination is strictly below the breakdown point and (2) the model is properly specified). The high-breakdown-point $S$-estimator for mixed-level models will be included in the next release. Once the parameters of the Gaussian core model have been robustly estimated, we consider robustly predicting the random effects and the small-area means. Further, the package is shipped with several useful utility functions. For more details see TEMPL et al. (2011a, Appendix).

## 4.9 Specific R-Code

### 4.9.1 Work package 2

R functions have been programmed for small area estimation (SAE) of indicators on poverty and social exclusion. The indicators include at-risk-of poverty rate, the Gini coefficient, relative median at-risk-of poverty gap and quintile share ratio (S20/S80 ratio). Design-based estimators include direct estimators that do not use auxiliary data. The more advanced indirect model-assisted, model-based and composite estimators use auxiliary data at unit level or at aggregated level and generalized linear mixed models. We have fitted most of the mixed models with R functions `nlme` and `glmer` (package `lme4`). In addition, R function multinom of package `nnet` has been used. Technical description of SAE methodology is in LEHTONEN et al. (2011). Annex 1 (Manual of R codes) of LEHTONEN et al. (2011) includes a more detailed description of R codes. The R program codes can be found in separate AMELI deliverable files.

### 4.9.2 Work package 3

Variance estimation with the linearised variance estimators, was done with the help of the `survey` package (cf. TILLÉ and MATEI, 2011), the definition of the necessary `survey.design` objects can be found in HULLIGER et al. (2011a, section 9). The functions used to compute both point and variance estimates for the linearised estimators are given in the appendix of HULLIGER et al. (2011a) (see *R Functions for Computing Point and Variance Estimates*).

### 4.9.3 Work package 4

Specific code for outlier detection of semi-continuous variables can be found in Todorov (2011) Todorov et al. (2011, see also) and in Meraner (2010). Todorov et al. (2011) was written in a collaborative manner with UNIDO, the latter one, Meraner (2010), was written within the AMELI project where also details can be found in Hulliger et al. (2011b).

### 4.9.4 Work package 8

Code for mapping and projection of coordinates are described in Deliverable 8.2 Templ et al. (2011c).

Package `sparktable` Kowarik et al. (2010) includes methods to generate scalable graphical tables including various types of sparklines for publication in web and in publications. It is mainly developed by Statistics Austria but with minor contribution from the AMELI team.

R code for checkerplots and further visualisation tools will be made soon available as a R package.

# Bibliography

**Alfons, A.**, **Burgard, J. P.**, **Filzmoser, P.**, **Hulliger, B.**, **Kolb, J.-P.**, **Kraft, S.**, **Münnich, R.**, **Schoch, T.** and **Templ, M.** (**2011**a): *The AMELI Simulation Study*. Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

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

**Alfons, A.**, **Holzer, J.** and **Templ, M.** (**2011**c): laeken: Estimation of indicators on social exclusion and poverty. R package version 0.3.
URL http://CRAN.R-project.org/package=laeken

**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.** (**2011**d): *Simulation of close-to-reality population data for household surveys with application to EU-SILC*. Statistical Methods & Applications, DOI 10.1007/s10260-011-0163-2, to appear.
URL http://dx.doi.org/10.1007/s10260-011-0163-2

**Alfons, A.**, **Templ, M.** and **Filzmoser, P.** (**2009**): *On the influence of imputation methods on Laeken indicators: simulations and recommandations*. UNECE Work Session on Statistical Data Editing, October 5–7, 2009, Neuchâtel, Switzerland.

**Alfons, A.**, **Templ, M.** and **Filzmoser, P.** (**2010**a): *Applications of Statistical Simulation in the Case of EU-SILC: Using the R Package simFrame*. Journal of Statistical Software, 37 (3), p. 17, supplementary paper.
URL http://www.jstatsoft.org/v37/i03/

**Alfons, A.**, **Templ, M.** and **Filzmoser, P.** (**2010**b): *Contamination models in the R package `simFrame` for statistical simulation*. **Aivazian, S.**, **Filzmoser, P.** and **Kharin, Y.** (editors) Computer Data Analysis and Modeling: Complex Stochastic Data and Systems, vol. 2, pp. 178–181, Minsk, ISBN 978-985-476-848-9.

**Alfons, A.**, **Templ, M.** and **Filzmoser, P.** (**2010**c): *An object-oriented framework for statistical simulation: The R package simFrame*. Journal of Statistical Software, 37 (3), pp. 1–36.
URL http://www.jstatsoft.org/v37/i03/

**Alfons, A.**, **Templ, M.**, **Filzmoser, P.** and **Holzer, J.** (**2010**d): *A comparison of robust methods for Pareto tail modeling in the case of Laeken indicators.* **Borgelt, C.**, **González-Rodríguez, G.**, **Trutschnig, W.**, **Lubiano, M.**, **Gil, M.**, **Grzegorzewski, P.** and **Hryniewicz, O.** (editors) Combining Soft Computing and Statistical Methods in Data Analysis, *Advances in Intelligent and Soft Computing*, vol. 77, pp. 17–24, Heidelberg: Springer, ISBN 978-3-642-14745-6.

**Béguin, C.** and **Hulliger, B.** (**2004**): *Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations.* Journal of the Royal Statistical Society, Series A, 167 (2), pp. 275–294.

**Béguin, C.** and **Hulliger, B.** (**2008**): *The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data.* Survey Methodology, 341 (1), pp. 91–103.

**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

**Chambers, J.** (**1998**): Programming with Data. New York: Springer, ISBN 0-387-98503-4.

**Chambers, J.** (**2008**): Software for Data Analysis: Programming with R. New York: Springer, ISBN 978-0-387-75935-7.

**Fabrizi, E.**, **Ferrante, M. R.** and **Pacei, S.** (**2007**): *Small area estimation of average household income based on unit level models for panel data.* Survey Methodology, 33 (2), pp. 187–198.

**Graf, M.**, **Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Hulliger, B.**, **Lehtonen, R.**, **Meindl, B.**, **Münnich, R.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Wenger, A.** and **Zins, S.** (**2011**a): *State-of-the-art of Laeken Indicators.* Research Project Report WP1 – D1.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

**Graf, M.** and **Nedyalkova, D.** (**2010**): GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation. R package version 1.0.
URL http://CRAN.R-project.org/package=GB2

**Graf, M.**, **Nedyalkova, D.**, **Münnich, R.**, **Seger, J.** and **Zins, S.** (**2011**b): *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

**Graf, M.**, **Wenger, A.** and **Nedyalkova, D.** (**2011**c): *Description and Quality of the User Data Base.* Research Project Report WP5 – D5.1, 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**a): *Report on the Simulation Results.* Research Project Report WP7 – D7.1, 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

**Hulliger, B.** and **Lussmann, D.** (**2008**): *Bewertung der Nachhaltigkeits- und Umwelt-Indikatoren.* Bericht im Auftrag des Bundesamts für Statistik Sektion Umwelt, Nachhaltigkeit, Landwirtschaft, Institute for Competitiveness and Communication Hochschule für Wirtschaft Fachhochschule Nordwestschweiz.

**Hulliger, B.**, **Lussmann, D.**, **Kohler, F.**, **Mayerat, A.-M.** and **de Montmollin, A.** (**2008**): *Evaluation of Indicators on Environment and Sustainability.* European Conference on Quality in Official Statistics, Rome: Eurostat and ISTAT.

**Judkins, D. R.** and **Liu, J.** (**2000**): *Correcting the Bias in the Range of a Statistic Across Small Areas.* Journal of Official Statistics, 16 (1), p. 1–13.

**Kowarik, A.**, **Meindl, B.** and **Zechner, S.** (**2010**): sparkTable: Sparklines and graphical tables for tex and html. R package version 0.1.3.
URL http://CRAN.R-project.org/package=sparkTable

**Lehtonen, R.**, **Särndal, C.-E.** and **Veijanen, A.** (**2003**): *The effect of model choice in estimation for domains, including small domains.* Survey Methodology, 29 (1), pp. 33–44.

**Lehtonen, R.**, **Särndal, C.-E.** and **Veijanen, A.** (**2005**): *Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains.* Statistics in Transition, 7 (3), pp. 649–673.

**Lehtonen, R.** and **Veijanen, A.** (**2009**): *Design-based methods of estimation for domains and small areas.* **Rao, C. R.** and **Pfeffermann, D.** (editors) Handbook of Statistics 29, *Sample Surveys: Inference and Analysis*, vol. 29B, pp. 219–249, Amsterdam: Elsevier.

**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

**Meraner, A.** (**2010**): Outlier Detection for Semi-continuous Variables. Diplomarbeit, Institut f. Statistik und Wahrscheinlichkeitstheorie, Technische Universität, Wien.

**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** (**2011**): 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

**Templ, M.** and **Alfons, A.** (**2010**): *Disclosure risk of synthetic population data with application in the case of EU-SILC.* **Domingo-Ferrer, J.** and **Magkos, E.** (editors) Privacy in Statistical Databases, *Lecture Notes in Computer Science*, vol. 6344, pp. 174–186, Heidelberg: Springer, ISBN 978-3-642-15837-7.

**Templ, M.**, **Alfons, A.**, **Filzmoser, P.**, **Graf, M.**, **Holzer, J.**, **Hulliger, B.**, **Kowarik, A.**, **Kraft, S.**, **Lehtonen, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Veijanen, A.** and **Zins, S.** (**2011**a): *R Packages plus Manual.* Research Project Report WP10 – D10.3, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

**Templ, M.**, **Alfons, A.**, **Filzmoser, P.**, **Hulliger, B.** and **Lussmann, D.** (**2011**b): *Visualisation Tools.* Research Project Report WP8 – D8.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

**Templ, M.**, **Alfons, A.**, **Filzmoser, P.**, **Hulliger, B.** and **Lussmann, D.** (**2011**c): *Visualisation Tools.* Research Project Report WP8 – D8.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

**Templ, M.**, **Alfons, A.** and **Kowarik, A.** (**2011**d): VIM: Visualization and Imputation of Missing Values. R package version 2.0.1.
URL http://CRAN.R-project.org/package=VIM

**Templ, M.**, **Kowarik, A.** and **Filzmoser, P.** (**2011**e): *Iterative stepwise regression imputation using standard and robust methods.* Computational Statistics & Data Analysis, 55 (10), pp. 2793 – 2806, ISSN 0167-9473, doi:DOI:10.1016/j.csda.2011.04.012.
URL http://www.sciencedirect.com/science/article/pii/S0167947311001411

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

**Todorov, V.** (**2011**): rrcovNA: Scalable Robust Estimators with High Breakdown Point for Incomplete Data. R package version 0.4-02.
URL http://CRAN.R-project.org/package=rrcovNA

**Todorov, V.**, **Templ, M.** and **Filzmoser, P.** (**2011**): *Detection of multivariate outliers in business survey data with incomplete information.* Advances in Data Analysis and Classification, 5, pp. 37–56.
URL http://dx.doi.org/10.1007/s11634-010-0075-2

**Tufte, E.** (**2001**): The Visual Display of Quantitative Information. Graphics Press.