A Framework for Producing Small Area Estimates Based on Area-Level Models in R

by Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Schmidt, Nicola Salvati and Timo Schmid

Abstract The R package **emdi** facilitates the estimation of regionally disaggregated indicators using small area estimation methods and provides tools for model building, diagnostics, presenting, and exporting the results. The package version 1.1.7 includes unit-level small area models that rely on access to micro data. The area-level model by Fay and Herriot (1979) and various extensions have been added to the package since the release of version 2.0.0. These extensions include (a) area-level models with back-transformations, (b) spatial and robust extensions, (c) adjusted variance estimation methods, and (d) area-level models that account for measurement errors. Corresponding mean squared error estimators are implemented for assessing the uncertainty. User-friendly tools like a stepwise variable selection, model diagnostics, benchmarking options, high quality maps and results exportation options enable a complete analysis procedure. The functionality of the package is illustrated by examples based on synthetic data for Austrian districts.

1 Introduction

Small area estimation (SAE) enables better insight at smaller scales, for which it has gained importance in both academic and applied research. Among others, SAE is used for estimating socio-economic measures like income, poverty and health or indicators for agriculture (Datta et al., 1991; Tzavidis et al., 2012; Zhang et al., 2015; Pratesi, 2016). Economic or political decision-makers and official statistics practitioners especially benefit from reliable estimation of disaggregated indicators and thus SAE methods. Existing surveys were often not planned for analysis at disaggregated levels and show only small sample sizes, which leads to a low precision of the estimates. SAE methods can be employed to avoid expensive and time-consuming enlargements of the sample size of surveys. The idea is to combine data sources with model-based approaches. Existing survey data will be enriched by auxiliary information, e.g., from census or register data, to improve the accuracy of the indicator estimation on an area- or domain- level. The terms area and domain are used interchangeably and refer either to a geographic area or to any subpopulation of a population of interest, like socio-demographic groups. Among others, Pfeffermann (2013), Rao and Molina (2015), Tzavidis et al. (2018) and Jiang and Rao (2020) give comprehensive overviews of SAE methods.

The main goal of the package **emdi** is the simplification of estimating these regionally disaggregated indicators. The package version 1.1.7 contains direct estimation based exclusively on survey data and model-based estimation using the unit-level empirical best predictor (EBP) method (Molina and Rao, 2010). The EBP approach is powerful since it enables the simultaneous estimation of various indicators. For this, it relies on unit-level information, i.e. information about each unit in each domain. Though survey data often provides unit-level information, access to census or register data at unit-level is less likely. Hence, area-level models provide a valuable alternative, with the following benefits: First, only area-level aggregates are needed to estimate the regional indicators. Second, area-level models can consider the survey design by integrating the sampling weights. Third, the computation is faster compared to the computational intensive EBP approach.

Various R packages that employ different area-level models are available on the Comprehensive R Archive Network (CRAN). The package **smallarea** (Nandy, 2015) offers several variance estimation methods for the standard Fay-Herriot (FH) model: maximum likelihood (ML), residual maximum likelihood (REML), and both Prasad-Rao and Fay-Herriot method-of-moment. Estimation of unknown sampling variances is also offered." The ability to estimate unit- and area-level models under heteroscedasticity is implemented by the **JoSAE** package (Breidenbach, 2018). Robust estimation of area-level models with spatial and/or temporal structures in the random effects is supported by package **saeRobust** (Warnholz, 2022). The **mcmcsae** package (Boonstra, 2021) also takes spatial and temporal correlation of the random effects into account, but fits unit- and area-level models via Markov Chain Monte Carlo simulation. Estimation of univariate and multivariate FH models is possible with package **msae** (Permatasari and Ubaidillah, 2022). The package **hbsae** (Boonstra, 2022) allows for the fitting of unit- and area-level models by frequentist or hierarchical Bayesian approaches. The possibility of estimating FH models and some of its extensions in a Bayesian framework is also given by the **BayesSAE** package (Developer, 2018). The **tipsae** package (De Nicolò and Gardini, 2022) provides estimation and mapping tools within a Bayesian setting for proportions that are defined

	Package								
Area-level model	*smallarea*	*JoSAE*	*sae*	*saeRobust*	*msae*	*hbsae*	*BayesSAE*	*saeME*	*emdi*
Standard variance estimation	√				√	√			√
Adjusted variance estimation									√
Unknown sampling variances	√								
Heteroscedasticity		√							
Spatial correlation			√						√
Spatio-temporal correlation			√						
Robust				√					√
Robust, spatial correlation				√					√
Robust, (spatio-)temporal correlation				√					
Multivariate					√				
Bayesian formulation						√	√		
Measurement error								√	√
Transformation (log, arcsin)									√

Table 1: Overview of selected implemented area-level models in R packages available on CRAN.

on the unit interval. The **mme** package (Lopez-Vizcaino et al., 2019) implements Gaussian area-level multinomial mixed-effects models in the SAE context. The **saeME** package (Mubarak and Ubaidillah, 2022) can fit an area-level model when the auxiliary variables are measured with error. The **NSAE** package (Chandra et al., 2022) can fit stationary and nonstationary FH models. One of the commonly used packages is the **sae** package (Molina and Marhuenda, 2015). It includes a wide range of area-level models (the standard FH model with REML, ML and FH method-of-moment model fitting and a spatial and a spatio-temporal extension of the FH model) and unit-level models (the nested error linear regression model of Battese et al. (1988) and the EBP approach). Table 1 gives an overview of selected packages and the implemented methodology.
Besides packages that include particular area-level models, the packages **saeMSPE** (Xiao et al., 2022) and **SAEval** (Fasulo, 2022) offer different analytical- and resampling-based MSE estimators and tools for diagnostics and graphical evaluation of SAE models, respectively.

The latest version of package **emdi** 2.1.3 combines a wide range of SAE models with several tools that enable a complete analysis, and therefore adds to the space of useful packages, for the following reasons:

- None of the existing packages contains such a variety of different area-level models.
- In addition to models that are already available in existing R packages, **emdi** includes: adjusted variance estimation methods and transformation options for the standard FH model. Adjusted variance estimation methods are of particular importance when working in a non-Bayesian framework. In a Bayesian context, the variance will always be estimated as strictly positive, so packages providing a Bayesian approach do not need adjusted variance estimation methods.
- Package **emdi** offers user-friendly tools that go beyond model estimation: diagnostic tools with both summary and graphical results, benchmarking options, high-quality geographical visualization of results, and export of results to Excel and OpenDocument Spreadsheet formats.
- Plus a stepwise variable selection algorithm for area-level models is included in **emdi** to allow the user to build a model based on information criteria.

Thus, since package version 2.0.0, version 1.1.7 has been extended by various area-level models, but stays in line with the user-friendly orientation of the existing version.

The structure of the paper can be described as follows. The section Statistical methodology introduces the statistical methods implemented in the package. The example data sets included in the package are presented in the section Data sets. The section Functionality and case studies provides an illustrative description of the functions using the example data sets. The section Estimation procedure for the standard Fay-Herriot model guides the reader from model-building to diagnostics of a standard FH model and creating maps of the results. The section Estimation of the extended area-level models follows with short descriptions of how to build the different extended area-level models. Finally, the section Conclusion and outlook summarizes our contributions and gives an outlook.

2 Statistical methodology

Area-level models for the estimation of indicators like means, totals or shares have been added to the package since the release of version 2.0.0. These comprise the area-level model by Fay and Herriot (1979) and several extensions of this standard model which account for issues that may come up in real data applications. To measure the precision of those models, MSE estimators have been integrated following the literature.

Standard Fay-Herriot model

Throughout the paper, a finite population U is assumed that consists of N units that are subdivided into D domains or areas of specific sizes $N_1, ..., N_D$. Then a random sample of size n can be drawn from U and partitioned into D areas with $n_1, ..., n_D$ observations per domain.

The FH model links area-level direct estimators based on survey data to covariates aggregated on an area level that may stem from administrative data (e.g. register or census) or alternative data sources (e.g. satellite, social media or mobile phone data). The FH model is composed of two levels. The first level is the sampling model

$$\hat{\theta}_i^{\text{Dir}} = \theta_i + e_i, \quad i = 1, \dots, D.$$

$\hat{\theta}_i^{\text{Dir}}$ is an unbiased direct estimator for a population indicator of interest θ_i, for instance, a mean or a ratio. e_i represents independent and normally distributed sampling errors with $e_i \overset{ind}{\sim} N\left(0, \sigma_{e_i}^2\right)$. Though the model assumes known sampling variances, in practical applications $\sigma_{e_i}^2$ are usually unknown, and have to be estimated from the unit-level sample data (Rivest and Vandal, 2003; Wang and Fuller, 2003; You and Chapman, 2006). Package **emdi** provides a non-parametric bootstrap for estimating the variances of the direct estimator (Alfons and Templ, 2013). To allow for complex survey designs, sampling weights (w) can be considered in the direct estimation (Horvitz and Thompson, 1952). For example, an estimator for the population mean θ_i of a continuous variable of interest y for each area i is estimated by

$$\hat{\theta}_i^{\text{Dir}} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

where the index j indicates an individual with $j = 1, ..., n_i$ in the i-th area. The second FH level links the target indicator θ_i linearly to area-specific covariates $\boldsymbol{x}_i$,

$$\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i,$$

where $\boldsymbol{\beta}$ is a vector of unknown fixed-effect parameters, and u_i is an independent and identically normally distributed random effect with $u_i \overset{iid}{\sim} N\left(0, \sigma_u^2\right)$.
The combination of the sampling and the linking model leads to a special linear mixed model

$$\hat{\theta}_i^{\text{Dir}} = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i + e_i, \quad i = 1, \dots, D. \tag{1}$$

The empirical best linear unbiased estimators $\hat{\boldsymbol{\beta}}$ are computed by weighted least square theory. The empirical best linear unbiased predictor (EBLUP) of θ_i is obtained by substituting the variance parameter σ_u^2 with an estimate. The resulting estimator can then be written as

$$\begin{aligned} \hat{\theta}_i^{\text{FH}} &= \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \hat{\gamma}_i \hat{\theta}_i^{\text{Dir}} + (1 - \hat{\gamma}_i)\, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}. \end{aligned} \tag{2}$$

The EBLUP/FH estimator can be understood as a weighted average of the direct estimator $\hat{\theta}_i^{\text{Dir}}$ and a regression-synthetic part $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$. The estimated shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ puts more weight on the direct estimator when the sampling variance is small and vice versa. Areas for which no direct estimation results exist, because the sample size is zero or the results may not be published, are called out-of-sample domains. For those domains, the prediction reduces to the regression-synthetic component $\hat{\theta}_{i,\text{out}}^{\text{FH}} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ (Rao and Molina, 2015).

Estimation methods for σ_u^2
The variance of the random effects has to be estimated. Commonly used approaches are the FH method-of-moment estimator (Fay and Herriot, 1979), the ML, and the REML estimators (Rao and Molina, 2015). The likelihood methods are known to perform more efficiently than the methods of moments (Rao and Molina, 2015). The commonly used methods can produce negative variance

estimates that should be strictly positive. In the estimation methods mentioned above, negative variance estimates are set to zero ($\hat{\sigma}_u^2 = \max\left(\tilde{\sigma}_u^2, 0\right)$) resulting in zero estimates of the shrinkage factor γ_i. Therefore, no weight is put on the direct estimator, ignoring its possible reliability. This poses a problem, especially when the number of areas is small. To avoid this so-called over-shrinkage problem, Li and Lahiri (2010) and Yoshimori and Lahiri (2014) proposed methods that adjust the respective likelihoods of the standard ML and REML approaches by some factor:

$$L_{\text{adj}}\left(\sigma_u^2\right) = A \times L\left(\sigma_u^2\right),$$

where A denotes the adjustment factor and $L(\sigma_u^2)$ the given likelihood function. The proposed adjustment factors are:

- by Li and Lahiri (2010): $A = \sigma_u^2$,
- by Yoshimori and Lahiri (2014): $A = \left(\tan^{-1}\left(\sum_{i=1}^{D} \gamma_i\right)\right)^{1/D}$.

Simulation studies conducted by Yoshimori and Lahiri (2014) showed that the adjusted Yoshimori-Lahiri methods are preferable when the variance of the random effect is small relative to the sampling variance. Otherwise, the adjusted Li-Lahiri methods are recommended. Package **emdi** offers six different variance estimation methods: standard ML (`ml`) and REML (`reml`), and adjusted ML and REML following either Li and Lahiri (2010) (`amrl`, `ampl`) or Yoshimori and Lahiri (2014) (`amrl_yl`, `ampl_yl`).

Extended area-level models

In real data applications, problems might occur that were not theoretically expected. There may also be some violation of the assumptions of the standard FH model, e.g., normality and independence of the error terms. The following section outlines the extensions of the standard FH model that are implemented in the package **emdi** to address these issues.

Transformations

When working with right-skewed data like income, wealth or business data, the assumptions of a linear relation between the response and the explanatory variables and normality of both error terms (u_i and e_i) of the FH model may be violated. Applying a log-transformation could be a reasonable solution to meet these model assumptions (Neves et al., 2013; Kreutzmann et al., 2022). In the **emdi** package, the direct estimates and their variances are transformed following Neves et al. (2013):

$$\hat{\theta}_i^{\text{Dir*log}} = \log\left(\hat{\theta}_i^{\text{Dir}}\right),$$
$$\text{var}\left(\hat{\theta}_i^{\text{Dir*log}}\right) = \left(\hat{\theta}_i^{\text{Dir}}\right)^{-2} \text{var}\left(\hat{\theta}_i^{\text{Dir}}\right),$$

where the *log notation stands for the logarithmic transformed scale. To obtain the FH estimator on the transformed scale $\hat{\theta}_i^{\text{FH*log}}$, $\hat{\theta}_i^{\text{Dir}}$ is substituted by $\hat{\theta}_i^{\text{Dir*log}}$ and $\text{var}(\hat{\theta}_i^{\text{Dir*log}})$ serves as estimate for the sampling variances ($\sigma_{e_i}^2$) in Equation 2. Since the logarithm is a nonlinear transformation, the final FH estimates on the original scale require a bias-corrected back-transformation (Slud and Maiti, 2006; Sugawasa and Kubokawa, 2017). The **emdi** package provides two options:

1. A *crude* method (`bc_crude`) that takes the properties of the log-normal distribution into account:

$$\hat{\theta}_i^{\text{FH, crude}} = \exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\text{MSE}\left(\hat{\theta}_i^{\text{FH*log}}\right)\right\}.$$

2. A bias correction suggested by Slud and Maiti (2006) (`bc_sm`) that further regards the bias due to the random effects:

$$\hat{\theta}_i^{\text{FH, Slud-Maiti}} = \exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\hat{\sigma}_u^2\left(1 - \hat{\gamma}_i^{\text{*log}}\right)\right\}.$$

The FH estimator on the transformed scale is denoted by $\hat{\theta}_i^{\text{FH*log}}$ and, accordingly $\text{MSE}(\hat{\theta}_i^{\text{FH*log}})$ stands for a MSE estimator on the transformed scale, e.g., the Prasad-Rao or Datta-Lahiri MSE (cf. following subsection). The Slud-Maiti back-transformation is derived for the ML variance estimation of the random effect and is implemented for in-sample domains in **emdi**. In the presence of out-of-sample domains, the *crude* method can be applied, which allows to use also other variance estimation methods.

Another transformation provided by the **emdi** package is the arcsin transformation, which is widely used when the direct estimator of the FH model is a ratio (Casas-Cordero et al., 2016; Schmid et al., 2017). The **emdi** package automatically transforms the direct estimates and the sampling variances as suggested by Jiang et al. (2001):

$$\hat{\theta}_i^{\text{Dir*arcsin}} = \sin^{-1}\left(\sqrt{(\hat{\theta}_i^{\text{Dir}})}\right),$$
$$\text{var}\left(\hat{\theta}_i^{\text{Dir*arcsin}}\right) = 1/\left(4\tilde{n}_i\right),$$

where the *arcsin denotes the arcsin transformed scale, and $\tilde{n}_i$ the effective sample size: the sample size adjusted by the sampling design (Jiang et al., 2001). The FH model is estimated using Equation 2 and, if necessary, the results are truncated to the interval $[0, \pi/2]$ to ensure final results between 0 and 1. To obtain final estimates on the original scale, the final estimation results must be subjected to a back-transformation. Two different back-transformations are available in **emdi**:

1. A *naive* back-transformation (`naive`):

$$\hat{\theta}_i^{\text{FH, naive}} = \sin^2\left(\hat{\theta}_i^{\text{FH*arcsin}}\right).$$

2. A back-transformation with bias-correction (bc) following Sugawasa and Kubokawa (2017) and Hadam et al. (2020):

$$\hat{\theta}_i^{\text{FH, bc}} = \int_{-\infty}^{\infty} \sin^2(t)\, \frac{1}{2\pi\frac{\hat{\sigma}_u^2\sigma_{\hat{e}_i}^2}{\hat{\sigma}_u^2+\sigma_{\hat{e}_i}^2}} \exp\left(-\frac{\left(t-\hat{\theta}_i^{\text{FH*arcsin}}\right)^2}{2\frac{\hat{\sigma}_u^2\sigma_{\hat{e}_i}^2}{\hat{\sigma}_u^2+\sigma_{\hat{e}_i}^2}}\right) dt.$$

Spatial FH model

The standard FH model assumes independence of the random effects. However, when working with geographical areas, assuming correlated random effects to incorporate a certain neighbouring structure can be valuable. The **emdi** package contains the spatial FH model introduced by Petrucci and Salvati (2006) that considers a simultaneously autoregressive process of order one, SAR(1). Compared to the standard model, the estimation differs mainly by discarding the assumptions of independent random effects and estimating a spatial autoregressive coefficient (ρ) which takes values between -1 and 1. Greater absolute values of (ρ) indicate a stronger relationship with the neighboring areas. The random effect u_i in Equation 1 is replaced by

$$\boldsymbol{u} = \rho_1 \boldsymbol{W}\boldsymbol{u} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}_D, \sigma_1^2 \boldsymbol{I}_D\right), \tag{3}$$

with $\boldsymbol{W}$ being the $D \times D$ row standardized proximity matrix that describes the neighbourhood structure of the areas, $\boldsymbol{0}_D$ a vector of zeros and $\boldsymbol{I}_D$ the $D \times D$ identity matrix. The random effects $\boldsymbol{u}$ of Equation 3 follow a SAR(1). When normality of the random effects is assumed, the model can be fitted by ML and REML. The application of spatial FH models should be considered when no geographic auxiliary variables are available to capture the spatial relation, or when ρ_1 is larger than 0.5 (Bertarelli et al., 2021). Even before estimating the model, the **emdi** package enables testing for spatial correlation by Moran's I and Geary's C statistics (Cliff and Ord, 1981; Pratesi and Salvati, 2008). While Moran's I mimics a typical correlation coefficient whose values range from -1 and 1, Geary's C takes values between 0 and 2 (0: positive, 1: no, 2: negative spatial autocorrelation). The two statistics behave inversely to each other.

Robust area-level models

In the case of influential outlying observations, the **emdi** package allows for robust versions of the standard and the spatial FH model. The theory is extensively studied in Warnholz (2016), wherein the robust estimation procedure for linear mixed models suggested by Sinha and Rao (2009) was extended to area-level models. The model fitting can be understood as a robustified ML version that also contains an influence function with a tuning constant `k`. 1.345 is recommended as an initial value for the tuning constant (Sinha and Rao, 2009). When non-symmetric outliers are expected to influence the robust estimation, a bias correction should be involved. This correction can be controlled by a multiplier constant (`mult_constant`). For further details, we also refer to Chambers et al. (2014) and Schmid et al. (2016).

Measurement error model

The standard FH model is based on the assumption that the covariates are measured without error (Fay and Herriot, 1979). This characteristic is typically assumed because census or register data are used as auxiliary information. However, when the covariate information stems from larger

Model	Type of MSE	Reference
Standard FH (depending on variance estimation of σ_u^2)		
`ml/ampl_yl`	Analytical	Datta and Lahiri (2000)
`reml/amrl_yl`	Analytical	Prasad and Rao (1990)
`ampl/amrl`	Analytical	Li and Lahiri (2010)
`ml/reml` (out-of-sample)	Analytical	Rao and Molina (2015)
Transformations		
`log` (depending on back-transformation)		
`bc_crude`	Analytical	Rao and Molina (2015)
`bc_sm`	Analytical	Slud and Maiti (2006)
`arcsin` (depending on back-transformation)		
`naive`	Jackknife	Jiang et al. (2001)
	Weighted Jackknife	Jiang et al. (2001); Chen and Lahiri (2002)
	Parametric bootstrap	Hadam et al. (2020)
`bc`	Parametric bootstrap	Hadam et al. (2020)
Spatial FH (depending on variance estimation)		
`ml/reml`	Analytical	Singh et al. (2005)
`ml/reml`	Parametric bootstrap	Molina et al. (2009)
`reml`	Nonparametric bootstrap	Molina et al. (2009)
Robust FH		
	Pseudolinear	Warnholz (2016)
	Parametric bootstrap	Warnholz (2016)
FH with ME		
	Jackknife	Jiang et al. (2002)

Table 2: Overview of the MSE estimation options of the `fh` function.

surveys or alternative data sources, this assumption can be violated. The **emdi** package includes an implementation of the measurement error (ME) model developed by Ybarra and Lohr (2008). To account for the ME in the covariates $\boldsymbol{x}_i$, they modified the shrinkage factor as follows:

$$\gamma_i = \frac{\sigma_u^2 + \boldsymbol{\beta}^\top \boldsymbol{C}_i \boldsymbol{\beta}}{\sigma_u^2 + \boldsymbol{\beta}^\top \boldsymbol{C}_i \boldsymbol{\beta} + \sigma_{e_i}^2},$$

where the $\boldsymbol{C}_i$ stands for the variance-covariance matrix of the covariates, which is a required prerequisite for the model. The modified shrinkage factor pulls more weight on the direct estimator when the variances of the covariates are large. A modified weighted least squares method and a moment estimator were used to estimate $\boldsymbol{\beta}$s and the σ_u^2, respectively. Additional details are available in Ybarra and Lohr (2008).

Mean squared error estimation

To evaluate the accuracy of the EBLUP estimates, the MSE is the most common measure used in SAE (Rao and Molina, 2015). The **emdi** package offers a variety of MSE estimators stemming from both analytical determination and resampling strategies, like the bootstrap and jackknife methods. Table 2 gives an overview of the included MSE approaches. For each area-level model presented in the previous sections, the provided MSE types are shown. The quoted references detail extensive formulas and derivations. As an additional measure of variability of the direct and FH estimates, within various functions and methods of the **emdi** package, the coefficient of variation (CV) is provided: $CV = \sqrt{\widehat{\text{MSE}}(\hat{\theta}_i)}/\hat{\theta}_i$, where $\hat{\theta}_i$ either stands for $\hat{\theta}_i^{\text{Dir}}$ or $\hat{\theta}_i^{\text{FH}}$.

3 Data sets

The **emdi** package version 1.1.7 contains a sample 'eusilcA_smp' and a population data set 'eusilcA_pop' at a household level. The generation process for both data sets is extensively described in Kreutzmann et al. (2019). Our process is nearly equivalent, but we do not produce out-of-sample domains for the

Variable	Meaning
Sample data set	
`Domain`	Austrian districts
`Mean`	Mean of the equivalized household income
`MTMED`	Share of households who earn more than the national median income
`Cash`	Mean employee cash or near cash income
`Var_Mean`	Variance of equivalized household income
`Var_MTMED`	Variance of share of households who earn more than the national median income
`Var_Cash`	Variance of employee cash or near cash income
`n`	Effective sample sizes
Population data set	
`Domain`	Austrian districts
`eqsize`	Equivalized household size according to the modified OECD scale
`cash`	Employee cash or near cash income
`self_empl`	Cash benefits or losses from self-employment (net)
`unempl_ben`	Unemployment benefits (net)
`age_ben`	Old-age benefits (net)
`surv_ben`	Survivor's benefits (net)
`sick_ben`	Sickness benefits (net)
`dis_ben`	Disability benefits (net)
`rent`	Income from rental of a property or land (net)
`fam_allow`	Family/children related allowances (net)
`house_allow`	Housing allowances (net)
`cap_inv`	Interest, dividends, profit from capital investments in unincorporated business (net)
`tax_adj`	Repayments/receipts for tax adjustment (net)
`ratio_n`	Ratios of the population size per area and the total population size

Table 3: Variables of the aggregated data sets. The `Domain` variables are factors, the rest of the variables are numeric. Except for the variables `Domain` and `ratio_n`, the observations of all variables of the population data set consist of the mean values per district.

area-level version of the data sets. The Austrian European Union Statistics on Income and Living Conditions (EU-SILC) synthetic 2006 data set (`eusilcP`) sourced from the **simFrame** package (Alfons et al., 2010) serves as basis for our data sets. The lowest regional level in the `eusilcP` data set consists of the nine Austrian states. Based on certain population size and income criteria, households were allocated to 94 Austrian districts resulting in the synthetic population data set 'eusilcA_pop'. For the 'eusilcA_smp' data set, a sample was drawn following a stratified random sampling process using the districts as strata. To show the usage of the FH model and its extensions, area-level data is required. The area-level survey and population data sets, 'eusilcA_smpAgg' and 'eusilcA_popAgg', are obtained by aggregation on the district level with the help of the `direct` function defined by the **emdi** package. The direct estimates in 'eusilcA_smpAgg' are the weighted mean equivalized household income `Mean`, the ratio of households that earn more than the national median income (`MTMED`) and their variances. These are based on the equivalized household income `eqIncome` in 'eusilcA_smp', defined as the total income of a household divided by the size of the household, with household size equalized by the modified equivalence scale of the Organisation for Economic Co-operation and Development (OECD) (Hagenaars et al., 1994). Additionally, the mean of the variable `cash`, its variance and the sample sizes are included in 'eusilcA_smpAgg', being required by the model extensions. The population data set 'eusilcA_popAgg' contains a variety of variables that describe different income sources of households, and a variable `ratio_n` that describes the ratios of the population sizes per area and the total population size. The variable `Domain` exists in both data sets and identifies the different districts. Both data sets have an observation for each of the 94 Austrian districts, with the sample data set 'eusilcA_smpAgg' containing eight variables and the population data set 'eusilcA_popAgg' containing fifteen. Table 3

provides an overview of all included variables of the sample and population data sets. For the creation of the proximity matrix used in the spatial FH model and the usage of the `map_plot` function, a shape file is needed. A shape file 'shape_austria_dis' (`.rda` format, `"SpatialPolygonsDataFrame"`) for the 94 districts of Austria is provided. This file was sourced from the SynerGIS website (Bundesamt für Eich- und Vermessungswesen, 2017). The data set 'eusilcA_prox', an example proximity matrix, has also been added to the **emdi** package. The creation of 'eusilcA_prox' is described in the following section.

4 Functionality and case studies

While the theoretical background of the implemented area-level models has been introduced in the section Statistical methodology, the focus of this section lies on the functionality and the workflow of their usage in R. All of the contained area-level models can be applied by one function: `fh`. Table 4 gives an overview of the 20 input arguments of function `fh`, together with a short description and any default settings. Not every argument needs a specification for every estimated model. Depending

Argument	Description	Default
`fixed`	Formula of fixed-effects part of linear mixed model	
`vardir`	Domain-specific sampling variances of the direct estimates	
`combined_data`	Combined sample and census data set	
`domains`	Domain indentifier for `combined_data`	NULL
`method`	Model fitting method	reml
`interval`	Lower and upper limit for the variance estimation	NULL
`k`	Tuning constant for robust estimation	1.345
`mult_constant`	Bias correction multiplier constant for robust estimation	1
`transformation`	Type of transformation	no
`backtransformation`	Type of back-transformation	NULL
`eff_smpsize`	Effective sample sizes for the arcsin transformation	NULL
`correlation`	Correlation of random effects	no
`corMatrix`	Proximity matrix for the spatial model	NULL
`Ci`	Array of the variance-covariance matrix of the explanatory variables for each area for the ME model	NULL
`tol`	Tolerance value for the variance estimation	0.0001
`maxit`	Maximum number of iterations for the variance estimation	100
`MSE`	MSE estimation	FALSE
`mse_type`	Type of MSE estimator	analytical
`B`	Numbers of bootstrap iterations for computation of a bootstrap MSE and information criteria by Marhuenda et al. (2014)	c(50,0)
`seed`	Seed for random number generator	123

Table 4: Input arguments of function `fh`.

on the area-level model, different arguments have to be determined (see Table 6 in Appendix A). Figure 1 demonstrates the estimation possibilities of a standard FH model (for the extended area-level models see Figure 6 in Appendix A). In line with the `direct` and `ebp` functions of package version 1.1.7, the S3 object system is used for function `fh` (Chambers and Hastie, 1992). All three return objects of class `"emdi"`. The application of function `direct` leads to a `"direct"` object, and of functions `ebp` and `fh` to objects of classes `"ebp"` and `"fh"`, respectively. Though all of the returned objects contain ten components, not every component is available for each estimation method. In these cases they are indicated as `NULL` (see Table 5). Furthermore, the `model` component differs for the two classes `"ebp"` and `"fh"`. The components of objects of class `"fh"` are provided in Table 7 in Appendix B. Not all of

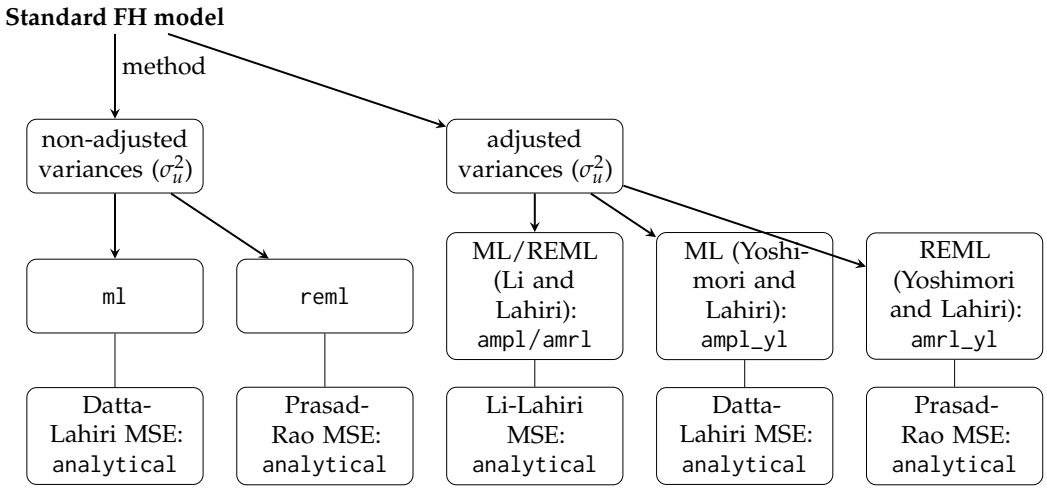

Figure 1: Overview of the standard FH model and adjusted variance estimation methods.

the components are available for every area-level model, e.g., the shrinkage factors per domain are not provided for the spatial and robust model extensions, as they do not have an intuitive interpretation in those cases. Due to the consistent structure, all functions and methods of **emdi** version 1.1.7 can be applied to objects of class `"fh"`. Additionally, new functions and methods are available for the area-level models. Furthermore, a variety of methods that are available in base R and used by other model fitting R packages are included in the latest package version 2.1.3 for the different `"emdi"` objects. Two examples of the new generic functions used are `coef` and `logLik`. Figure 2 demonstrates the steps of a full data analysis procedure and the respective functions, from model building and diagnostics to presenting the results. The section Estimation procedure for the standard Fay-Herriot model demonstrates the procedure shown in Figure 2 by applying the standard FH model to the Austrian EU-SILC data described in the section Data sets. To demonstrate how the different extended area-level models are fitted with function `fh`, the section Estimation of the extended area-level models follows.

	Name	Description	Available for		
			`direct`	`ebp`	`fh`
1	`ind`	Point estimates per area	√	√	√
2	`MSE`	Variance/MSE estimates per area	√	√	√
3	`transform_param`	Transformation and shift parameters		√	
4	`model`	Fitted model		√	√
5	`framework`	List for data description	√	√	√
6	`transformation`	Type of transformation		√	√
7	`method`	Estimation method		√	√
8	`fixed`	Formula of fixed effects		√	√
9	`call`	Function call	√	√	√
10	`successful_bootstraps`	Number of successful bootstraps	√		√

Table 5: The ten `"emdi"` object components distinguished in `"direct"`, `"ebp"` and `"fh"`. More detailed information is provided by the package documentation.

Estimation procedure for the standard Fay-Herriot model

The aim of this example is to estimate the equivalized income for the 94 Austrian districts. The package and the example data sets are loaded as follows:

```
> library("emdi")
> data("eusilcA_popAgg")
> data("eusilcA_smpAgg")
```

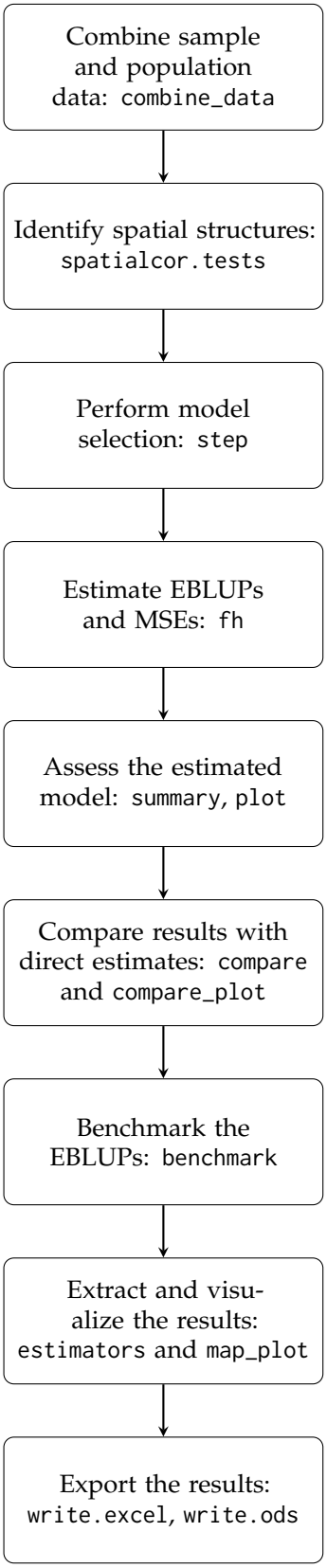

Figure 2: Estimation procedure for area-level models.

Combine input data

The function `fh` requires one data set (argument `combined_data`) that comprises the sample and population data. Thus, the data set must contain all variables of the formula object `fixed`, the variances of the direct estimates and, optionally, a domain identifier. For cases where the sample and population data are only available separately, a merging function `combine_data` is provided. The necessary arguments are the two data sets and characters specifying the domain indicator for the respective data sets.

```
> combined_data <- combine_data(
+   pop_data = eusilcA_popAgg, pop_domains = "Domain",
+   smp_data = eusilcA_smpAgg, smp_domains = "Domain")
```

Identify spatial structures

With the help of a proximity matrix, Moran's I and Geary's C test statistics can be computed to identify spatial structures by the `spatialcor.tests` command. For the creation of the proximity matrix, the shapefile must be loaded. We load the Austrian shapefile that is provided and merge it to the sample data set by using the respective domain identifiers with the help of the `merge` method from the **sp** package (Pebesma and Bivand, 2005). Before merging, we sort the Austrian shapefile by the domains in the sample data.

```
> library("sp")
> load_shapeaustria()
> shape_austria_dis <- shape_austria_dis[
+   order(shape_austria_dis\$PB),]
> austria_shape <- merge(shape_austria_dis,
+   eusilcA_smpAgg, by.x = "PB", by.y = "Domain",
+   all.x = F)
```

Then the `poly2nb` and `nb2mat` functions of the **spdep** package (Bivand and Wong, 2018) are used. While `poly2nb` generates a list of neighbours that share joint boundaries, `nb2mat` computes a weights matrix. The `style` argument has to be set to `W`, as a row standardized proximity matrix is required.

```
> library("spdep")
> rel <- poly2nb(austria_shape,
+   row.names = austria_shape$PB)
> eusilcA_prox <- nb2mat(rel, style = "W",
+   zero.policy = TRUE)
```

Thus, a row standardized proximity matrix is generated that initially had weights of one if an area shares a boundary with another area and zero if not. Function `spatialcor.tests` makes use of the `moran.test` and `geary.test` functions with their respective default settings, from the **spdep** package. The input arguments are the created matrix and the direct estimates.

```
> spatialcor.tests(direct = combined_data$Mean,
+   corMatrix = eusilcA_prox)

 Statistics     Value      p.value
1 Moran's I 0.2453677 5.607958e-05
2 Geary's C 0.6238681 2.473294e-03
```

Since the output indicates only a weak positive spatial autocorrelation, the following estimation procedure does not consider the integration of a correlation structure for the random effects.

Perform model selection
Besides theoretical considerations on which auxiliary variables should be part of the model, the decision for the best model should be based on information criteria like the Akaike or Bayesian information criterion (AIC, BIC). Many applications use selection techniques based on linear regression (Casas-Cordero et al., 2016; Schmid et al., 2017). Instead, the **emdi** package provides the AIC, BIC, Kullback information criterion (KIC) and their bootstrap and bias corrected versions (AICc, AICb1, AICb2, KICc, KICb1, KICb2) especially developed for FH models by Marhuenda et al. (2014). These criteria are also included in the **sae** package, but the **emdi** package enables a stepwise variable selection procedure based on the chosen information criteria, comparable to the `step` function for `lm` models of package **stats**. The most important input arguments are an object of class `"fh"` and the direction of the stepwise search (`both`, `backward`, `forward`). In this example, the default setting `backward` and the `KICb2` information criterion is used. In the `fixed` argument of the `fh` function, the variables employee cash (`cash`), cash benefits from self-employment (`self_empl`) and unemployment benefits (`unempl_ben`) are included. For a valid comparison of models based on information criteria, the model fitting method must be `ml`. To activate the estimation of the information criteria by Marhuenda et al. (2014), we set the number of bootstrap iterations to 50. The output shows the stepwise removal of variables until the lowest KICb2 is reached, the function call and an overview of the estimated coefficients of the final recommended model.

```
> fh_std <- fh(fixed = Mean ~ cash + self_empl + unempl_ben, vardir = "Var_Mean",
+   combined_data = combined_data, domains = "Domain", method = "ml", B = c(0,50))
> step(fh_std, criteria = "KICb2")

Start: KICb2 = 1709.42
Mean ~ cash + self_empl + unempl_ben

            df  KICb2
- unempl_ben 1 1708.3
<none>         1709.4
- self_empl  1 1763.0
- cash       1 1808.6

Step: KICb2 = 1708.33
Mean ~ cash + self_empl

            df  KICb2
<none>         1708.3
- self_empl  1 1765.3
- cash       1 1816.1


Call:
fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
    combined_data = combined_data,
    domains = "Domain", method = "ml", B = c(0, 50))

Coefficients:
             coefficients  std.error t.value   p.value
(Intercept)   3070.51231  635.94290  4.8283 1.377e-06 ***
cash             1.05939    0.07049 15.0288 < 2.2e-16 ***
self_empl        1.74564    0.22017  7.9284 2.219e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

KICb2 is minimised when the variable `unempl_ben` is removed. Therefore, the formula `Mean` $\sim$ `cash + self_empl` is used in the following.

Estimate EBLUPs and MSEs
The standard FH model is built. In addition to the `fixed` part, required arguments are `vardir` and `combined_data`. We specify the `domains` (if the `domains` argument is set to `NULL`, the domains are numbered consecutively) and activate the estimation of the MSE and of the information criteria by Marhuenda et al. (2014).

```
> fh_std <- fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean", combined_data =
+   combined_data, domains = "Domain", method = "ml", MSE = TRUE, B = c(0,50))
```

Assess the estimated model

In many publications using FH models, model diagnostics are discussed only briefly, if at all. One reason for this might be the lack of an existing implementation of desired diagnostic measures in R or other statistical software. The `summary` method of **emdi** provides additional information about the data and model components, in particular the chosen estimation methods, the number of domains, the log-likelihood, the information criteria by Marhuenda et al. (2014), the adjusted R^2 of a standard linear model and the adjusted R^2 especially for FH models proposed by Lahiri and Suntornchost (2015). Additionally, measures to validate model assumptions about the standardized realized residuals and the random effects are provided: skewness and kurtosis (`skewness` and `kurtosis` of package **moments**, Komsta and Novomestky, 2015) of the standardized realized residuals and the random effects and the test statistics with corresponding p value of the Shapiro-Wilks-test for normality of both error terms. As the introduced area-level models do not assume a homoscedastic sampling distribution, the realized residuals ($\hat{e}_i$) are standardized by $\hat{e}_i^{\text{std}} = \hat{e}_i / \sigma_{e_i}$ for the `summary` and `plot` methods. The `summary` output differs slightly for the different implemented area-level models. For example, log-likehoods and thus information criteria are not available in theory for the robust and the ME model.

```
> summary(fh_std)

Call:
fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
    combined_data = combined_data,
    domains = "Domain", method = "ml", MSE = TRUE, B = c(0, 50))

Out-of-sample domains: 0
In-sample domains: 94

Variance and MSE estimation:
Variance estimation method: ml
Estimated variance component(s): 1371195
MSE method: datta-lahiri

Coefficients:
           coefficients std.error  t.value   p.value
(Intercept)  3070.51231 635.94290   4.8283 1.377e-06 ***
cash            1.05939   0.07049  15.0288 < 2.2e-16 ***
self_empl       1.74564   0.22017   7.9284 2.219e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Explanatory measures:
    loglike      AIC    AICc    AICb1    AICb2      BIC      KIC
1 -847.8303 1703.661 1703.91 1715.758 1703.461 1713.834 1707.661
      KICc    KICb1    KICb2     AdjR2     FH_R2
1 1708.783 1720.632 1708.335 0.9212817 0.9482498

Residual diagnostics:
                        Skewness Kurtosis Shapiro_W Shapiro_p
Standardized_Residuals  0.3004662 3.971216 0.9840810 0.3119346
Random_effects         -0.4113238 3.086048 0.9839858 0.3072834

Transformation: No transformation
```

The output of the example shows that all domains have survey information and the variance of σ_u^2 amounts to 1371195. Further, all of the included auxiliary variables are quite significant and their explanatory power is large with an adjusted R^2 (for FH models) of around 0.95. The results of the Shapiro-Wilk-test indicate that normality is not rejected for both errors. Graphical residual diagnostics are possible by the `plot` method.

```
> plot(fh_std)
```

Figure 3 shows normal quantile-quantile (Q-Q) plots of the standardized realized residuals and random effects (Figure 3a) as well as plots of the kernel densities of the distribution of both error terms and, for comparison, a standard normal distribution (Figure 3b and 3c). Like in **emdi** version 1.1.7, the user is free to modify the interface of the plots. The `label` and `color` arguments are easy to edit. Additionally, the overall appearance of the plots are changeable by the `gg_theme` argument, as the plots are built with the **ggplot2** package (Wickham, 2016). We refer to the package documentation

for a detailed description of how to customize the `plot` arguments. Figure 3 supports the results of the normality tests provided in the `summary` output, the distribution of the standardized random effects may be slightly skewed (Figure 3c). If one is not be satisfied with the results, applying a log-transformation could improve the distribution of the error terms.

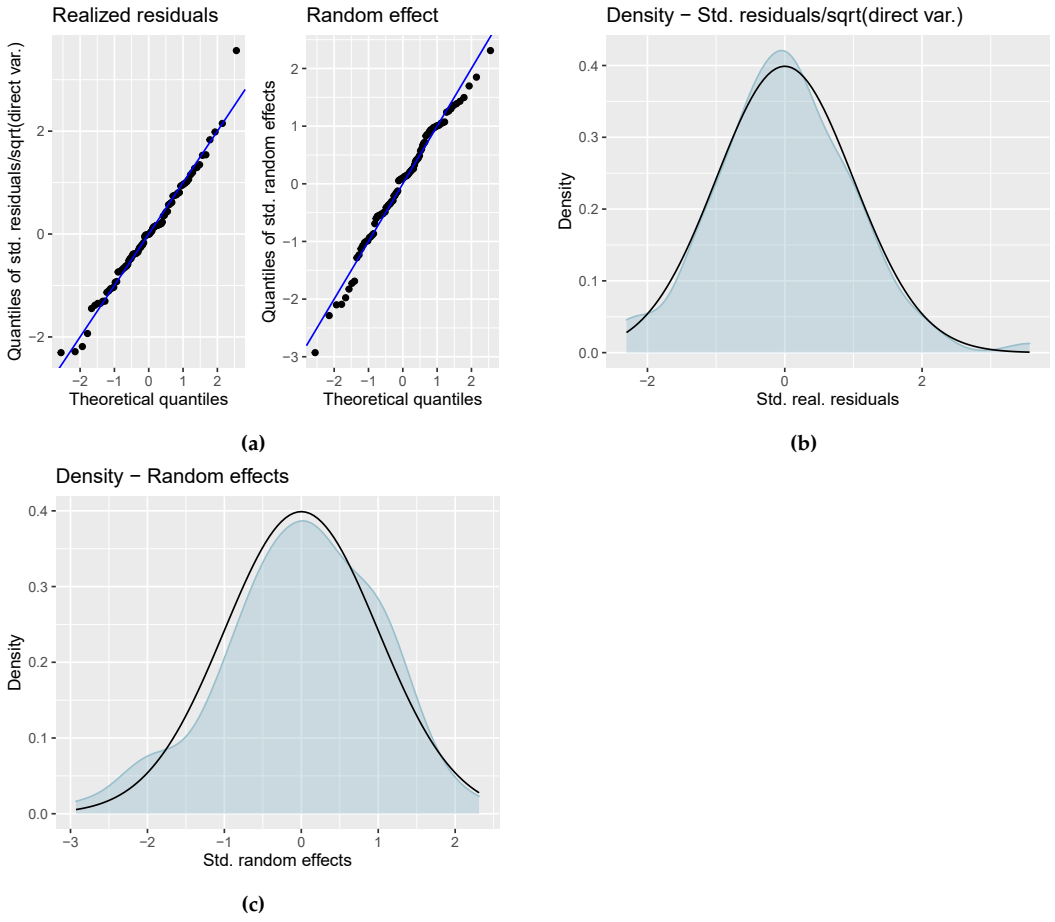

Figure 3: Output of `plot(fh_std)`: (a) normal quantile-quantile (Q-Q) plots of the standardized realized residuals and random effects, (b) and (c): kernel densities of the distribution of the standardized realized residuals and random effects (blue) in comparison to a standard normal distribution (black).

Compare results with direct estimates

The FH results should be consistent with the direct estimates for domains with a small direct MSE and/or large sample sizes. Further, the precision of the direct estimates should be improved by using auxiliary information. The comparison of the direct and model-based (FH) estimates can be done graphically by the generic function `compare_plot`. For the `fh` method the required input argument is an object of class `"fh"`. When the default settings of the command are used, the output consists of two plots: a scatter plot proposed by Brown et al. (2001) and a line plot. Besides the direct and FH estimates, the plot contains the fitted regression and the identity line. These two lines should not differ too much. Preferably, the model-based (FH) estimates should track the direct estimates within the line plot especially for domains with a large sample size or small MSE of the direct estimator. The points are ordered by decreasing MSE of the direct estimates. In addition, the input arguments `MSE` and `CV` can be set to `TRUE` leading to two extra plots, respectively. The MSE/CV estimates of the direct and model-based (FH) estimates are compared first via boxplots and second via ordered scatter plots (ordered by increasing CV of the direct estimates). Like for the `plot` command, a variety of customization options are offered, e.g., the label options (`label`), the format of the points (`shape`) and the style of the line (`line_type`).

```
> compare_plot(fh_std, CV = TRUE, label = "no_title")
```

Except one high value, the fitted regression and identity line of the scatter plot (Figure 4a) are relatively close. Note that the high value corresponds to the domain Eisenstadt (Stadt) with a very small sample size of 10 and the highest MSE of the direct estimates, so the direct estimator is very uncertain. Also the direct estimates are well tracked by the model-based (FH) estimates within the line plot (Figure 4b).

The boxplot (Figure 4c) and the ordered scatter plot (Figure 4d) show that the precision of the direct estimates could be improved by the usage of the FH model in terms of CVs. Additionally, all of the CV values are less than 20% which is a common rule of the UK Office for National Statistics in order to determine whether estimation results should be published (Miltiadou, 2020).

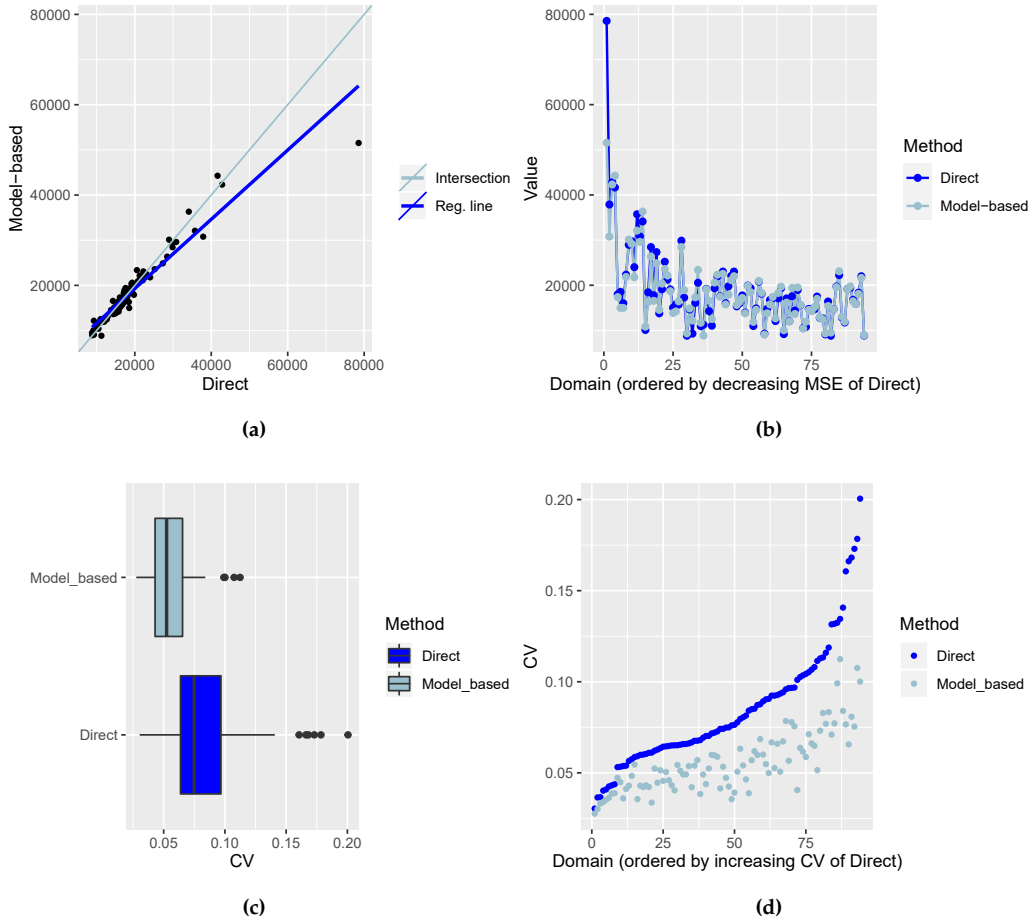

Figure 4: Output of `compare_plot(fh_std)`: (a) and (b) scatter and line plots of direct and model-based point estimates, (c) and (d) boxplot and scatter plots of the CV estimates of the direct and model-based (FH) estimates.

Further on, the function `compare` enables the user to compute a goodness of fit diagnostic (Brown et al., 2001) and a correlation coefficient of the direct estimates and the estimates of the regression-synthetic part of the FH model (Chandra et al., 2015). Following Brown et al. (2001), the difference between the model-based estimates and the direct estimates should not be significant (null hypothesis). The Wald test statistic is specified as

$$W\left(\hat{\theta}_i^{\mathrm{FH}}\right) = \sum_{i=1}^{D} \frac{\left(\hat{\theta}_i^{\mathrm{Dir}} - \hat{\theta}_i^{\mathrm{FH}}\right)^2}{\widehat{\mathrm{var}}\left(\hat{\theta}_i^{\mathrm{Dir}}\right) + \widehat{\mathrm{MSE}}\left(\hat{\theta}_i^{\mathrm{FH}}\right)}$$

and is approximately χ^2-distributed with D degrees of freedom. When working with out-of-sample domains, those are not taken into account, because the direct estimates and their variances are missing. The input argument of function `compare` is an `"fh"` object.

```
> compare(fh_std)

Brown test

Null hypothesis: EBLUP estimates do not differ significantly from the
      direct estimates

  W.value Df p.value
 46.97181 94 0.9999874
```

```
Correlation between synthetic part and direct estimator: 0.94
```

The results of the goodness of fit statistic and the correlation coefficient confirm what the scatter and the line plot already indicated. In the example the null hypothesis is not rejected and the correlation coefficient indicates a strong positive correlation (0.94) between the direct and model-based (FH) estimates.

Benchmarking for consistent estimates
The idea of benchmarking is that the aggregated FH estimates should sum up to estimates of a higher regional level (τ):

$$\sum_{i=1}^{D} \xi_i \hat{\theta}_i^{\text{FH, bench}} = \tau,$$

where ξ_i stands for the share of the population size of each area in the total population size (N_i/N). In our example, the EBLUP estimates could get aggregated on a national level and then compared to or benchmarked with the Austrian mean equivalized income. The **emdi** package contains a benchmark function that allows the user to select three different options suggested by Datta et al. (2011). A general estimator of the three options can be written as follows:

$$\hat{\theta}_i^{\text{FH, bench}} = \hat{\theta}_i^{\text{FH}} + \left(\sum_{i=1}^{D} \frac{\xi_i^2}{\phi_i} \right)^{-1} \left(\tau - \sum_{i=1}^{D} \xi_i \hat{\theta}_i^{\text{FH}} \right) \frac{\xi_i}{\phi_i}.$$

Depending on the weight ϕ_i, the formula leads to different benchmarking options. If ϕ_i equals ξ_i, all FH estimates are adjusted by the same value (`raking`). A ratio adjustment (`ratio`) is being conducted if $\phi_i = \xi_i / \hat{\theta}_i^{\text{FH}}$. For the last option (`MSE_adj`), $\phi_i = \xi_i / \widehat{\text{MSE}}\left(\hat{\theta}_i^{\text{FH}}\right)$. While the first option is a relatively naive approach, the latter two conduct larger adjustments for the areas with larger FH and MSE estimates, respectively. Thus, for the benchmark function the following arguments have to be specified: an object of class `"fh"`, a `benchmark` value, a vector containing the ξ_is (`share`) and the `type` of benchmarking. The output is a data frame with an extra column `FH_Bench` for the benchmarked EBLUP values. If the optional argument `overwrite` is set to `TRUE`, the benchmarked results are added to the `"fh"` object and the MSE estimates of the non benchmarked FH estimates are set to `NULL`. For the used example, the benchmark value is calculated by taking the mean of the variable `eqIncome` of the 'eusilcA_smp' data frame. The ξ_is can be found in 'eusilcA_popAgg' as `ratio_n`.

```
> fh_bench <- benchmark(fh_std, benchmark = 20140.09, share = eusilcA_popAgg$ratio_n,
+   type = "ratio")
> head(fh_bench)
```

```
              Domain   Direct       FH FH_Bench Out
1          Amstetten 14768.57 14242.04 14480.61   0
2              Baden 21995.72 21616.40 21978.49   0
3            Bludenz 12069.59 12680.38 12892.79   0
4     Braunau am Inn 10770.53 11925.82 12125.59   0
5            Bregenz 35731.20 32101.69 32639.43   0
6 Bruck-Mürzzuschlag 23027.37 22523.50 22900.79   0
```

It is recognizable that for the first six Austrian districts the original estimates are slightly modified by the benchmarking.

Extract and visualize the results
To gain an overview of the point, MSE and CV results of the direct estimates compared to the model-based (FH) results the generic function `estimators` (Kreutzmann et al., 2019) can be used, but differences among areas or hotspots of special interest are usually easier to detect on maps. With function `map_plot`, the **emdi** package offers a user-friendly way to produce maps since creating maps can often become a time consuming task. The input arguments mainly consist of an object of class `"emdi"` and a spatial polygon of a shape file. The only issue that might come up is if domain identifiers in the data do not match to the respective identifiers of the shape file. In those cases, the input argument `map_tab` is required, which is a data frame that contains the matching of the domain indentifiers of the population and the shape file data sets. For detailed instructions, we refer to Kreutzmann et al. (2019) and to the help page of function `map_plot`.

For producing maps of the 94 Austrian districts, the Austrian shape file has to be loaded. In addition to the `"emdi"` object, the `"SpatialPolygonsDataFrame"` object (`map_obj`) and a domain indicator (`map_dom_id`) have to be specified. The `map_tab` argument is not necessary since the identifiers match in our example. To allow for an easier comparison of the results, we adjust the scales of the maps

using the scale_points argument.

```
> load_shapeaustria()
> map_plot(object = fh_std, MSE = TRUE, map_obj = shape_austria_dis,
+   map_dom_id = "PB", scale_points = list(Direct = list(
+   ind = c(8000, 60000), MSE = c(200000, 10000000)), FH = list(
+   ind = c(8000, 60000), MSE = c(200000, 10000000))))
```

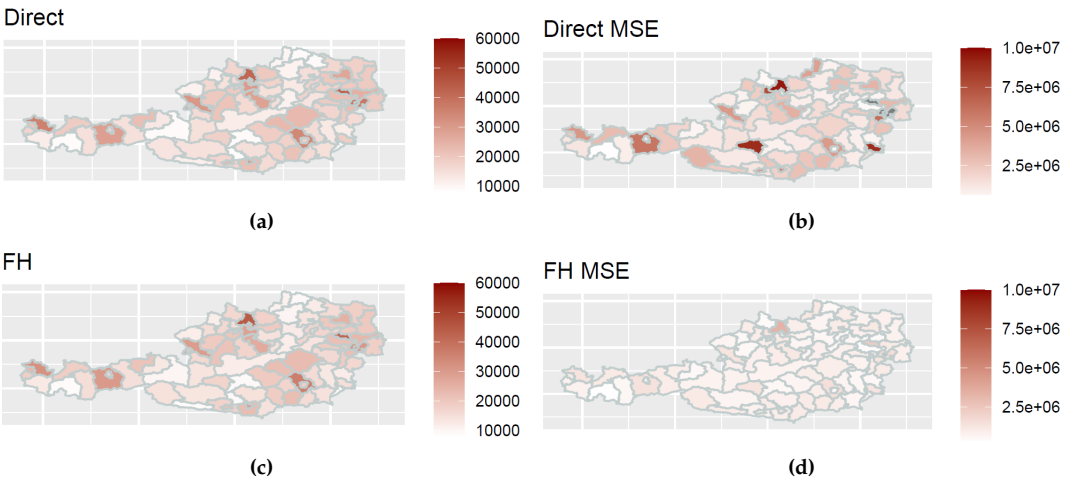

Figure 5: Output of map_plot: Maps of the direct and FH estimates ((a) and (c)) with corresponding MSE estimates ((b) and (d)).

Figures 5a and 5c show the distribution of the estimated (direct vs. model-based) equivalized income across Austria. It is striking that white and light red tones dominate the map, indicating relatively low mean incomes of the districts. But in contrast, districts Eisenstadt (Stadt), Urfahr-Umgebung and Mödling stand out having the largest incomes. Urfahr-Umgebung is also eye-catching when having a look at the MSE estimates (Figures 5b and 5d). The MSE of the direct and the FH estimates are quite high. Probably a single wealthy household raised the mean income and also the variance. Figure 5b contains some districts with MSEs larger than the customized scaling (gray areas). Without the scaling it would have been hard to identify any differences in Figure 5d.

Export the results
Some users might have an interest to store the results separately or to use them for presentations. Excel and OpenDocument Spreadsheets provide many opportunities for that. In contrast to some existing R packages, the **emdi** functions write.excel/write.ods do not only export the estimation results, but also the output of summary. Usage of the functions is comprehensively described in Kreutzmann et al. (2019).

Estimation of the extended area-level models

FH model with transformation
If the indicator of interest needs a transformation, either log or arcsin, in addition to the function used in the previous subsection, the arguments transformation and backtransformation must be specified. If, for example, the share of households per area that earn more than the national median income (MTMED) is the indicator of interest, an arcsin transformation can be used. The bias-corrected back-transformation bc is chosen in the example. Two more arguments are needed when using an arcsin transformation: the name of the variable describing the effective sample sizes (eff_smpsize) which needs to be contained in the combined_data frame. Because of having chosen the bias-corrected back-transformation, the only possible mse_type is boot, if the MSE estimation is activated.

```
> fh_arcsin <- fh(fixed = MTMED ~ cash + age_ben + rent + house_allow,
+   vardir = "Var_MTMED", combined_data = combined_data, domains = "Domain",
+   transformation = "arcsin", backtransformation = "bc", eff_smpsize = "n",
+   MSE = TRUE, mse_type = "boot")
```

Spatial FH model
If the spatial correlation tests indicated a spatial correlation of the domains, a spatial FH model for incorporating the spatial structure in the model could be used. For that, the correlation has to be

set to spatial and the example proximity matrix has to be given to the model within the corMatrix argument. The possible variance estimation methods are ml and reml.

```
> fh_spatial <- fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
+   combined_data = combined_data, domains = "Domain", correlation = "spatial",
+   corMatrix = eusilcA_prox, MSE = TRUE)
```

Robust FH model
If extreme values could influence the estimation, the application of a robust model might be appropriate. Within the robust framework, package **emdi** allows the user to choose between a standard and a spatial model (defaults to correlation = "no"). The estimation method must be reblup or reblupbc which includes a bias correction that can be modified by the argument mult_constant. Further, the tuning constant k defaults to 1.345 as proposed by Sinha and Rao (2009) and Warnholz (2016) and can be changed if desired. The functions of the package **saeRobust** are utilized for the robust extensions. An exemplary call with pseudolinear MSE estimation looks like this:

```
> fh_robust <- fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
+   combined_data = combined_data, domains = "Domain", method = "reblup",
+   MSE = TRUE, mse_type = "pseudo")
```

Measurement error model
If other data sources than register data, e.g., data from larger surveys or big data sources are used as auxiliary information, the ME model should be applied. For the estimation of the ME model, the model fitting method must be set to me and the only possible MSE estimation method is jackknife. The most complex input argument consists of the creation of the MSE array Ci. The variability of the auxiliary variables that is taken into account by the ME model is expressed by the variance-covariance matrices per domain (Ci). For example, for three covariates a, b and c the array should look like

$$\boldsymbol{C}_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathrm{var}_i(a) & \mathrm{cov}_i(a,b) & \mathrm{cov}_i(a,c) \\ 0 & \mathrm{cov}_i(a,b) & \mathrm{var}_i(b) & \mathrm{cov}_i(b,c) \\ 0 & \mathrm{cov}_i(a,c) & \mathrm{cov}_i(b,c) & \mathrm{var}_i(c) \end{pmatrix}, \; i = 1, ..., D.$$

The first row and column contain zeros, because the intercept is considered. The variances and covariances can be computed by standard approaches like, for example, the Horvitz-Thompson estimator.

For the Austrian EUSILC data example, the equalized income can also be explained by a variable of the sample data set. The code below demonstrates how the MSE array Ci is created for one covariate (variable Cash and its variance Var_Cash) and how the final ME model is built.

```
> P <- 1
> M <- 94
> Ci_array <- array(data = 0, dim = c(P + 1, P + 1, M))
> Ci_array[2,2, ] <- eusilcA_smpAgg$Var_Cash

> fh_yl <- fh(fixed = Mean ~ Cash, vardir = "Var_Mean",
+   combined_data = eusilcA_smpAgg, domains = "Domain", method = "me",
+   Ci = Ci_array, MSE = TRUE, mse_type = "jackknife")
```

5 Conclusion and outlook

In this paper, we have presented how the **emdi** package version 1.1.7 has been extended with various area-level models. Along with the well-known FH model, adjusted variance estimation methods and transformation options are offered to the user. In addition, spatial, robust, and ME model extensions of the standard model allow the user to address various issues that arise in practical data applications. All of these methods can be estimated conveniently by using a single function that provides EBLUP and the respective MSE estimates to measure their precision. Especially in the section Functionality and case studies, it is clear that the package does not only contain tools for estimation of the different SAE models. Instead, it additionally provides user-friendly tools to enable a whole data analysis procedure: 1. starting with model building and estimation, moving on to 2. model assessment and diagnostics, 3. presentation of the results, and finishing with 4. exporting the results to Excel or OpenDocument Spreadsheet.

For future package versions, it is planned to expand the options in the field of area-level models. In some practical applications, the incorporation of random effects is redundant. Therefore, an area-level estimator that considers a preliminary testing for the random effects following Molina et al. (2015)

will be included. Since version 2.0.0 **emdi** accounts for spatial structures of the random effects. Future developments may also account for out-of-sample EBLUP and MSE estimation for the spatial model proposed by Saei and Chambers (2005) and for temporal and spatio-temporal extensions (Rao and Yu, 1994; Marhuenda et al., 2013). For the existing ME model, a bootstrap MSE estimation option may be added to the package since the Jackknife MSE estimator may produce negative MSE estimates (Marchetti et al., 2015). Furthermore, cross-validation options additional to the model assessment via information criteria and the R^2 will be investigated.

6 Acknowledgments

The work of Kreutzmann and Schmid has been supported by the German Research Foundation within the project QUESSAMI (281573942) and by the MIUR-DAAD Joint Mobility Program (57265468). The numerical results are not official estimates and are only produced for illustrating the methods. The authors are indebted to the Editor-in-Chief, Associate Editor and the referees for comments that significantly improved the article.

Bibliography

A. Alfons and M. Templ. Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, 54(15):1–25, 2013. URL https://doi.org/10.18637/jss.v054.i15. [p3]

A. Alfons, M. Templ, and P. Filzmoser. An object-oriented framework for statistical simulation: The R package simFrame. *Journal of Statistical Software*, 37(3):1–36, 2010. URL https://doi.org/10.18637/jss.v037.i03. [p7]

G. E. Battese, R. M. Harter, and W. A. Fuller. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36, 1988. URL https://doi.org/10.1080/01621459.1988.10478561. [p2]

G. Bertarelli, F. Schirripa Spagnolo, N. Salvati, and M. Pratesi. Small area estimation of agricultural data. In *Spatial Econometric Methods in Agricultural Economics Using R*. CRC book, 2021. [p5]

R. S. Bivand and D. W. S. Wong. Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3):716–748, 2018. URL https://doi.org/10.1007/s11749-018-0599-x. [p10]

H. J. Boonstra. *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*, 2021. URL https://CRAN.R-project.org/package=mcmcsae. R package version 0.7.0. [p1]

H. J. Boonstra. *hbsae: Hierarchical Bayesian Small Area Estimation*, 2022. URL https://CRAN.R-project.org/package=hbsae. R package version 1.2. [p1]

J. Breidenbach. *JoSAE: Unit-Level and Area-Level Small Area Estimation*, 2018. URL https://CRAN.R-project.org/package=JoSAE. R package version 0.3.0. [p1]

G. Brown, R. Chambers, P. Heady, and D. Heasman. Evaluation of small area estimation methods - an application to unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium*, 2001. [p13, 14]

Bundesamt für Eich- und Vermessungswesen. Verwaltungsgrenzen (VGD) - 1:250.000 Bezirksgrenzen, Daten vom 01.04.2017 von SynerGIS, 2017. URL http://data-synergis.opendata.arcgis.com/datasets/bb4acc011100469185d2e59fa4cae5fc_0. [accessed: 07.02.2018]. [p8]

C. Casas-Cordero, J. Encina, and P. Lahiri. Poverty mapping for the chilean comunas. In *Analysis of Poverty by Small Area Estimation*, pages 379–403. John Wiley & Sons, 2016. URL https://doi.org/10.1002/9781118814963.ch20. [p5, 11]

J. Chambers and T. Hastie, editors. *Statistical Models in S*. Chapman & Hall, London, 1992. [p8]

R. Chambers, H. Chandra, N. Salvati, and N. Tzavidis. Outlier robust small area estimation. *Journal of the Royal Statistical Society B*, 76(1):47–69, 2014. URL https://doi.org/10.1111/rssb.12019. [p5]

H. Chandra, N. Salvati, and R. Chambers. A spatially nonstationary Fay-Herriot model for small area estimation. *Journal of the Survey Statistics and Methodology*, 3(2):109–135, 2015. URL https://doi.org/10.1093/jssam/smu026. [p14]

H. Chandra, N. Salvati, R. Chambers, and S. Guha. *NSAE: Nonstationary Small Area Estimation*, 2022. URL https://CRAN.R-project.org/package=NSAE. R package version 0.4.0. [p2]

S. Chen and P. Lahiri. A weighted jackknife mspe estimator in small-area estimation. In *Proceeding of the Section on Survey Research Methods*, pages 473–477, 2002. American Statistical Association. [p6]

A. Cliff and J. Ord. *Spatial Processes: Models and Applications*. Pion, London, 1981. [p5]

G. Datta, M. Ghosh, R. Steorts, and J. Maples. Bayesian benchmarking with applications to small area estimation. *TEST*, 20(3):574–588, 2011. URL https://doi.org/10.1007/s11749-010-0218-y. [p15]

G. S. Datta and P. Lahiri. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10(2):613–627, 2000. URL http://www.jstor.com/stable/24306735. [p6]

G. S. Datta, R. E. Fay, and M. Ghosh. Hierarchical and empirical bayes multivariate analysis in small area estimation. In *Proceedings of Bureau of the Census 1991 Annual Research Conference*, pages 63–79. US Bureau of the Census, 1991. [p1]

S. De Nicolò and A. Gardini. *tipsae: Tools for Handling Indices and Proportions in Small Area Estimation*, 2022. URL https://CRAN.R-project.org/package=tipsae. R package version 0.0.6. [p1]

C. S. Developer. *BayesSAE: Bayesian Analysis of Small Area Estimation*, 2018. URL https://CRAN.R-project.org/package=BayesSAE. R package version 1.0-2. [p1]

A. Fasulo. *SAEval: Small Area Estimation Evaluation*, 2022. URL https://CRAN.R-project.org/package=SAEval. R package version 0.1.5. [p2]

R. E. Fay and R. A. Herriot. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277, 1979. URL https://doi.org/10.1080/01621459.1979.10482505. [p1, 3, 5]

S. Hadam, N. Würz, and A.-K. Kreutzmann. Estimating regional unemployment with mobile network data for functional urban areas in Germany. *Refubium - Freie Universität Berlin Repository*, pages 1–28, 2020. URL https://doi.org/10.17169/refubium-26791. [p5, 6]

A. Hagenaars, K. de Vos, and M. Zaidi. *Poverty Statistics in the Late 1980s: Research Based on Mirco-data*. Office for the Official Publications of the European Communities, 1994. [p7]

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. URL https://doi.org/10.1080/01621459.1952.10483446. [p3]

J. Jiang and J. S. Rao. Robust small area estimation: An overview. *Annual Review of Statistics and Its Application*, 7:337–360, 2020. URL https://doi.org/10.1146/annurev-statistics-031219-041212. [p1]

J. Jiang, P. Lahiri, S.-M. Wan, and C.-H. Wu. Jackknifing in the Fay-Herriot model with an example. In *Proceedings of the Seminar on Funding Opportunity in Survey Research Council of Professional Associations on Federal Statistics*, pages 75–97, 2001. [p5, 6]

J. Jiang, P. Lahiri, and S.-M. Wan. A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, 30(6):1782–1810, 2002. URL https://doi.org/10.1214/aos/1043351257. [p6]

L. Komsta and F. Novomestky. *moments: Moments, cumulants, skewness, kurtosis and related tests*, 2015. URL https://CRAN.R-project.org/package=moments. R package version 0.14. [p12]

A.-K. Kreutzmann, S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7):1–33, 2019. URL https://doi.org/10.18637/jss.v091.i07. [p6, 15, 16]

A.-K. Kreutzmann, P. Marek, M. Runge, N. Salvati, and T. Schmid. The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany. *Journal of Applied Statistics*, 49(13):3278–3299, 2022. URL https://doi.org/10.1080/02664763.2021.1941805. [p4]

P. Lahiri and J. Suntornchost. Variable selection for linear mixed models with applications in small area estimation. *The Indian Journal of Statistics*, 77-B(2):312–320, 2015. URL https://www.jstor.org/stable/43694416. [p12]

H. Li and P. Lahiri. An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analyis*, 101(4):882–902, 2010. URL https://doi.org/10.1016/j.jmva.2009.10.009. [p4, 6]

E. Lopez-Vizcaino, M. Lombardia, and D. Morales. *mme: Multinomial Mixed Effects Models*, 2019. URL https://CRAN.R-project.org/package=mme. R package version 0.1-6. [p2]

S. Marchetti, C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281, 2015. URL https://doi.org/10.1515/jos-2015-0017. [p18]

Y. Marhuenda, I. Molina, and D. Morales. Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58:308–325, 2013. URL https://doi.org/10.1016/j.csda.2012.09.002. [p18]

Y. Marhuenda, D. Morales, and M. del Camen Pardo. Information criteria for Fay-Herriot model selection. *Computational Statistics and Data Analysis*, 70:268–280, 2014. URL https://doi.org/10.1016/j.csda.2013.09.016. [p8, 11, 12, 24]

M. Miltiadou. Measuring and reporting reliability of labour force survey and annual population survey estimates force survey and annual population survey estimates, 2020. URL https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/measuringandreportingreliabilityoflabourforcesurveyandannualpopulationsurveyestimates. UK Office for National Statistics, [accessed: 05.06.2020]. [p14]

I. Molina and Y. Marhuenda. sae: An R package for small area estimation. *The R Journal*, 7(1):81–98, 2015. URL https://doi.org/10.32614/rj-2015-007. [p2]

I. Molina and J. Rao. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38 (3):369–385, 2010. URL https://doi.org/10.1002/cjs.10051. [p1]

I. Molina, N. Salvati, and M. Pratesi. Bootstrap for estimating the mse of the spatial eblup. *Computational Statistics*, 24:441–458, 2009. URL https://doi.org/10.1007/s00180-008-0138-4. [p6]

I. Molina, J. Rao, and G. Datta. Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, 41(1):1–19, 2015. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14161-eng.htm. [p17]

M. Mubarak and A. Ubaidillah. *saeME: Small Area Estimation with Measurement Error*, 2022. URL https://CRAN.R-project.org/package=saeME. R package version 1.3. [p2]

A. Neves, D. Silva, and S. Correa. Small domain estimation for the Brazilian service sector survey. *ESTADÍSTICA*, 65(185):13–37, 2013. [p4]

E. J. Pebesma and R. S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005. URL https://CRAN.R-project.org/doc/Rnews/. [p10]

N. Permatasari and A. Ubaidillah. *msae: Multivariate Fay Herriot Models for Small Area Estimation*, 2022. URL https://CRAN.R-project.org/package=msae. R package version 0.1.5. [p1]

A. Petrucci and N. Salvati. Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics*, 11(2):169–182, 2006. URL https://doi.org/10.1198/108571106X110531. [p5]

D. Pfeffermann. New important developments in small area estimation. *Statistical Science*, 28(1):40–68, 2013. URL https://doi.org/10.1214/12-STS395. [p1]

N. Prasad and J. Rao. The estimation of the mean squared error of small-area estimation. *Journal of the American Statistical Association*, 85(409):163–171, 1990. URL https://doi.org/10.1080/01621459.1990.10475320. [p6]

M. Pratesi, editor. *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons, 2016. URL https://doi.org/10.1002/9781118814963. [p1]

M. Pratesi and N. Salvati. Small area estimation: The eblup estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17(1):113–141, 2008. URL https://doi.org/10.1007/s10260-007-0061-9. [p5]

J. N. K. Rao and I. Molina. *Small Area Estimation*. John Wiley & Sons, 2015. URL https://doi.org/10.1002/9781118735855. [p1, 3, 6]

J. N. K. Rao and M. Yu. Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, 22(4):511–528, 1994. URL https://doi.org/10.2307/3315407. [p18]

L.-P. Rivest and N. Vandal. Mean squared error estimation for small areas when the small area variances are estimated. In *Proceedings of International Conference of Recent Advanced Survey Sampling*, pages 197–206, 2003. [p3]

A. Saei and R. Chambers. Out of sample estimation for small areas using area level data. *Southampton Statistical Sciences Research Institute Methodology Working Paper*, M05/11, 2005. URL http://eprints.soton.ac.uk/id/eprint/14327. Southampton Statistical Sciences Research Institute, UK. [p18]

T. Schmid, N. Tzavidis, R. Münnich, and R. Chambers. Outlier robust small area estimation under spatial correlation. *Scandinavian Journal of Statistics*, 43(3):806–826, 2016. URL https://doi.org/10.1111/sjos.12205. [p5]

T. Schmid, F. Bruckschen, N. Salvati, and T. Zbiranski. Constructing sociodemographic indicators for national statistical institutes using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society A*, 180(4):1163–1190, 2017. URL https://doi.org/10.1111/rssa.12305. [p5, 11]

B. B. Singh, K. Shukla, and D. Kundu. Spatio-temporal models in small area estimation. *Survey Methodology*, 31(2):183–195, 2005. URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20050029053. [p6]

S. Sinha and J. Rao. Robust small area estimation. *The Canadian Journal of Statistics*, 37(3):381–399, 2009. URL https://doi.org/10.1002/cjs.10029. [p5, 17]

E. Slud and T. Maiti. Mean-squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society B*, 68(2):239–257, 2006. URL https://doi.org/10.1111/j.1467-9868.2006.00542.x. [p4, 6]

S. Sugawasa and T. Kubokawa. Transforming response values in small area prediction. *Computational Statistics and Data Analysis*, 114:47–60, 2017. URL https://doi.org/10.1016/j.csda.2017.03.017. [p4, 5]

N. Tzavidis, R. Chambers, N. Salvati, and H. Chandra. Small area estimation in practice an application to agricultural business survey data. *Journal of the Indian Society of Agricultural Statistics*, 66(1):213–228, 2012. URL https://ro.uow.edu.au/eispapers/758/. [p1]

N. Tzavidis, L.-C. Zhang, A. Luna Hernandez, T. Schmid, and N. Rojas-Perilla. From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society A*, 181(4):927–979, 2018. URL https://doi.org/10.1111/rssa.12364. [p1]

K. Ushey, J. McPherson, J. Cheng, A. Atkins, and J. Allaire. *packrat: A Dependency Management System for Projects and their R Package Dependencies*, 2022. URL https://CRAN.R-project.org/package=packrat. R package version 0.8.0. [p25]

J. Wang and W. A. Fuller. The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98:716–723, 2003. URL https://doi.org/10.1198/016214503000000620. [p3]

S. Warnholz. *Small Area Estimation Using Robust Extensions to Area Level Models*. PhD thesis, Freie Universität Berlin, 2016. URL https://doi.org/10.17169/refubium-13904. [p5, 6, 17]

S. Warnholz. *saeRobust: Robust Small Area Estimation*, 2022. URL https://CRAN.R-project.org/package=saeRobust. R package version 0.3.0. [p1]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL https://ggplot2.tidyverse.org. [p12]

P. Xiao, X. Liu, Y. Liu, and S. Liu. *saeMSPE: Compute MSPE Estimates for the Fay Herriot Model and Nested Error Regression Model*, 2022. URL https://CRAN.R-project.org/package=saeMSPE. R package version 1.0. [p2]

L. M. R. Ybarra and S. L. Lohr. Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931, 2008. URL https://doi.org/10.1093/biomet/asn048. [p6]

M. Yoshimori and P. Lahiri. A new adjusted maximum likelihood method for the Fay-Herriot small area model. *Journal of Multivariate Analysis*, 124:281–294, 2014. URL https://doi.org/10.1016/j.jmva.2013.10.012. [p4]

Y. You and B. Chapman. Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32(1):97–103, 2006. URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263. [p3]

X. Zhang, J. Holt, S. Yun, H. Lu, K. Greenlund, and J. Croft. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*, 182(2):127–137, 2015. URL https://doi.org/10.1093/aje/kwv002. [p1]

7 Appendix A: Area-level model options and corresponding input arguments

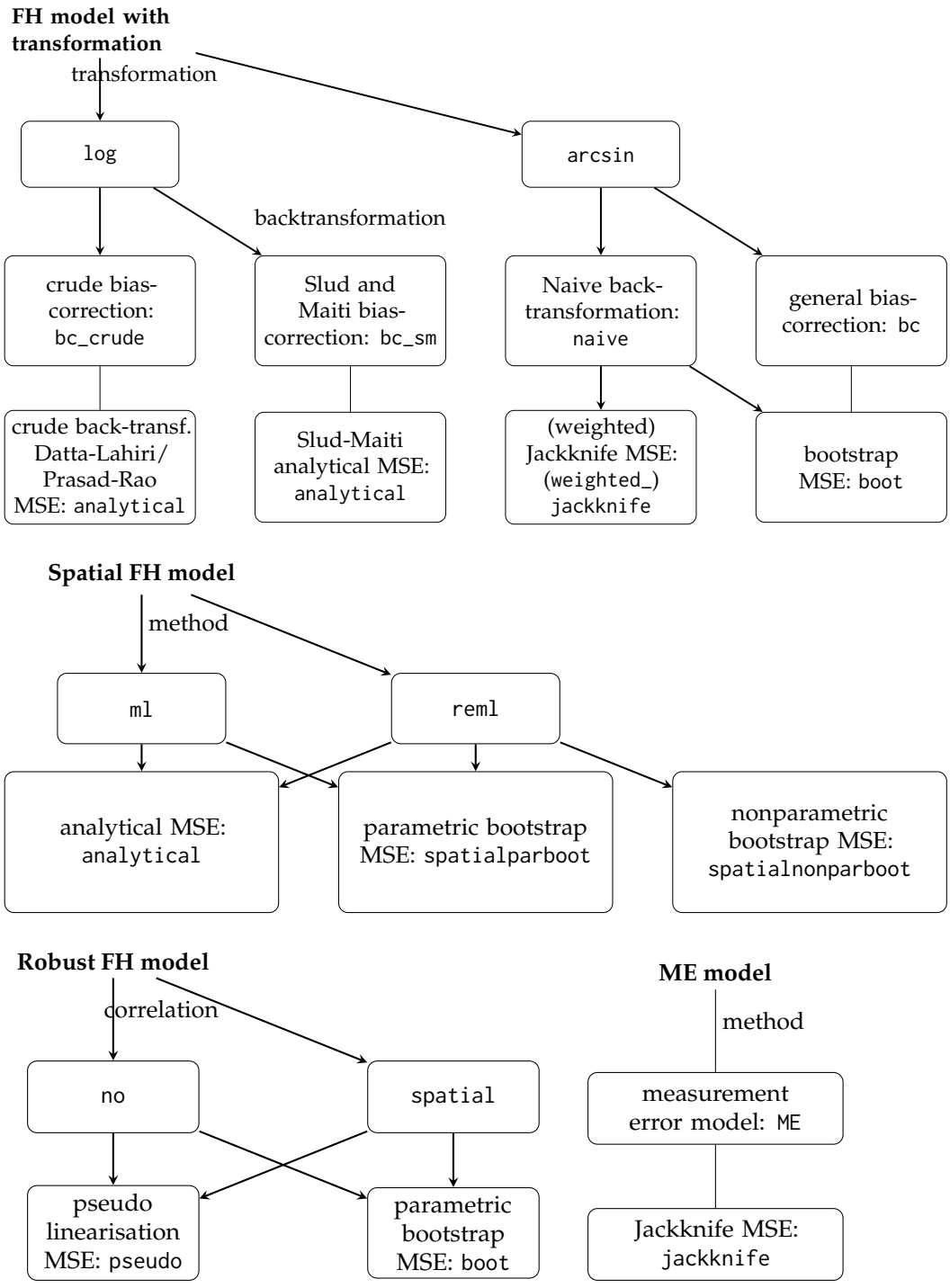

Figure 6: Overview of extended area-level models and combinations of estimation methods.

Argument	FH model				
	Standard	Transformed	Spatial	Robust	ME
`fixed`	√	√	√	√	√
`vardir`	√	√	√	√	√
`combined_data`	√	√	√	√	√
`domains`	(√)	(√)	(√)	(√)	(√)
`method`	√	√	√	√	√
`interval`	(√)	(√)			
`k`				√	
`mult_constant`				√	
`transformation`	√	√	√	√	√
`backtransformation`		√			
`eff_smpsize` (only if `transformation = "arcsin"`)		√			
`correlation`	√	√	√	√	√
`corMatrix` (only if `correlation = "spatial"`)			√	√	
`Ci`					√
`tol`			√	√	√
`maxit`			√	√	√
`MSE`	√	√	√	√	√
`mse_type` (only if `MSE = TRUE`)	√	√	√	√	√
`B`	(√)	√	√	√	
`seed`	(√)	(√)	(√)	(√)	

Table 6: Required √ and optional (√) input arguments of function `fh` for the different area-levels models. `B`: Only if bootstrap MSE is chosen. When the standard FH model is applied, `B` is required for the computation of the information criteria by Marhuenda et al. (2014) (optionally).

8 Appendix B: Output of the model component

Name	Short description	Available for				
		Standard	Transformed	Spatial	Robust	ME
`coefficients`	Estimated regression coefficients	√	√	√	√	√
`variance`	Estimated variance of the random effects/ estimated spatial correlation parameter	√	√	√	√	√
`random_effects`	Random effects per domain	√	√	√	√	√
`real_residuals`	Realized residuals per domain	√	√	√	√	√
`std_real_residuals`	Standardized realized residuals per domain	√	√	√	√	√
`gamma`	Shrinkage factors per domain	√	√			√
`model_select`	Model selection and accuracy criteria	√	√	√		
`correlation`	Selected correlation structure of the random effects	√	√	√	√	√
`k`	Tuning constant				√	
`mult_constant`	Multiplier constant for bias correction				√	
`seed`	Seed of the random number generator	√	√	√	√	

Table 7: Components of the output component `model` for models of class `"fh"`.

9 Reproducibility

For the computation of the results in this paper we worked with R version 4.2.2 on a 64-bit platform under Microsoft Windows 10 with the installed packages listed in Table 8. Using the package **packrat** (Ushey et al., 2022) a snapshot of the corresponding repository was created that is available from the GitHub folder (`https://github.com/SoerenPannier/emdi.git`). We suggest the following steps:

- Install Git.
- Create a new project in RStudio.
- Choose checkout from version control and select Git.
- Insert the repository URL: `https://github.com/SoerenPannier/emdi.git`.
- Let **packrat** complete the initialization process.
- Restart RStudio.
- Enter the R command `packrat::restore()`.
- After finishing the installation process all packages are installed as provided in Table 8.

Sylvia Harmening
Institute for Statistics and Econometrics, School of Business & Economics, Freie Universität Berlin
Garystr. 21, 14195 Berlin
Germany
sylvia.harmening@fu-berlin.de

Ann-Kristin Kreutzmann
Institute for Statistics and Econometrics, School of Business & Economics, Freie Universität Berlin
Garystr. 21, 14195 Berlin
Germany
ann-kristin.kreutzmann@fu-berlin.de

Sören Schmidt
Institute for Statistics and Econometrics, School of Business & Economics, Freie Universität Berlin
Garystr. 21, 14195 Berlin
Germany
soeren.pannier@fu-berlin.de

Nicola Salvati
Department of Economics and Management, University of Pisa
Via C. Ridolfi, 10 56124 Pisa
Italy
nicola.salvati@unipi.it

Timo Schmid
Institute of Statistics, Otto-Friedrich-Universität Bamberg
Feldkirchenstr. 21, 96052 Bamberg
Germany
timo.schmid@uni-bamberg.de

Package	Version	Package	Version	Package	Version
aoos	0.5.0	highr	0.9	RColorBrewer	1.1-3
assertthat	0.2.1	HLMdiag	0.5.0	Rcpp	1.0.9
backports	1.4.1	hms	1.1.1	RcppArmadillo	0.11.2.0.0
BBmisc	1.12	isoband	0.2.5	readODS	1.7.0
bit	4.0.4	janitor	2.1.0	readr	2.1.2
bit64	4.0.5	jsonlite	1.8.0	rematch	1.0.1
boot	1.3-28	knitr	1.39	rematch2	2.1.2
brew	1.0-7	labeling	0.4.2	reshape2	1.4.4
brio	1.1.3	laeken	0.5.2	rgeos	0.5-9
cachem	1.0.6	lifecycle	1.0.1	rlang	1.0.4
callr	3.7.1	lubridate	1.8.0	roxygen2	7.2.1
cellranger	1.1.0	magrittr	2.0.3	rprojroot	2.0.3
checkmate	2.1.0	maptools	1.1-4	s2	1.1.0
classInt	0.4-7	MASS	7.3-58	saeRobust	0.3.0
cli	3.3.0	memoise	2.0.1	scales	1.2.0
clipr	0.8.0	modules	0.10.0	sf	1.0-8
colorspace	2.0-3	moments	0.14.1	simFrame	0.5.4
commonmark	1.8.0	MuMIn	1.47.1	snakecase	0.11.0
cpp11	0.4.2	munsell	0.5.0	sp	1.5-0
crayon	1.5.1	nlme	3.1-158	spData	2.0.1
data.table	1.14.2	openxlsx	4.2.5	spdep	1.2-4
DBI	1.1.3	operator.tools	1.6.3	stringi	1.7.8
deldir	1.0-6	packrat	0.8.1	stringr	1.4.0
desc	1.4.1	parallelMap	1.5.1	terra	1.5-34
diagonals	6.4.0	pbapply	1.5-0	testthat	3.1.4
diffobj	0.3.5	pillar	1.8.0	tibble	3.1.8
digest	0.6.29	pkgconfig	2.0.3	tidyr	1.2.0
dplyr	1.0.9	pkgload	1.3.0	tidyselect	1.1.2
e1071	1.7-11	plyr	1.8.7	tzdb	0.3.0
ellipsis	0.3.2	praise	1.0.0	units	0.8-0
emdi	2.1.3	prettyunits	1.1.1	utf8	1.2.2
evaluate	0.15	processx	3.7.0	vctrs	0.4.1
fansi	1.0.3	progress	1.2.2	viridisLite	0.4.0
farver	2.1.1	proxy	0.4-27	vroom	1.5.7
fastmap	1.1.0	ps	1.7.1	waldo	0.4.0
formula.tools	1.7.1	purrr	0.3.4	withr	2.5.0
fs	1.5.2	R.cache	0.16.0	wk	0.6.0
generics	0.1.3	R.methodsS3	1.8.2	xfun	0.31
ggplot2	3.3.6	R.oo	1.25.0	xml2	1.3.3
ggrepel	0.9.1	R.rsp	0.45.0	yaml	2.3.5
glue	1.6.2	R.utils	2.12.0	zip	2.2.0
gridExtra	2.3	R6	2.5.1		
gtable	0.3.0	raster	3.5-21		

Table 8: Installed packages for the computation of the results in this paper.