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

Sylvia Harmening
Freie Universität Berlin

Ann-Kristin Kreutzmann
Freie Universität Berlin

Sören Pannier
Freie Universität Berlin

Nicola Salvati
University of Pisa

Timo Schmid
Freie Universität Berlin

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 which may be challenging due to confidentiality constraints. In contrast, area-level models are less demanding with respect to (a) data requirements, as only aggregates are needed for estimating regional indicators, and (b) computational resources, and enable the incorporation of design-based properties. Therefore, the area-level model (Fay and Herriot 1979) and various extensions have been added to version 2.0.2 of the package **emdi**. These extensions include amongst others (a) transformed 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 function, model diagnostics, benchmarking options, high quality maps and export options of the results enable the user a complete analysis procedure - from model building to diagnostics. The functionality of the package is demonstrated by illustrative examples based on synthetic data for Austrian districts.

Keywords: Fay-Herriot models, official statistics, survey statistics, small area estimation.

1. Introduction

Small area estimation (SAE) has gained importance not only in research but also in many fields of application to get a better insight of indicators at a small-scale level. 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). Especially official statistics and economic or political decision makers benefit from reliable estimation of disaggregated indicators and thus SAE methods. Existing surveys were often not planned for these disaggregated levels and show only small sample sizes which often lead 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 estimation of the indicators on area- or domain- level. The terms area and domain can be used interchangably 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. Even 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. First, only area-level aggregates are needed for the estimation of 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 different variance estimation methods (maximum likelihood (ML), residual maximum likelihood (REML), Prasad-Rao- and Fay-Herriot method-of-moment) for the standard Fay-Herriot (FH) model and a function to estimate unknown sampling variances. The opportunity of estimating unit- and area-level models under heteroscedasticity is provided by the **JoSAE** package (Breidenbach 2018). The package **saery** (Lefler *et al.* 2014) provides functions for the estimation of temporal FH models. The robust estimation of area-level models with spatial and/or temporal structures in the random effects is supported by package **saeRobust** (Warnholz 2018). The estimation of univariate and multivariate FH models is possible with package **msae** (Permatasari and Ubaidillah 2020). The package **hbsae** (Boonstra 2012) 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 (Shi 2018). Further on, the **mme** package (Lopez-Vizcaino *et al.* 2019) allows the building of Gaussian area-level multinominal mixed-effects models in the SAE context. Package **saeME** (Mubarak and Ubaidillah 2020) comprises an area-level model when the auxiliary variables are measured with error. 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 the packages and the implemented methodology. Package **emdi** version 2.0.2 expands the existing 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 also area-level models that are not available in existing packages: adjusted variance estimation methods and transformation options for the standard FH model.
- Package **emdi** offers user-friendly tools that go beyond model estimation for the new and

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

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

existing methods like specific diagnostic tools both in form of a summary and graphical diagnostics, and the comparison of the model-based with direct estimates and their respective mean squared error (MSE) estimates. Furthermore, benchmarking options, geographically visualization of the results in form of high quality maps, and export of the results to Excel and OpenDocument Spreadsheet are provided.

- 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, the newly introduced package version 2.0.2 extends the current version 1.1.7 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. Section 2 introduces the statistical methods implemented in the package. The included example data sets are presented in Section 3. Section 4 provides an illustrative description of the functions using the example data sets. While Section 4.1 guides the reader from model building to model diagnostics of a standard FH model and exporting the results to Excel, Section 4.2 follows with relatively short descriptions of how to build the different extended area-level models. Finally, Section 5 concludes 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 new package release (2.0.2). These comprise the area-level model by Fay and Herriot

(1979) and several extensions of this standard model to account for issues that may come up in real data applications. To measure the precision of those models, respective MSE estimators have been integrated following the literature.

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

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

$\hat{\theta}_i^{\text{Dir}}$ is an unbiased direct estimator for a population indicator of interest θ_i, for instance a mean or a ratio. e_i stands for independent and normally distributed sampling errors with $e_i \overset{ind}{\sim} N(0, \sigma^2_{e_i})$. Even though the model assumes known sampling variances, in practical applications $\sigma^2_{e_i}$ 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 continous 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 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, u_i is an independent and identically normally distributed random effect with $u_i \overset{iid}{\sim} N(0, \sigma^2_u)$.
The combination of the sampling and the linking model leads to a special linear mixed model

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

The empirical best linear unbiased estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ 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^2_u 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 $\boldsymbol{\sigma_u^2}$

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 are supposed to be strictly positive. In the estimation methods mentioned above, negative variance estimates are set to zero ($\hat{\sigma}_u^2 = max(\tilde{\sigma}_u^2, 0)$) 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 a factor:

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

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`), adjusted ML and REML following Li and Lahiri (2010) (`amrl`, `ampl`) and Yoshimori and Lahiri (2014) (`amrl_yl`, `ampl_yl`).

2.2. Extended area-level models

In real data applications problems might occur that were theoretically not expected or assumptions of the standard FH model, e.g., normality and independency of the error terms, may be violated. The following Section outlines the extensions of the standard FH model that are implemented in package **emdi**.

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.* 2019a). In package **emdi**, the direct estimates and their variances are transformed following Neves *et al.* (2013):

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

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 $\mathsf{VAR}(\hat{\theta}_i^{\text{Dir*log}})$ 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 correction after the back-transformation (Slud and Maiti 2006; Sugawasa and Kubokawa 2017). Package **emdi** allows to choose 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. Section 2.3). The Slud-Maiti back-transformation is derived for the ML variance estimation of the random effect and cannot be applied in the presence of out-of-sample domains, because the back-transformation contains the estimate of the shrinkage factor on domain level. In those cases, the *crude* method can be applied which allows to use also other variance estimation methods.

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

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

where the *arcsin denotes the arcsin transformed scale and $\tilde{n}_i$ the effective sample size which can be described as the sample size adjusted by the sampling design (Jiang *et al.* 2001). The FH model is estimated using Equation 2 and the results are additionally truncated to the interval $[0, \pi/2]$ to ensure results between 0 and 1, if needed. 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):

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

Spatial FH model

The standard FH model assumes independency of the random effects. When working with geographical areas, assuming correlated random effects to incorporate a certain neighbouring structure can be valuable. Package **emdi** 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. The higher the absolute value, the stronger the relationship with the neighboring areas. The random effect u_i in Equation 1 is replaced by

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

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 (`ml`) and REML (`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.* 2019). Even before estimating the model, package **emdi** enables the testing for spatial correlation by the Moran's I and Geary's C statistics (Cliff and Ord 1981; Pratesi and Salvati 2008). While Moran's I mimics an usual 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). Both statistics behave inversely to each other.

Robust area-level models

For the case of influential outlying observations, package **emdi** allows for robust versions of the standard and the spatial FH model. The theory is extensively studied in Warnholz (2016) that extended the robust estimation procedure for linear mixed models suggested by Sinha and Rao (2009) to area-level models. The model fitting can be understood as a robustified ML version that also contains an influence function together with a tuning constant `k`. The recommendation is to set the tuning constant to 1.345 (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 multiplyer constant (`mult_constant`) that is used for the bias correction. 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 surveys or alternative data sources this assumption can be violated. Package **emdi** 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 needs to be given to the model. The modified shrinkage factor pulls more weight on the direct estimator when the variances of the covariates are large. For the estimation of the $\boldsymbol{\beta}$s and the σ_u^2, they used a modified method of weighted least squares and a moment estimator, respectively. Additional details are available in Ybarra and Lohr (2008).

2.3. 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). Package **emdi** offers a variety of MSE estimators stemming from both analytical determination and resampling strategies like boostrap and jackknife methods. Table 2 gives an overview about the included MSE approaches. For each area-level model presented in Sections 2.1 and 2.2, the provided MSE type(s) is (are) shown. Please refer to the quoted references for extensive formulas and derivations. As additional measure of variability of the direct and FH estimates, within various functions and methods of package **emdi**, the coefficient of variation (CV) is provided: $CV = \sqrt{\widehat{MSE}(\hat{\theta}_i)}/\hat{\theta}_i$, where $\hat{\theta}_i$ either stands for $\hat{\theta}_i^{\text{Dir}}$ or $\hat{\theta}_i^{\text{FH}}$.

3. Data sets

The version 1.1.7 of package **emdi** contains a sample (`eusilcA_smp`) and a population data set (`eusilcA_pop`) at a household level. The data generating processs for both data sets is extensively described in Kreutzmann *et al.* (2019b). Besides the modification of not producing out-of-sample domains for the area-level version of the data sets, the process is almost equivalent. As basis for the data sets serves the synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) data set (`eusilcP`) from 2006 of the **simFrame** package (Alfons *et al.* 2010). 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 of the package **emdi**. 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 equiv-

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.

alized household income `eqIncome` in `eusilcA_smp` corresponding to the total income of a household divided by the size of the household that is equalised 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` since these are used in the model extensions. The population data set `eusilcA_popAgg` contains a variety of variables that describe different income sources of households and a variable that describes the ratios of the population sizes per area and the total population size `ratio_n`. The variable `Domain` exists in both data sets and indentifies the different districts. Both data sets have 94 observations standing for the 94 Austrian districts, the sample data set `eusilcA_smpAgg` contains eight variables and the population data set `eusilcA_popAgg` 15. Table 3 provides an overview of all included variables of the sample and population data set. For the creation of the proximity matrix used in the spatial FH model and also for 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. It stems from the SynerGIS website (Bundesamt für Eich- und Vermessungswesen 2017). The data set `eusilcA_prox` comprising an exemplary proximity matrix is also added to package **emdi**. The creation of `eusilcA_prox` is described in Section 4.1.

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.

4. Functionality and case studies

While the theoretical background of the implemented area-level models has been introduced in Section 2, the focus of Section 4 lies on the functionality and the work flow 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 default settings if specified. Not every argument needs a specification for every estimated model. Depending on the area-level model, different arguments have to be determined (see Table 6 in Appendix A). The flow diagram of Figure 1 demonstrates the estimation possibilities of a standard FH model introduced in Section 2.1. In line with the `direct` and `ebp` functions of package version 1.1.7,

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 multiplyer 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 iteration for the variance estimation	100
`MSE`	MSE estimation	FALSE
`mse_type`	Type of MSE estimator	analytical
`B`	Numbers of bootstrap iteration 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`.

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. Even though all of the returned objects contain ten components, not every component is available for each estimation method such that 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 for the objects of class '`fh`' are provided in Table 7 in Appendix B. Not all of 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 enable an intuitive interpretation. Due to the consistent structure, all functions and methods of **emdi** version 1.1.7 can be applied to

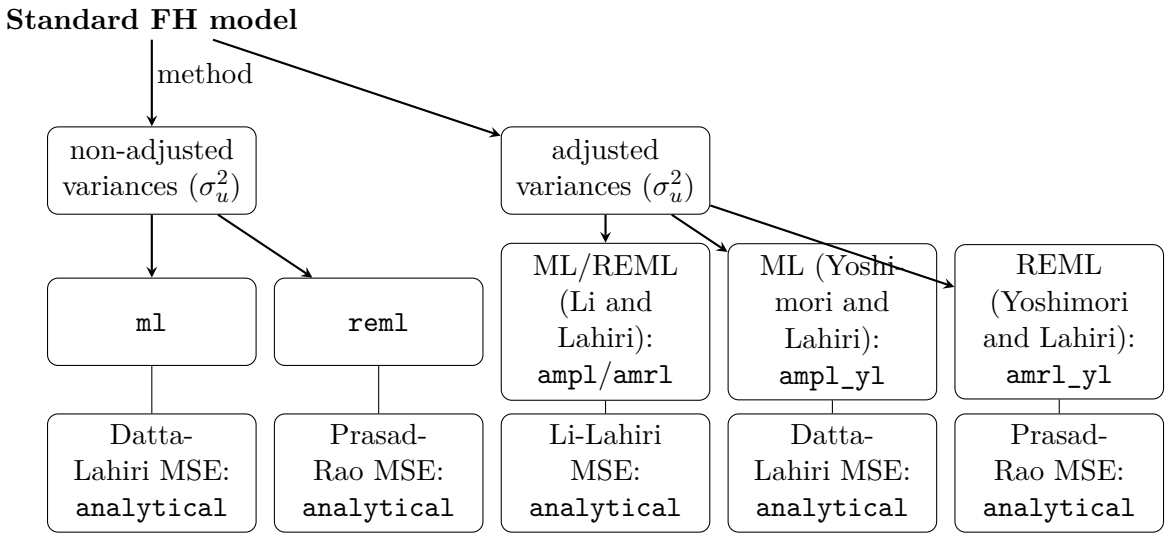


Figure 1: Overview of the standard FH model and adjusted variance estimation methods.

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 package version 2.0.2 for the different '`emdi`' objects. New generic functions comprise for example `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. Section 4.1 explains the procedure shown in Figure 2 step by step for the standard FH model by using the Austrian EU-SILC data described in Section 3. To understand how the different extended area-level models are fitted with function `fh`, Section 4.2 shortly gives instructions.

4.1. Estimation procedure for the standard Fay-Herriot model

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

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

Combine input data

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

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

	Name	Description	Available for		
			direct	ebp	fh
1	ind	Point estimates per area	√	√	√
2	MSE	Variance/MSE estimaties 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 distuingished in 'direct', 'ebp' and 'fh'. More detailed information are provided by the package documentation.

Identify spatial structures

With the help of a proximity matrix, the 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 has to be loaded. We load the Austrian shapefile that is provided by package **emdi** for our example and merge it to the sample data set by using the respective domain identifiers with the help of the `merge` method from package **sp** (Pebesma and Bivand 2005). Before merging, we sort the Austrian shapefile corresponding to the order of the domains in the sample data.

```
R> library("sp")
R> load_shapeaustria()
R> shape_austria_dis <- shape_austria_dis[order(shape_austria_dis$PB),]
R> 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.

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

Thus, a row standardized proximity matrix is generated that initially had weights amounting to one if an area shares a boundary with another area and to zero when the respective areas

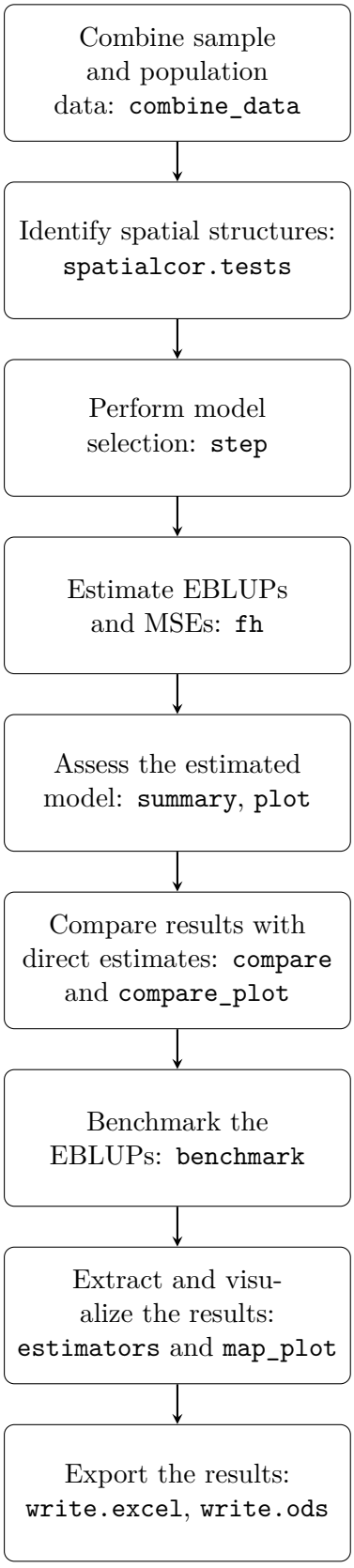


Figure 2: Estimation procedure for area-level models.

are not neighbours. Function `spatialcor.tests` makes use of the `moran.test` and `geary.test` functions with their respective default settings of package **spdep**. The input arguments are the created matrix and the direct estimates.

```
R> 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 of 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, package **emdi** provides the AIC, BIC, the 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 package **sae**, but package **emdi** enables a stepwise variable selection procedure based on the chosen information criteria comparable to the `step` function for `lm` models of package **stats** (R Core Team 2020). 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 equivalized household size (`eqsize`), 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` has to 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.

```
R> 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))
R> 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 the lowest when the variable `unempl_ben` is removed. Therefore, the formula `Mean ~ 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).

```
R> 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 not or only little discussed. One reason for this might be the lack of existing implementation of those 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 R^2 and the adjusted R^2 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.

```
R> 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        R2     AdjR2
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 significant even on a small significance level and their explanatory power is large with an adjusted R^2 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.

```
R> 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 the **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 would not be satisfied with the results, applying a log-transformation could improve the distribution of the error terms.

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. Both 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/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 firstly via boxplots and secondly 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`).

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

Except of 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)

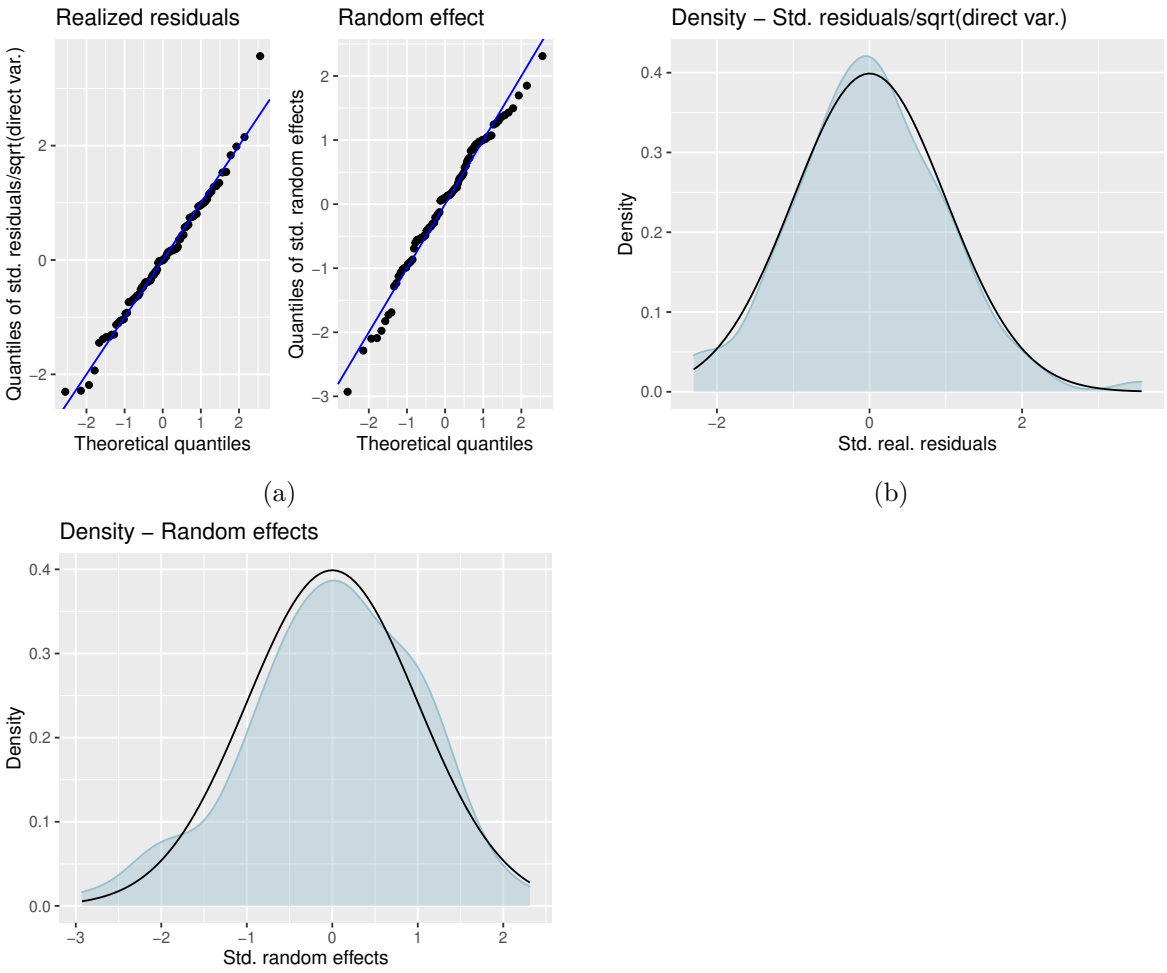


Figure 3: Output of `plot(fh_std)`: (a) normal quantile-quantile (Q-Q) plots of the standardized realized residuals and random effects, (b) and (c): kernel densities of the distribution of the standardized realized residuals and random effects (blue) in comparison to a standard normal distribution (black).

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).

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

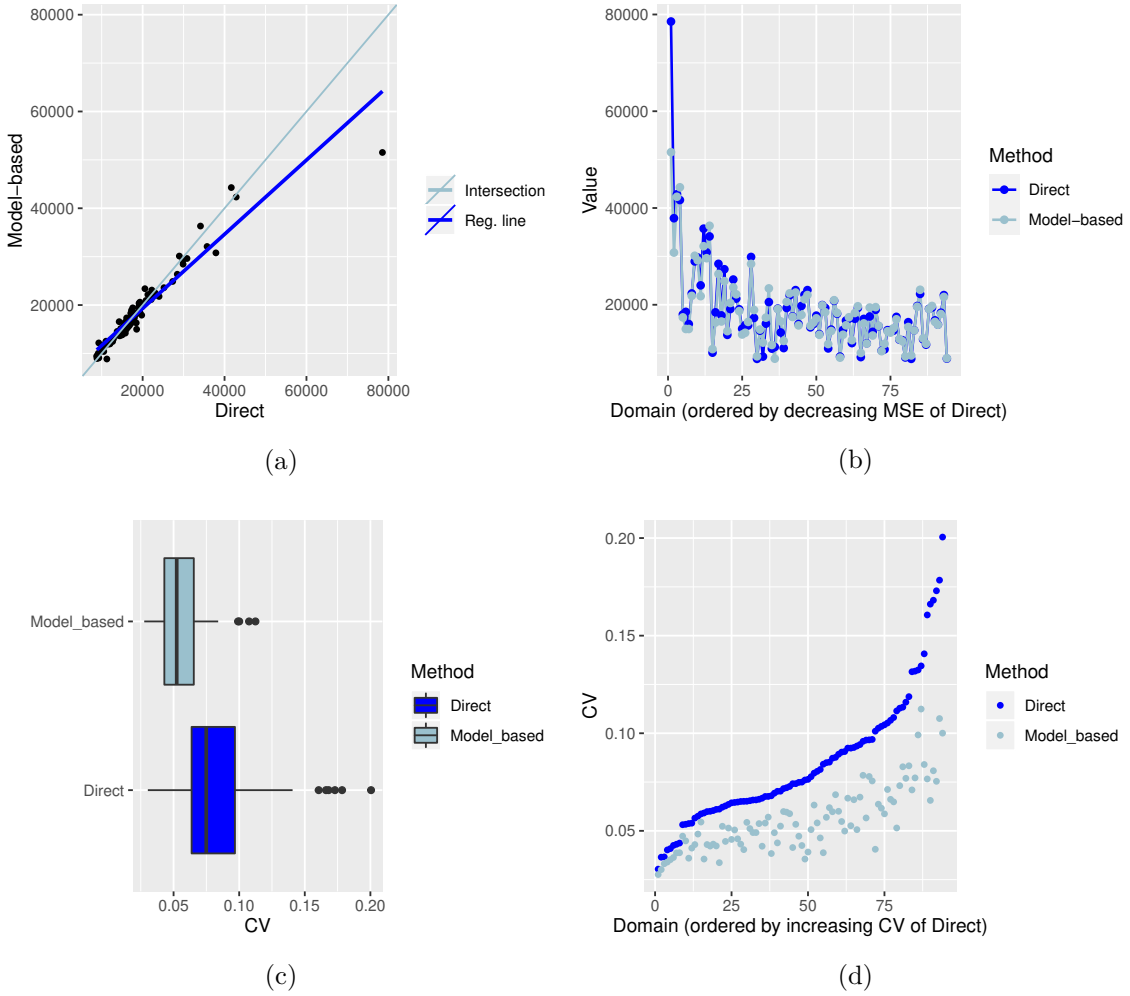


Figure 4: Output of `compare_plot(fh_std)`: (a) and (b) scatter and line plots of direct and model-based point estimates, (c) and (d) boxplot and scatter plots of the CV estimates of the direct and model-based (FH) estimates.

be significant (null hypothesis). The Wald test statistic is specified as

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

and is 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.

```
R> 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. Package **emdi** 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`), $\phi_i = \xi_i/\widehat{\text{MSE}}(\hat{\theta}_i^{\text{FH}})$. 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`.

```
R> fh_bench <- benchmark(fh_std, benchmark = 20140.09,
+    share = eusilcA_popAgg$ratio_n, type = "ratio")
R> 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 recognizeable that for the first six Austrian districts the original estimates are slightly modified by the benchmarking.

Extract and visualize the results

The generic function `estimators` offers an easy way to overview the point, MSE and CV results of the direct estimates compared to the model-based (FH) results. The following output shows the point and MSE results for the first six domains in Austria.

```
R> head(estimators(fh_std, MSE = TRUE))
```

```
               Domain   Direct Direct_MSE       FH    FH_MSE
1          Amstetten 14768.57   926167.4 14242.04  599010.6
2              Baden 21995.72   446534.3 21616.40  356586.1
3            Bludenz 12069.59  1243265.0 12680.38  716040.1
4     Braunau am Inn 10770.53  1029502.4 11925.82  643500.2
5            Bregenz 35731.20  4467316.4 32101.69 1302156.0
6 Bruck-Mürzzuschlag 23027.37  1971664.0 22523.50  906339.2
```

While the highest equalized income of the considered domains was found in Bregenz, the lowest was estimated for Braunau am Inn. The MSE estimates of the EBLUPs are always lower than those of the direct estimates, indicating that the precision of the direct estimates could be improved with the help of the FH model.

Differences among the areas or hotspots of special interest are easier to detect on maps. With function `map_plot`, package **emdi** 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.* (2019b) 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.

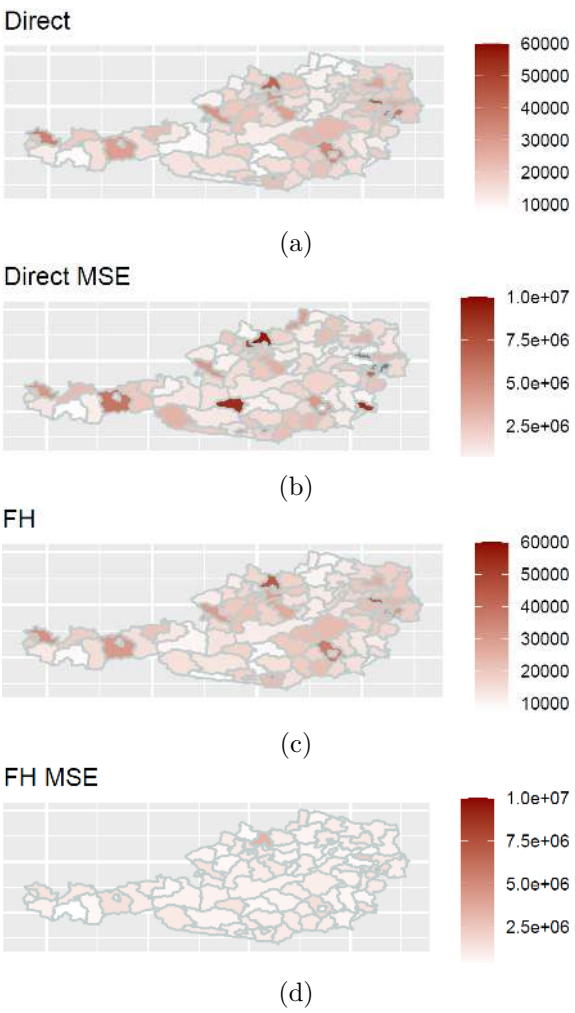


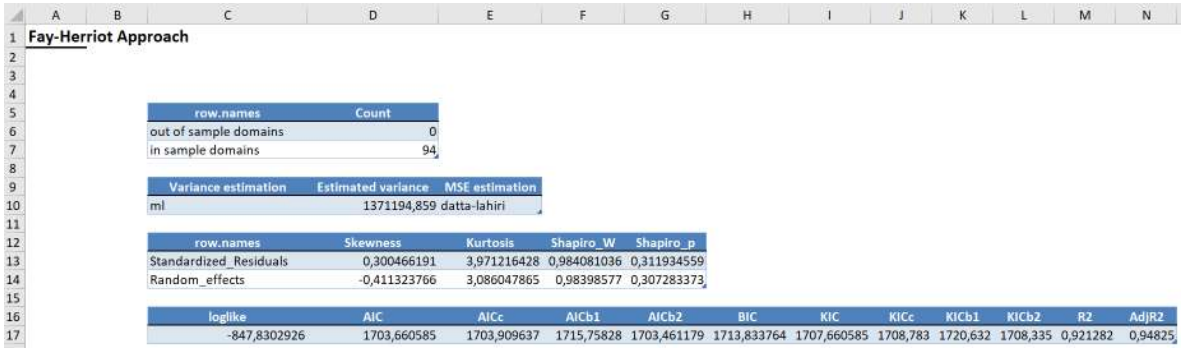
Figure 5: Output of `map_plot()`: Maps of the direct and FH estimates ((a) and (c)) with corresponding MSE estimates ((b) and (d)).

```
R> load_shapeaustria()
R> 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))))
```

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 like for example 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 provides many opportunities for that. Compared to some existing R packages, the **emdi** function `write.excel` does not only export the estimation results to Excel, but also the output of `summary`. The input arguments are again similar to the `estimators` command except that the newly created path and filename of the spreadsheet `file` has to be specified. The output consists of a new Excel file which shows the `summary` output on the first sheet and the estimation results on the second sheet. The package **openxlsx** (Walker 2020) has been used for the linkage with Excel. When working with Microsoft Windows an extra zipping applications for R is necessary for the usage of package **openxlsx** (Walker 2020). Thus, the user is recommended to install RTools. For Linux and macOS zipping application are automatically installed. Using a similar syntax, the results can also be exported to OpenDocument



Fay-Herriot Approach

row.names	Count
out of sample domains	0
in sample domains	94

Variance estimation	Estimated variance	MSE estimation
ml	1371194,859	datta-lahiri

row.names	Skewness	Kurtosis	Shapiro_W	Shapiro_p
Standardized_Residuals	0,300466191	3,971216428	0,984081036	0,311934559
Random_effects	-0,411323766	3,086047865	0,98398577	0,307283373

loglike	AIC	AICc	AICb1	AICb2	BIC	KIC	KICc	KICb1	KICb2	R2	AdjR2
-847,8302926	1703,660585	1703,909637	1715,75828	1703,461179	1713,833764	1707,660585	1708,783	1720,632	1708,335	0,921282	0,94825

(a)

	A	B	C	D	E	F	G
1	Domain	Direct	Direct_MSE	Direct_CV	FH	FH_MSE	FH_CV
2	Amstetten	14768,56933	926167,3714	0,065163787	14242,04457	599010,649	0,054343165
3	Baden	21995,72487	446534,2852	0,030380095	21616,39582	356586,0515	0,027624784
4	Bludenz	12069,59239	1243265,013	0,092382403	12680,37578	716040,1177	0,066732371
5	Braunau am Inn	10770,53331	1029502,352	0,094205544	11925,8169	643500,244	0,067264547
6	Bregenz	35731,19812	4467316,434	0,059152864	32101,68983	1302156,03	0,035547054
7	Bruck-Mürzzuschlag	23027,3744	1971664,032	0,06097784	22523,49503	906339,1859	0,042267795
8	Bruck an der Leitha	25209,50992	3135150,031	0,070236807	23590,33007	1069157,926	0,043831558
9	Deutschlandsberg	21271,28902	3000062,465	0,081427545	22159,12123	1055862,452	0,046371499
10	Dornbirn	20552,06381	2374522,488	0,074977802	23382,35334	986784,1129	0,042483751

(b)

Figure 6: Extract of the Excel spreadsheets created by `write.excel`: (a) `summary` Output, (b) estimation results.

Spreadsheets by the command `write.ods`. The difference to `write.excel` is that multipe files are created. The output of the FH model is exemplarily exported to Excel.

```
R> write.excel(fh_std, file = "fh_std_output.xlsx", MSE = TRUE,
+     CV = TRUE)
```

Figure 6 provides an insight of the output.

4.2. Estimation of the extended area-level models

This section is dedicated to the model building of the extensions of the standard FH model (see Section 2.2) implemented in **emdi**. Figure 7 in Appendix A provides an overview of the options that can be chosen and Table 6 summarizes which arguments have to be specified for the respective models.

FH model with transformation

If the indicator of interest needs a transformation, either log or arcsin, in addition to the function used in Section 4.2, 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.

```
R> 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

In case the spatial correlation tests conducted in Section 4.1 would have 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 proximity matrix exemplarily created in Section 4.1 has to be given to the model within the `corMatrix` argument. The possible variance estimation methods are `"ml"` and `"reml"`.

```
R> 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 equal `"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** Warnholz (2018) are utilized for the robust extensions. An exemplary call with pseudolinear MSE estimation looks like this:

```
R> 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 has to 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

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

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. In R the array is computed by

```
P <- number of covariates
M <- number of areas

Ci_array <- array(data = 0, dim = c(P + 1, P + 1, M))

for(i in 1:M){
  Ci_array[2,2,i] <- Var_a[i]
  Ci_array[3,3,i] <- Var_b[i]
  Ci_array[4,4,i] <- Var_c[i]
  Ci_array[3,2,i] <- Ci_array[2,3,i] <- Cov_ab[i]
  Ci_array[4,2,i] <- Ci_array[2,4,i] <- Cov_ac[i]
  Ci_array[4,3,i] <- Ci_array[3,4,i] <- Cov_bc[i]
}
```

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.

```
R> P <- 1
R> M <- 94
R>
R> Ci_array <- array(data = 0, dim = c(P + 1, P + 1, M))
R>
R> for(i in 1:M){
+    Ci_array[2,2,i] <- eusilcA_smpAgg$Var_Cash[i]
+  }
R>
R> 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 by various area-level models. Besides 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 Section 4 it becomes clear that the package does not only contain the 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. The **emdi** version 2.0.2 accounts for spatial structures of the random effects. Future developments will 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 will 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.

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.

References

Alfons A, Templ M (2013). "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package **laeken**." *Journal of Statistical Software*, **54**(15), 1–25. doi:10.18637/jss.v054.i15.

Alfons A, Templ M, Filzmoser P (2010). "An Object-Oriented Framework for Statistical Simulation: The R package **simFrame**." *Journal of Statistical Software*, **37**(3), 1–36. doi:10.18637/jss.v037.i03.

Battese GE, Harter RM, Fuller WA (1988). "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. doi:10.1080/01621459.1988.10478561.

Bertarelli G, Schirripa Spagnolo F, Salvati N, Pratesi M (2019). "Small Area Estimation of Agricultural Data." In *Spatial Econometric Methods in Agricultural Economics Using R*. CRC book. Accepted to be published.

Bivand RS, Wong DWS (2018). "Comparing Implementations of Global and Local Indicators of Spatial Association." *TEST*, **27**(3), 716–748. doi:10.1007/s11749-018-0599-x.

Boonstra H (2012). ***hbsae**: Hierarchical Bayesian Small Area Estimation*. R package version 1.0, URL https://CRAN.R-project.org/package=hbsae.

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

Brown G, Chambers R, Heady P, Heasman D (2001). "Evaluation of Small Area Estimation Methods - An Application to Unemployment Estimates from the UK LFS." In *Proceedings of Statistics Canada Symposium*.

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

Casas-Cordero C, Encina J, Lahiri P (2016). *Poverty Mapping for the Chilean Comunas*, pp. 379–403. Wiley. doi:10.1002/9781118814963.ch20.

Chambers J, Hastie T (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.

Chambers R, Chandra H, Salvati N, Tzavidis N (2014). "Outlier Robust Small Area Estimation." *Journal of the Royal Statistical Society Series B*, **76**(1), 47–69. doi:10.1111/rssb.12019.

Chandra H, Salvati N, Chambers R (2015). "A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation." *Journal of the Survey Statistics and Methodology*, **3**(2), 109–135. doi:10.1093/jssam/smu026.

Chen S, Lahiri P (2002). "A Weighted Jackknife MSPE Estimator in Small-Area Estimation." In *Proceeding of the Section on Survey Research Methods*, pp. 473–477. American Statistical Association.

Cliff A, Ord J (1981). *Spatial Processes: Models and Applications*. Pion, London.

Datta G, Ghosh M, Steorts R, Maples J (2011). "Bayesian Benchmarking with Applications to Small Area Estimation." *TEST*, **20**(3), 574–588. doi:10.1007/s11749-010-0218-y.

Datta GS, Fay RE, Ghosh M (1991). "Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation." In *Proceedings of Bureau of the Census 1991 Annual Research Conference*, pp. 63–79. US Bureau of the Census.

Datta GS, Lahiri P (2000). "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems." *Statistica Sinica*, **10**(2), 613–627. URL http://www.jstor.com/stable/24306735.

Fay RE, Herriot RA (1979). "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. doi:10.1080/01621459.1979.10482505.

Hadam S, Würz N, Kreutzmann AK (2020). "Estimating Regional Unemployment with Mobile Network Data for Functional Urban Areas in Germany." *Refubium - Freie Universität Berlin Repository*, pp. 1–28. doi:10.17169/refubium-26791.

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

Horvitz D, Thompson D (1952). "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*, **47**(260), 663–685. doi:10.1080/01621459.1952.10483446.

Jiang J, Lahiri P, Wan SM (2002). "A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation." *The Annals of Statistics*, **30**(6), 1782–1810. doi:10.1214/aos/1043351257.

Jiang J, Lahiri P, Wan SM, Wu CH (2001). "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*, pp. 75–97.

Jiang J, Rao JS (2020). "Robust Small Area Estimation: An Overview." *Annual Review of Statistics and Its Application*, **7**, 337–360. doi:10.1146/annurev-statistics-031219-041212.

Komsta L, Novomestky F (2015). ***moments**: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.14, URL https://CRAN.R-project.org/package=moments.

Kreutzmann AK, Marek P, Salvati N, Schmid T (2019a). "Estimating Regional Wealth in Germany: How Different are East and West Really?" *Bundesbank Discussion Paper*. NO 35/2019, URL https://www.bundesbank.de/resource/blob/807786/ed9f7534c3ba2039aabeb685295cca1f/mL/2019-09-24-dkp-35-data.pdf.

Kreutzmann AK, Pannier S, Rojas-Perilla N, Schmid T, Templ M, Tzavidis N (2019b). "The R Package **emdi** for Estimating and Mapping Regionally Disaggregated Indicators." *Journal of Statistical Software*, **91**(7), 1–33. doi:10.18637/jss.v091.i07.

Lahiri P, Suntornchost J (2015). "Variable Selection for Linear Mixed Models with Applications in Small Area Estimation." *The Indian Journal of Statistics*, **77-B**(2), 312–320. URL https://www.jstor.org/stable/43694416.

Lefler M, Gonzalez D, Martin A (2014). ***saery**: Small Area Estimation for Rao and Yu Model*. R package version 1.0, URL https://CRAN.R-project.org/package=saery.

Li H, Lahiri P (2010). "An Adjusted Maximum Likelihood Method for Solving Small Area Estimation Problems." *Journal of Multivariate Analysis*, **101**(4), 882–902. doi:10.1016/j.jmva.2009.10.009.

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

Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L (2015). "Small Area Model-Based Estimators Using Big Data Sources." *Journal of Official Statistics*, **31**(2), 263–281. doi:10.1515/jos-2015-0017.

Marhuenda Y, Molina I, Morales D (2013). "Small Area Estimation with Spatio-Temporal Fay-Herriot Models." *Computational Statistics and Data Analysis*, **58**, 308–325. doi:10.1016/j.csda.2012.09.002.

Marhuenda Y, Morales D, del Camen Pardo M (2014). "Information Criteria for Fay-Herriot Model Selection." *Computational Statistics and Data Analysis*, **70**, 268–280. doi:10.1016/j.csda.2013.09.016.

Miltiadou M (2020). "Measuring and Reporting Reliability of Labour Force Survey and Annual Population Survey Estimates Force Survey and Annual Population Survey Estimates." UK Office for National Statistics, [accessed: 05.06.2020], URL https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/measuringandreportingreliabilityoflabourforcesurveyandannualpopulationsurveyestimates.

Molina I, Marhuenda Y (2015). "**sae**: An R Package for Small Area Estimation." *The R Journal*, **7**(1), 81–98. doi:10.32614/rj-2015-007.

Molina I, Rao J (2010). "Small Area Estimation of Poverty Indicators." *The Canadian Journal of Statistics*, **38**(3), 369–385. doi:10.1002/cjs.10051.

Molina I, Rao J, Datta G (2015). "Small Area Estimation Under a Fay-Herriot Model with Preliminary Testing for the Presence of Random Area Effects." *Survey Methodology*, **41**(1), 1–19. URL https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14161-eng.htm.

Molina I, Salvati N, Pratesi M (2009). "Bootstrap for Estimating the MSE of the Spatial EBLUP." *Computational Statistics*, **24**, 441–458. doi:10.1007/s00180-008-0138-4.

Mubarak M, Ubaidillah A (2020). ***saeME**: Small Area Estimation with Measurement Error*. R package version 1.2.4, URL https://CRAN.R-project.org/package=saeME.

Nandy A (2015). ***smallarea**: Fits a Fay-Herriot Model*. R package version 0.1, URL https://CRAN.R-project.org/package=smallarea.

Neves A, Silva D, Correa S (2013). "Small Domain Estimation for the Brazilian Service Sector Survey." *ESTADÍSTICA*, **65**(185), 13–37.

Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R." *R News*, **5**(2), 9–13. URL https://CRAN.R-project.org/doc/Rnews/.

Permatasari N, Ubaidillah A (2020). ***msae**: Multivariate Fay Herriot Models for Small Area Estimation*. R package version 0.1.2, URL https://CRAN.R-project.org/package=msae.

Petrucci A, Salvati N (2006). "Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment." *Journal of Agricultural, Biological and Environmental Statistics*, **11**(2), 169–182. doi:10.1198/108571106X110531.

Pfeffermann D (2013). "New Important Developments in Small Area Estimation." *Statistical Science*, **28**(1), 40–68. doi:10.1214/12-STS395.

Prasad N, Rao J (1990). "The Estimation of the Mean Squared Error of Small-Area Estimation." *Journal of the American Statistical Association*, **85**(409), 163–171. doi:10.1080/01621459.1990.10475320.

Pratesi M (ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. Wiley. doi:10.1002/9781118814963.

Pratesi M, Salvati N (2008). "Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects." *Statistical Methods and Applications*, **17**(1), 113–141. doi:10.1007/s10260-007-0061-9.

Rao JNK, Molina I (2015). *Small Area Estimation*. New York: Wiley. doi:10.1002/9781118735855.

Rao JNK, Yu M (1994). "Small-Area Estimation by Combining Time-Series and Cross-Sectional Data." *The Canadian Journal of Statistics*, **22**(4), 511–528. doi:10.2307/3315407.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rivest LP, Vandal N (2003). "Mean Squared Error Estimation for Small Areas when the Small Area Variances are Estimated." In *Proceedings of International Conference of Recent Advanced Survey Sampling*, pp. 197–206.

Saei A, Chambers R (2005). "Out of Sample Estimation for Small Areas using Area Level Data." *Southampton Statistical Sciences Research Institute Methodology Working Paper*, **M05/11**. Southampton Statistical Sciences Research Institute, UK, URL http://eprints.soton.ac.uk/id/eprint/14327.

Schmid T, Bruckschen F, Salvati N, Zbiranski T (2017). "Constructing Sociodemographic Indicators for National Statistical Institutes Using Mobile Phone Data: Estimating Literacy Rates in Senegal." *Journal of the Royal Statistical Society Series A*, **180**(4), 1163–1190. doi:10.1111/rssa.12305.

Schmid T, Tzavidis N, Münnich R, Chambers R (2016). "Outlier Robust Small Area Estimation Under Spatial Correlation." *Scandinavian Journal of Statistics: Theory and Applications*, **43**(3), 806–826. doi:10.1111/sjos.12205.

Shi C (2018). ***BayesSAE**: Bayesian Analysis of Small Area Estimation*. R package version 1.0-2, URL https://CRAN.R-project.org/package=BayesSAE.

Singh BB, Shukla K, Kundu D (2005). "Spatio-Temporal Models in Small Area Estimation." *Survey Methodology*, **31**(2), 183–195. URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20050029053.

Sinha S, Rao J (2009). "Robust Small Area Estimation." *The Canadian Journal of Statistics*, **37**(3), 381–399. doi:10.1002/cjs.10029.

Slud E, Maiti T (2006). "Mean-Squared Error Estimation in Transformed Fay-Herriot Models." *Journal of the Royal Statistical Society Series B*, **68**(2), 239–257. doi:10.1111/j.1467-9868.2006.00542.x.

Sugawasa S, Kubokawa T (2017). "Transforming Response Values in Small Area Prediction." *Computational Statistics and Data Analysis*, **114**, 47–60. doi:10.1016/j.csda.2017.03.017.

Tzavidis N, Chambers R, Salvati N, Chandra H (2012). "Small Area Estimation in Practice an Application to Agricultural Business Survey Data." *Journal of the Indian Society of Agricultural Statistics*, **66**(1), 213–228. URL https://ro.uow.edu.au/eispapers/758/.

Tzavidis N, Zhang LC, Luna Hernandez A, Schmid T, Rojas-Perilla N (2018). "From Start to Finish: A Framework for the Production of Small Area Official Statistics." *Journal of the Royal Statistical Society: Series A*, **181**(4), 927–979. doi:10.1111/rssa.12364.

Ushey K, McPherson J, Cheng J, Atkins A, Allaire J (2018). ***packrat**: A Dependency Management System for Projects and their R Package Dependencies*. R package version 0.5.0, URL https://CRAN.R-project.org/package=packrat.

Walker A (2020). ***openxlsx**: Read, Write and Edit XLSX Files*. R package version 4.2.3, URL https://CRAN.R-project.org/package=openxlsx.

Wang J, Fuller WA (2003). "The Mean Squared Error of Small Area Predictors Constructed With Estimated Area Variances." *Journal of the American Statistical Association*, **98**, 716–723. doi:10.1198/016214503000000620.

Warnholz S (2016). *Small Area Estimation Using Robust Extensions to Area Level Models*. Ph.D. thesis. doi:10.17169/refubium-13904. Freie Universität Berlin.

Warnholz S (2018). ***saeRobust**: Robust Small Area Estimation*. R package version 0.2.0, URL https://CRAN.R-project.org/package=smallarea.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag.

Ybarra LMR, Lohr SL (2008). "Small Area Estimation When Auxiliary Information Is Measured with Error." *Biometrika*, **95**(4), 919–931. doi:10.1093/biomet/asn048.

Yoshimori M, Lahiri P (2014). "A New Adjusted Maximum Likelihood Method for the Fay-Herriot Small Area Model." *Journal of Multivariate Analysis*, **124**, 281–294. doi:10.1016/j.jmva.2013.10.012.

You Y, Chapman B (2006). "Small Area Estimation Using Area Level Models and Estimated Sampling Variances." *Survey Methodology*, **32**(1), 97–103. URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263.

Zhang X, Holt J, Yun S, Lu H, Greenlund K, Croft J (2015). "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. doi:10.1093/aje/kwv002.

A. Area-level model options and input arguments of function fh

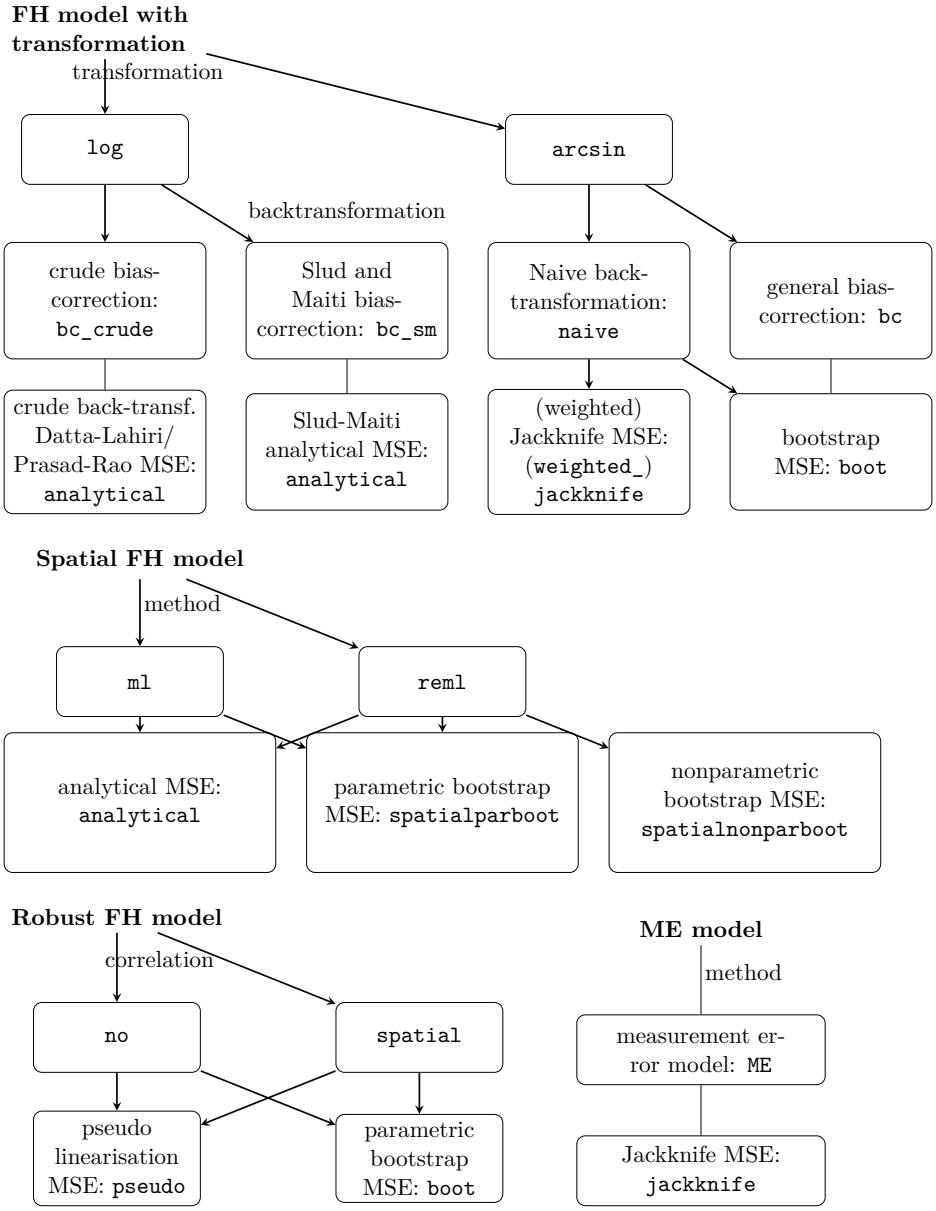


Figure 7: Overview of extended area-level models and combinations of estimation methods.

Argument	**FH model**				
	Standard	Transformed	Spatial	Robust	ME
`fixed`	√	√	√	√	√
`vardir`	√	√	√	√	√
`combined`	√	√	√	√	√
`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).

B. Output of the model component of an fh object

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`	Multiplyer constant for bias correction				√	
`seed`	Seed of the random number generator	√	√	√	√	

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

C. Reproducibility

For the computation of the results in this paper we worked with R version 4.0.3 on a 64-bit platform under Microsoft Windows 10 with the installed packages listed in Table 8. Using the package **packrat** (Ushey *et al.* 2018) 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.

Affiliation:

Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Pannier, Timo Schmid
Institute for Statistics and Econometrics
School of Business & Economics
Freie Universität Berlin
Garystr. 21, 14195 Berlin, Germany
E-mail: sylvia.harmening@fu-berlin.de, ann-kristin.kreutzmann@fu-berlin.de, soeren.pannier@fu-berlin.de, timo.schmid@fu-berlin.de

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

Package	Version	Package	Version	Package	Version
aoos	0.5.0	gtools	3.8.2	R.rsp	0.43.2
assertthat	0.2.1	highr	0.8	R.utils	2.9.2
backports	1.2.0	HLMdiag	0.4.0	R6	2.5.0
BBmisc	1.11	hms	1.0.0	raster	3.4-5
BH	1.75.0-0	isoband	0.2.3	RColorBrewer	1.1-2
boot	1.3-25	janitor	2.1.0	Rcpp	1.0.4.6
brew	1.0-6	jsonlite	1.7.2	RcppArmadillo	0.10.1.2.2
brio	1.1.1	knitr	1.31	readODS	1.7.0
cachem	1.0.1	labeling	0.4.2	readr	1.4.0
callr	3.5.1	laeken	0.5.1	rematch	1.0.1
cellranger	1.1.0	LearnBayes	2.15.1	rematch2	2.1.2
checkmate	2.0.0	lifecycle	0.2.0	reshape2	1.4.4
classInt	0.4-3	lubridate	1.7.9.2	rgeos	0.5-5
cli	2.2.0	magrittr	2.0.1	rlang	0.4.10
clipr	0.7.1	maptools	1.0-2	roxygen2	7.1.1
coda	0.19-4	markdown	1.1	rprojroot	2.0.2
colorspace	2.0-0	MASS	7.3-51.6	rstudioapi	0.13
commonmark	1.7	memoise	2.0.0	saeRobust	0.2.0
cpp11	0.2.5	mime	0.9	scales	1.1.1
crayon	1.3.4	modules	0.8.0	sf	0.9-7
DBI	1.1.1	moments	0.14	simFrame	0.5.3
deldir	0.2-9	MuMIn	1.43.17	snakecase	0.11.0
desc	1.2.0	munsell	0.5.0	sp	1.4-5
diffobj	0.3.3	nlme	3.1-148	spData	0.3.8
digest	0.6.25	openxlsx	4.2.3	spdep	1.1-5
dplyr	1.0.3	operator.tools	1.6.3	stringi	1.5.3
e1071	1.7-4	packrat	0.5.0	stringr	1.4.0
ellipsis	0.3.1	parallelMap	1.5.0	testthat	3.0.1
evaluate	0.14	pbapply	1.4-3	tibble	3.0.5
expm	0.999-6	pillar	1.4.7	tidyr	1.1.2
fansi	0.4.2	pkgbuild	1.2.0	tidyselect	1.1.0
farver	2.0.3	pkgconfig	2.0.3	units	0.6-7
fastmap	1.1.0	pkgload	1.1.0	utf8	1.1.4
formula.tools	1.7.1	plyr	1.8.6	vctrs	0.3.6
gdata	2.18.0	praise	1.0.0	viridisLite	0.3.0
generics	0.1.0	prettyunits	1.1.1	waldo	0.2.3
ggplot2	3.3.3	processx	3.4.5	withr	2.4.1
ggrepel	0.9.1	ps	1.5.0	xfun	0.20
glue	1.4.2	purrr	0.3.4	xml2	1.3.2
gmodels	2.18.1	R.cache	0.14.0	yaml	2.2.1
gridExtra	2.3	R.methodsS3	1.8.0	zip	2.1.1
gtable	0.3.0	R.oo	1.23.0	emdi	2.0.2

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