emdi 2.0.1: A Framework for Producing Small Area Estimates based on Area-Level Models in R

Sylvia Harmening[1], Ann-Kristin Kreutzmann[1], Sören Pannier[1], Nicola Salvati[2], Timo Schmid[1]

[1] Freie Universität Berlin
[2] University of Pisa

The Use of R in Official Statistics
December 2, 2020

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 urosconf.org

What's new in version 2.0.1?

- Package **emdi**: estimation of regionally disaggregated indicators
- Version 1.1.7 (Kreutzmann et al. 2019):
  - direct estimation based on survey data
  - model-based estimation using the unit-level empirical best predictor method (Molina and Rao 2010)
- **Version 2.0.1**: area-level model by Fay and Herriot (1979) and various extensions

Motivation

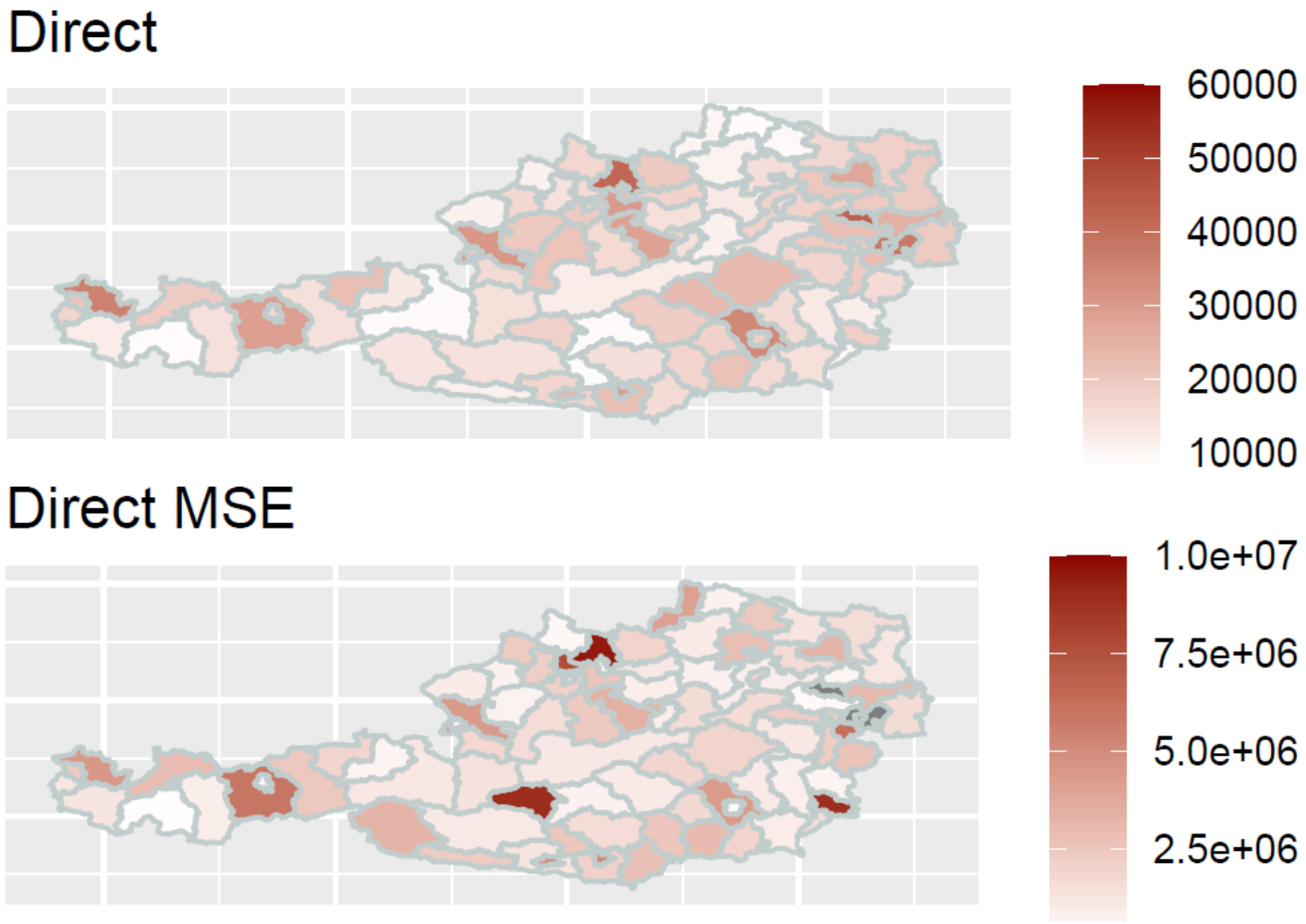


Figure: Direct estimation of equalized income for the 94 Austrian districts

The Fay-Herriot model

Assume

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

- The index i stands for the areas.
- $\hat{\theta}_i^{Dir}$ is a direct estimator based on survey information
- e_i are sampling errors with $e_i \overset{ind}{\sim} N(0, \sigma_{e_i}^2)$

Area-level linear mixed model:

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

- u_i are random effects with $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$
- $\boldsymbol{x}$ are covariates, $\boldsymbol{\beta}$ are regression coefficients

Reference: Fay & Herriot (1979)

The Fay-Herriot model

The EBLUP under the **Fay-Herriot** (FH) model is obtained by

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

- $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ denotes the shrinkage factor for area i.
- The parameter estimates for $\boldsymbol{\beta}$, u_i and σ_u^2 are obtained by ML theory.

Reference: Fay & Herriot (1979)

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 urosconf.org

Included extensions of the standard FH model

Variance estimation of the random effects:

- Zero variance estimates
  → adjusted variance estimation methods (Li and Lahiri 2010, Yoshimori and Lahiri 2014)

Transformations of the direct estimator:

- Violations of the model assumptions (e.g. right skewed data)
  → log transformation (back-transformation options: crude and one following Slud and Maiti, 2006)
- Direct estimator is a ratio
  → arcsin transformation (back-transformation options: naive and one following Hadam et al., 2020)

Included extensions of the standard FH model

Spatial FH model:

- Assumption of correlated random effects
  → spatial FH model by Petrucci and Salvati (2006) that considers a simultaneously autoregressive process of order one SAR(1)

Robust area-level models:

- Influential outlying observations
  → robust versions of the standard and the spatial FH models proposed by Warnholz (2016)

Measurement error model:

- Covariate information from surveys or alternative data sources
  → measurement error model developed by Ybarra and Lohr (2008)

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 urosconf.org

Estimation procedure for area-level models

1. Combine sample and population data: `combine_data`
2. Identify spatial structures: `spatialcor.tests`
3. Perform model selection: `step`
4. Estimate FHs and MSEs: `fh`
5. Assess the estimated model: `summary`, `plot`
6. Compare results with direct estimates: `compare` and `compare_plot`
7. Benchmark the FH estimates: `benchmark`
8. Extract and visualize the results: `estimators` and `map_plot`
9. Export the results: `write.excel`, `write.ods`

Example for the standard area-level model

Estimation of equalized income for the 94 Austrian districts

Data (included as example data in emdi 2.0.1):

- Survey data set `eusilcA_smpAgg`: direct estimates and its variances
- Population data set `eusilcA_popAgg`: auxiliary information

Aim

- Estimation of the equalized income
- Improving the precision of the direct estimates
- Model diagnostics
- Visualization of the results

Step 1: Combine sample and population data

```
1  # Load data
2  data("eusilcA_popAgg")
3  data("eusilcA_smpAgg")
4
5  # Combine data
6  combined_data <- combine_data(
7    pop_data = eusilcA_popAgg, pop_domains = "Domain",
8    smp_data = eusilcA_smpAgg, smp_domains = "Domain")
```

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 urosconf.org

Step 3: Perform model selection (1)

```
1   # Generate initial fh object
2   fh_std <- fh(fixed = Mean ~ cash + self_empl +
3     unempl_ben, vardir = "Var_Mean",
4     combined_data = combined_data, domains = "Domain",
5     method = "ml", B = c(0,200))
6
7   # Perform model selection
8   > step(fh_std, criteria = "KICb2")
9   Start: KICb2 = 1709.11
10  Mean ~ cash + self_empl + unempl_ben
11
12                df   KICb2
13  - unempl_ben   1 1708.0
14  <none>           1709.1
15  - self_empl    1 1762.4
16  - cash         1 1808.1
17
18  Step: KICb2 = 1707.99
19  Mean ~ cash + self_empl
```

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 uroscon.org

Step 3: Perform model selection (2)

```
                df   KICb2
<none>             1708.0
- self_empl     1 1764.6
- cash          1 1815.0

Call:
 fh(fixed = Mean ~ cash + self_empl,
    vardir = "Var_Mean", combined_data = combined_data,
    domains = "Domain", method = "ml", B = c(0, 200))

Coefficients:
             coefficients    std.error   t.value
(Intercept)  3070.512311 635.94290168  4.828283
cash            1.059385   0.07049025 15.028815
self_empl       1.745636   0.22017394  7.928443
                 p.value
(Intercept) 1.377153e-06
cash        4.754350e-51
self_empl   2.219112e-15
```

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 uroscon.org

Step 4: Estimate FHs and MSEs

```
# Estimate FHs and MSEs
fh_std <- fh(fixed = Mean ~ cash + self_empl,
             vardir = "Var_Mean",
             combined_data = combined_data,
             domains = "Domain",
             method = "ml",
             MSE = TRUE)
```

Freie Universität Berlin

UNIVERSITÀ DI PISA

uRos 2020 uroscon.org

Step 5: Assess the estimated model (1)

```
> 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, 200))

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)  3070.512311 635.94290168  4.828283 1.377153e-06
cash            1.059385   0.07049025 15.028815 4.754350e-51
self_empl       1.745636   0.22017394  7.928443 2.219112e-15
```

Freie Universität Berlin

UNIVERSITÀ DI PISA

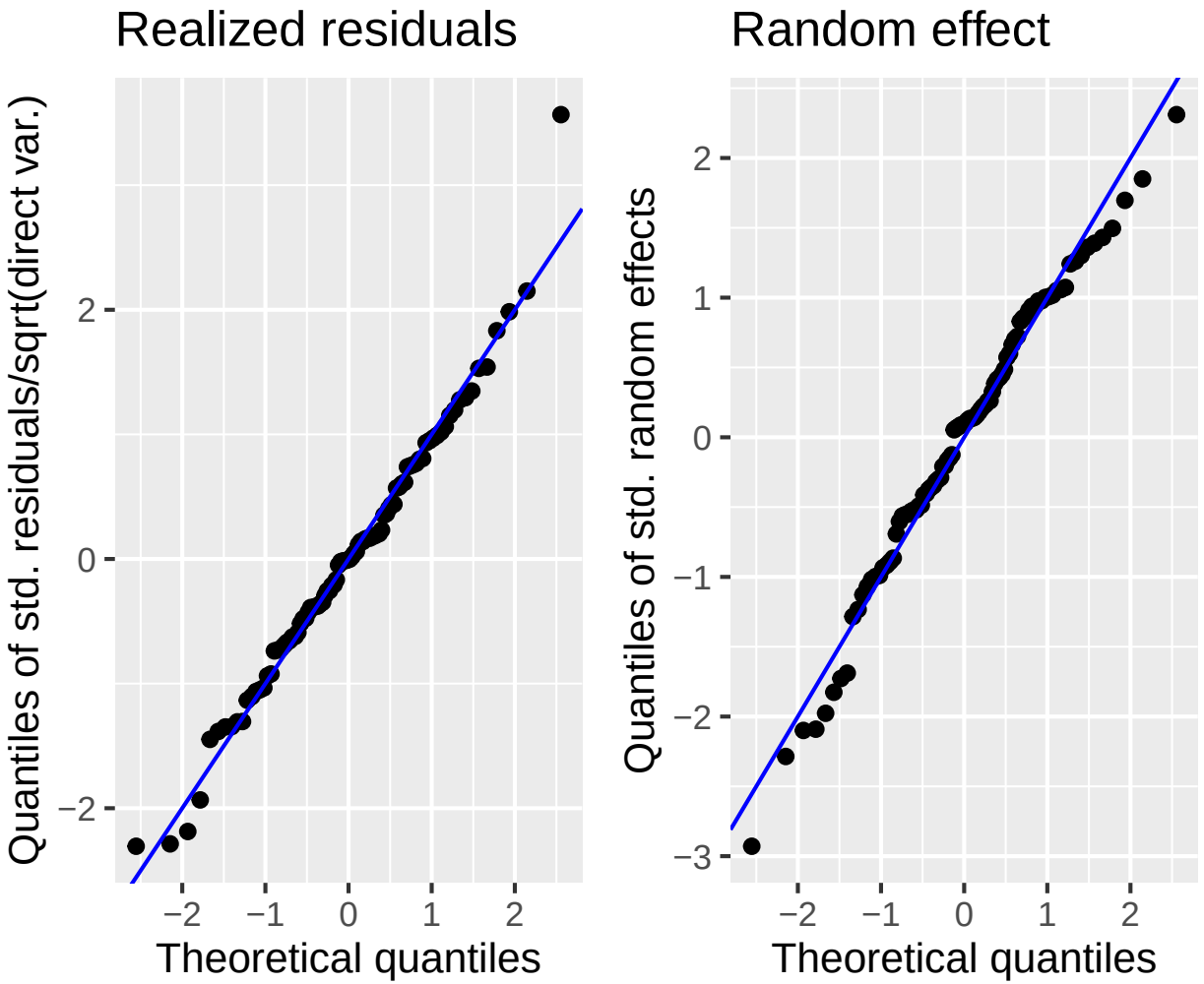
uRos 2020 uroscon.org

Step 5: Assess the estimated model (2)

```
Explanatory measures:
    loglike       AIC      AICc     AICb1     AICb2       BIC
1 -847.8303 1703.661 1703.551 1715.407 1703.866 1713.834
       KIC    KICc     KICb1     KICb2        R2      AdjR2
1 1707.661 1707.67 1719.526 1707.985 0.9212817 0.9482498

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

Transformation: No transformation
```

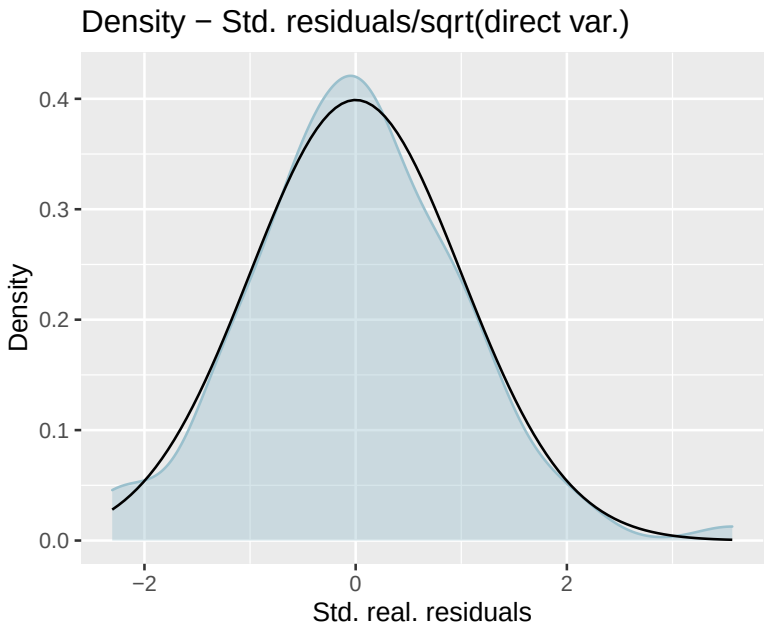
Step 5: Assess the estimated model (3)

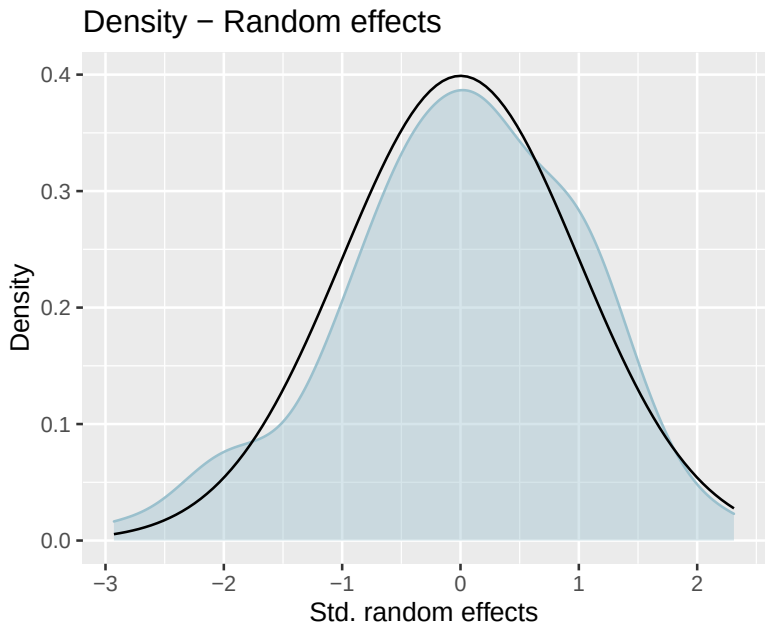
```
1    > plot(fh_std)
```



Step 5: Assess the estimated model (4)

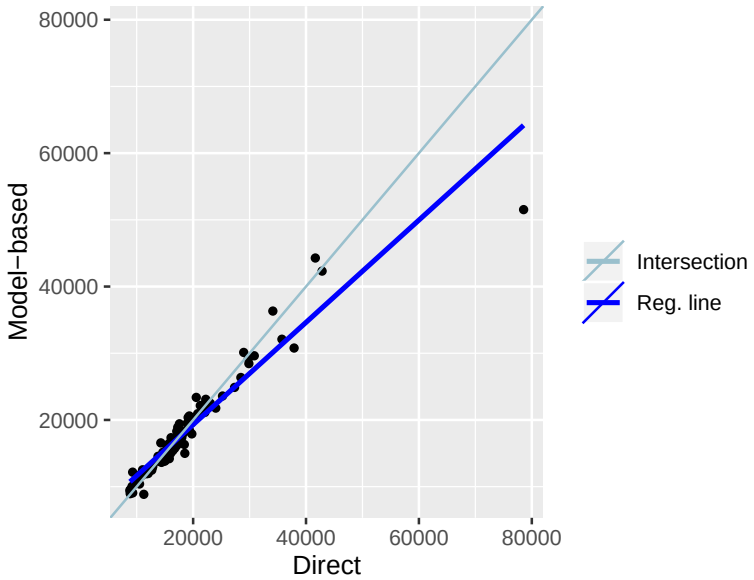
```
> plot(fh_std)
```

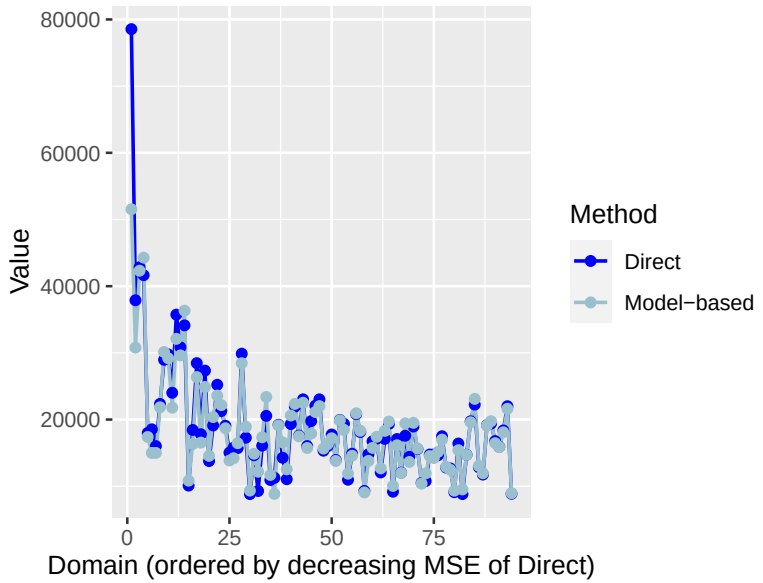




Step 6: Compare results with direct estimates (1)

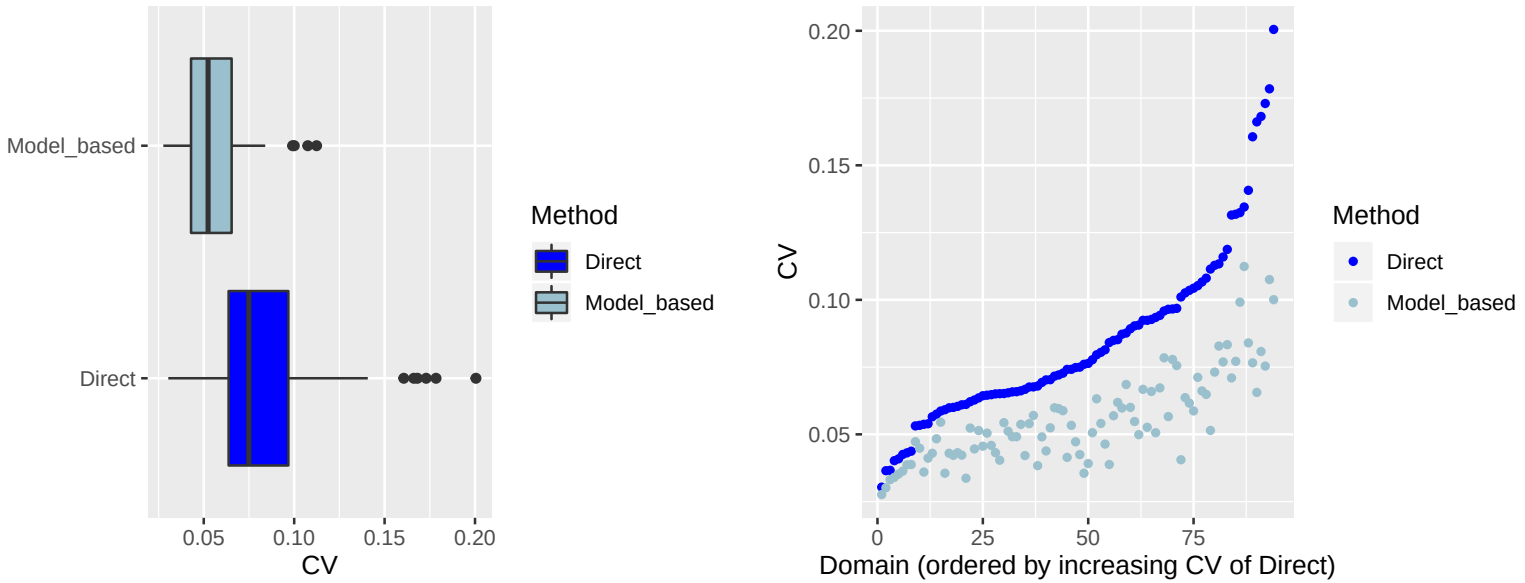
```
> compare_plot(fh_std, CV = TRUE)
```





Step 6: Compare results with direct estimates (2)

```
> compare_plot(fh_std, CV = TRUE)
```

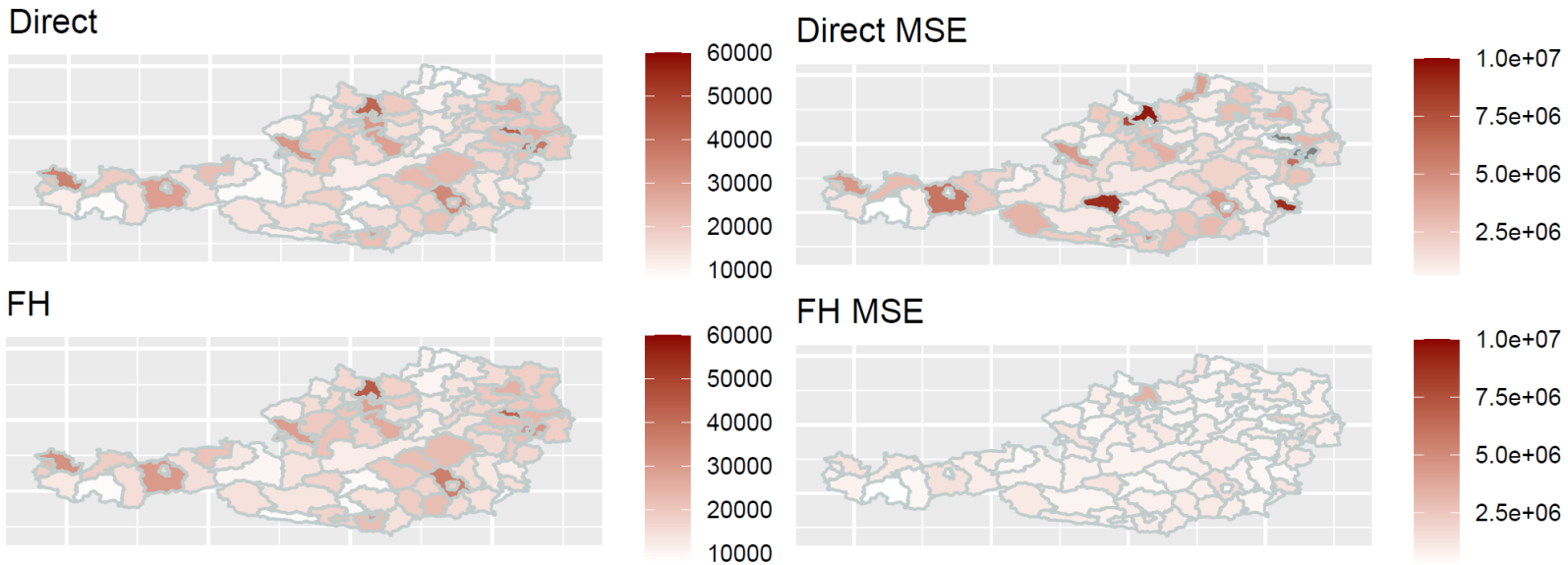


Step 6: Compare results with direct estimates (3)

```
> 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
```

Step 8: Visualize the results

```
1   > load_shapeaustria()
2   > 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))))
```



Step 9: Export the results

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

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

Summary

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
-847,8302926	1703,660585	1703,55124	1715,407312	1703,865937	1713,833764

Estimates

Domain	Direct	Direct_MSE	Direct_CV	FH	FH_MSE	FH_CV
Amstetten	14768,56933	926167,3714	0,065163787	14242,04457	599010,649	0,054343165
Baden	21995,72487	446534,2852	0,030380095	21616,39582	356586,0515	0,027624784
Bludenz	12069,59239	1243265,013	0,092382403	12680,37578	716040,1177	0,066732371

Conclusion

- Variety of area-level models allow the user to address various issues that arise in practical data applications.
- All methods can be estimated conveniently by using a single function that provides FH and MSE estimates.
- User-friendly tools are provided to enable a whole data analysis procedure.
- More information
  - `https://CRAN.R-project.org/package=emdi`
  - Newly added package vignette "A Framework for Producing Small Area Estimates Based on Area-Level Models in R"
  - Github repository emdiExamples (`https://github.com/akreutzmann/emdiExamples`)

References

Fay, R.E. and Herriot, R.A. (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., and Kreutzmann, A.K. (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.

Harmening, S., Kreutzmann, A.K., Pannier, S., Rojas-Perilla, N., Salvati, N., Schmid, T., Templ, M., Tzavidis, N. and Würz, N. (2020). emdi: Estimating and Mapping Disaggregated Indicators.

R package version 2.0.1, URL: https://CRAN.R-project.org/package=emdi

References

Kreutzmann, A.K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. and Tzavidis (2019). 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.

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

Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators.
The Canadian Journal of Statistics 38(3), 369-385.

Petrucci, A. and 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.

References

Slud, E. and 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.

Warnholz, S. (2016). Small Area Estimation Using Robust Extensions to Area Level Models.

Ph.D. thesis, Freie Universität Berlin. doi:10.17169/refubium-13904.

Ybarra, L.M.R. and Lohr, S.L. (2008). Small Area Estimation When Auxiliary Information Is Measured with Error.

Biometrika, 95(4), 919-931. doi:10.1093/biomet/asn048.

Yoshimori, M. and 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.

Thank you for your attention

Sylvia Harmening
sylvia.harmening@fu-berlin.de