The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators

Ann-Kristin Kreutzmann, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Nikos Tzavidis & Matthias Templ

Journal of Statistical Software, 2019

Motivation

- The demand for indicators on a disaggregated level is increasing in order to improve policy decisions
- Maps that combine the estimated indicators with geographical data are in favour for presenting these indicators
- User-friendly software tools can simplify the estimation of these indicators, the assessment of estimations and their visualization

Discussion and Outlook

- The package comprises all steps from estimation, assessment of estimation to presentation via maps and in excel
- It is especially simple to use the provided functions and thus to receive illustrative results
- **Further implementations:** More model-based small area estimation methods, a wider range of transformation methods and parallelization of the bootstrap computation

How the R package emdi supports estimating and mapping disaggregated indicators

Estimation Method

- Direct estimators
- Empirical Best Prediction (EBP)
- World Bank Method (coming soon)

Model Summary

- Description of the data
- Information about the transformation (if used)
- Residual diagnostics and coefficients of determination of the underlying model

Diagnostic Plots

- Q-Q plots of both error levels
- Kernel density estimates of both error levels
- Cooks distance plot for outlier detection
- Illustration of the model log likelihood depending on the transformation parameter

Selection of Indicators

- Select from various predefined or custom indicators
- Easily extract correspondent RMSE or CV

Visualization

- Built in mapping function
- Visualize estimates on a corresponding shape file

Export

- Export the model summary and estimates with their precisions into pre-formatted excel files

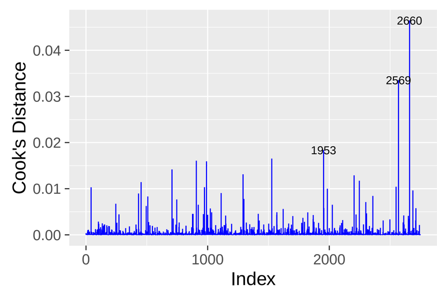
Receive point and MSE/variance estimates

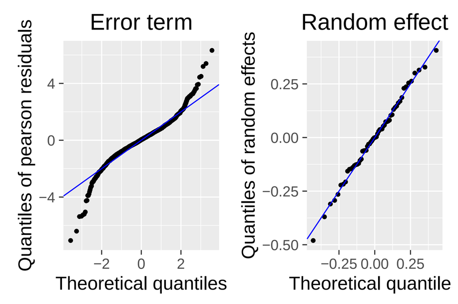
- The direct estimates correspond to the direct estimates in the **laeken** package and thus comprise important poverty and inequality measures used in European and worldwide poverty and social exclusion analysis: Head Count Ratio, Poverty Gap, Gini coefficient and Quintile Share Ratio
  ↪ `head_count()`, `poverty_gap()`, `gini()`, `quintile_share()`
- The implemented model-based small area estimation method is the Empirical Best Prediction (EBP) approach by Molina and Rao (2010). For the EBP, the mentioned poverty and inequality indicators, the mean, and several quantiles (10%, 25%, median, 75%, 90%) are returned. Furthermore, the user can define multiple individual indicators by the argument `custom_indicator`.
  ↪ `ebp()`
- Different transformations can be conducted in order to meet the Gaussian assumptions for model-based estimation methods: no transformation, log-transformation and Box-Cox transformation. For the latter, the optimal parameter is obtained by REML estimation.

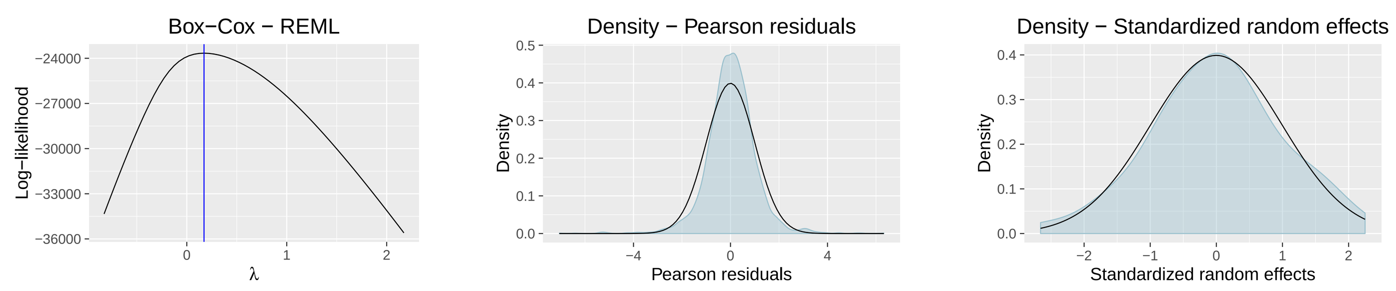
Model diagnostics

Graphical diagnostics contain the four graphs on the right and in case that the Box-Cox transformation is used also the plot below.

```
> plot(ebp)
```







Summarize estimation results

The `summary()` function gives information about the data sets, the data transformation, results of conducted normality checks and explanatory measures. Additionally, a summary especially of the underlying linear mixed model can be received by `summary(emdiObject$model)`.

```
 > summary(ebp)
Empirical Best Prediction

Call:
ebp(fixed = ictpc ~ pcocup + jnived + clase\_hog + pcpering +
  bienes + actcom,
pop_data = census, pop_domains =
"domain_id", smp_data = survey_mex, smp_domains =
"domain_id", pov_line = 903.04, transformation =
"box.cox", L = 50, MSE = T, B = 50, custom_indicator = list(
  my_max = function(y, pov_line) {max(y)}, my_min = function(y,
   pov_line) {min(y)}))

Out-of-sample domains:  67
In-sample domains:  58

Sample sizes:
Units in sample:  2748
Units in population:  219514
                     Min. 1st Qu. Median    Mean 3rd Qu.   Max.
Sample_domains          3      17     21   47.38   42.25    527
Population_domains    650     923   1161 1756.00 1447.00  13580

Transformation:
Transformation Method Optimal_lambda Shift_parameter
    box.cox   reml       0.172312               1

Explanatory measures:
Marginal_R2 Conditional_R2
0.4867911      0.4955212

Residual diagnostics:
                Skewness Kurtosis Shapiro_W    Shapiro_p
Error         -0.2426125 7.951944 0.9500250 1.238856e-29
Random_effect -0.1211658 3.003680 0.9936586 9.906020e-01

ICC:  0.01701086
```

Select indicators

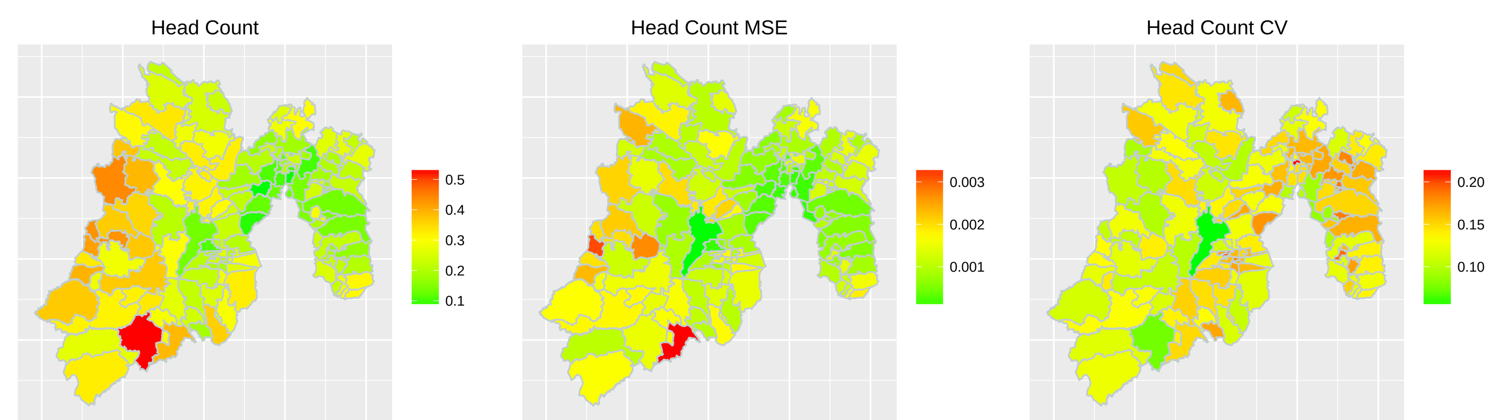
Function `estimators()` enables to select `all` indicators, groups of indicators (`Poverty` and `Inequality`) and each indicator separately.

```
 > estimators(object = ebp, MSE = T, CV = T, indicator = "Poverty")
Indicator/s: Head_Count, Poverty_Gap
                   Domain Head_Count Head_Count_MSE Head_Count_CV Poverty_Gap Poverty_Gap_MSE Poverty_Gap_CV
1                 Acambay 0.33861472   2.136365e-03    0.13649975  0.14101067    6.171988e-04     0.17618160
2                 Acolman 0.18812822   9.045488e-04    0.15986818  0.06795016    1.620265e-04     0.18732794
3                  Aculco 0.25606695   1.785476e-03    0.16501501  0.09855112    4.095351e-04     0.20534494
4  Almoloya de Alquisiras 0.30166667   1.713125e-03    0.13720415  0.12177054    4.566486e-04     0.17548857
5     Almoloya de Juárez 0.21157773   6.459450e-04    0.12012345  0.08261839    1.472854e-04     0.14689370
6       Almoloya del Río 0.18889381   8.946196e-04    0.15834396  0.07017192    1.947474e-04     0.19887149
       ...
```

Map estimation results

Function `map_plot()` combines shape files with estimated indicators. If the domain variables differ between the data set and the shape file a `data_frame` that fits the variables is necessary.

```
map_table <- data.frame(Domain = unique(census$domain_id),
                        mun = sort(shp_mex$mun))
 > map_plot(object = ebp, MSE = T, CV = T, map_obj =
  shp_mex,indicator = "Head_Count", map_dom_id =
  "mun",map_tab = map_table)
```

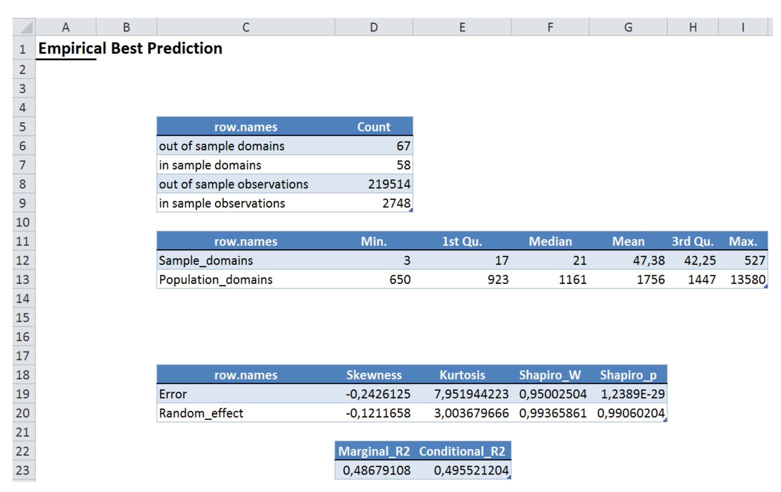


Export results to excel

Function `write_excel()` enables to use the results independently of the statistical software R by exporting results to excel.

```
 > write.excel(ebp, file ="to_excel.xlsx", indicator = "Poverty", MSE = T, CV = T)
```

	A	B	C	D	E	F	G
1	Domain	Head_Count	Head_Count_MSE	Head_Count_CV	Poverty_Gap	Poverty_Gap_MSE	Poverty_Gap_CV
2	Acambay	0,33861472	0,002136365	0,136499749	0,141010672	0,000617199	0,176181601
3	Acolman	0,18812822	0,000904549	0,159868185	0,067950162	0,000162027	0,187327941
4	Aculco	0,25606695	0,001785476	0,165015007	0,098551121	0,000409535	0,205344943
5	Almoloya de Alquisiras	0,30166667	0,001713125	0,137204154	0,121770544	0,000456649	0,175488569
6	Almoloya de Juárez	0,21157773	0,000645945	0,120123454	0,082618387	0,000147285	0,146893695



References

[1] 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.
[2] Molina, I., & Rao, J.N.K. (2010) *Small area estimation of poverty indicators.* Canadian Journal of Statistics, 38(3), 369–385.
[3] Gurka, M. J., Edwards, L. J., Muller, K. E. & Kupper, L. L. (2006) *Extending the Box–Cox Transformation to the Linear Mixed Model.* Journal of the Royal Statistical Society. 26(2), 211–252.

For further information

Ann-Kristin Kreutzmann, Sören Pannier — Freie Universität Berlin

Natalia Rojas-Perilla — United Arab Emirates University

Timo Schmid — Otto-Friedrich-Universität Bamberg

Nikos Tzavidis — University of Southampton

Matthias Templ — Zurich University of Applied Sciences