Small Area Estimation in R with Application to Mexican Income Data

Ann-Kristin Kreutzmann\textsuperscript{1}, Sören Pannier\textsuperscript{1}  
Natalia Rojas-Perilla\textsuperscript{1}, Timo Schmid\textsuperscript{1}  
Matthias Templ\textsuperscript{2}  
Nikos Tzavidis\textsuperscript{3}

\textsuperscript{1} Freie Universität Berlin  
\textsuperscript{2} Zürcher Hochschule für Angewandte Wissenschaften  
\textsuperscript{3} University of Southampton

New Techniques and Technologies for Statistics (NTTS)  
March 16, 2017
Why Using Small Area Estimation

Population of interest (or target population): population for which the survey is designed, here the State of Mexico (EDOMEX)

→ Direct estimators should be reliable for the target population

Gini - Direct

Figure: Gini coefficient for EDOMEX using direct estimation.
Why Using Small Area Estimation

Population of interest: Sub-population of the population (domain), planned or not in the survey design, here municipalities in EDOMEX

Direct estimators may be unreliable due to small sample sizes

Figure: Gini coefficients for municipalities in EDOMEX using direct estimation (left) and a Small Area Estimation (SAE) method (right). Municipalities filled with grey color represent areas with zero sample size.
Combining Different Data Sources

- In order to provide reliable estimates in all subdomains, efficient ways of *combining* information are required.

- Main idea:
  1. Usage of observed/collected data to fit a suitable model
     \[ Y \sim X \]

  2. Produce predictions in all the domains using available covariates

- Types of covariates:
  - Aggregate data provide covariates for every domain
  - Individual data provide covariates for every individual in every domain
R Packages for SAE

CRAN Task View: Official Statistics & Survey Methodology

- \texttt{nlme} and \texttt{lme4} for mixed-effects models
- \texttt{hbsae} for basic area- and unit-level models fitted by restricted maximum likelihood or hierarchical Bayes
- \texttt{rsae} for robust basic unit- and area-level models
- \texttt{JoSAE} for unit-level models and the generalized regression estimator (main purpose: documentation of function used in publications)
R Packages for SAE

Other packages

- **BayesSAE** for area-level models in Bayesian context
- **saeRobust** for robust area level models
- **saery** and **sae2** for area-level models with time effects
- **sae** for a wide variety of SAE methods including area-level and unit-level models for the mean as well as models for the estimation of non-linear parameters
- **emdi** for the estimation and visualization of non-linear indicators
Gini coefficients for municipalities in EDOMEX

How does a statistical institute receive the map of Gini coefficients for the municipalities in EDOMEX?

- We assume that survey and census data is available for EDOMEX
- SAE methods that enable the estimation of non-linear indicators as the Gini coefficient are implemented in the `sae` package and in package `emdi`
- Since package `emdi` supports the whole process from the estimation over model evaluation and visualization functions from this package are used for the following example
The method that is used to receive indicators for the municipalities is the Empirical Best Prediction (EBP) approach by Molina and Rao (2010).

```r
modelFit <- ebp(fixed = ictpc ~ pcocup + jnived + clase_hog + pcpering + bienes + actcom,
                pop_data = census, pop_domains = "domain_id",
                smp_data = survey, smp_domains = "domain_id",
                transformation = "box.cox",
                MSE = TRUE,
                custom_indicator =
                list(my_max = function(y, pov_line){max(y)},
                     my_min = function(y, pov_line){min(y)}),
                na.rm = TRUE)
```
Data and Model Diagnostics (1)

```r
> summary(modelFit)

Empirical Best Prediction

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

Sample sizes:
Units in sample: 2748
Units in population: 219514

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1stQu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rdQu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_domains</td>
<td>3</td>
<td>17</td>
<td>21</td>
<td>47.38</td>
<td>42.25</td>
<td>527</td>
</tr>
<tr>
<td>Population_domains</td>
<td>650</td>
<td>923</td>
<td>1161</td>
<td>1756</td>
<td>1447</td>
<td>13580</td>
</tr>
</tbody>
</table>
```
## Data and Model Diagnostics (2)

<table>
<thead>
<tr>
<th>Explanatory measures:</th>
<th>Marginal_R2</th>
<th>Conditional_R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4867911</td>
<td>0.4955212</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residual diagnostics:</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Shapiro_W</th>
<th>Shapiro_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>-0.2426125</td>
<td>7.951944</td>
<td>0.9500250</td>
<td>1.238856e-29</td>
</tr>
<tr>
<td>Random_effect</td>
<td>-0.1211658</td>
<td>3.003680</td>
<td>0.9936586</td>
<td>9.906020e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC:</td>
<td>0.01701086</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transformation:</th>
<th>Transformation Method</th>
<th>Optimal_lambda</th>
<th>Shift_parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>box.cox</td>
<td>reml</td>
<td>0.172312</td>
</tr>
</tbody>
</table>

Ann-Kristin Kreutzmann 10 (18)
Graphical Diagnostics

1. `plot(modelFit)`

- Error term
- Random effect
- Density – Pearson residuals
- Density – Standardized random effects

Cook's Distance Plot

Box–Cox – REML

Log–Likelihood

Index

\( \lambda \)
Selection of Indicators

- Function `estimators` helps to select the indicators the user is interested in
- Additionally, the coefficient of variation can be received
- The user can choose single indicators or groups of indicators

```r
> head(estimators(object = modelFit, indicator = "Gini", CV = TRUE))
```

<table>
<thead>
<tr>
<th>Domain</th>
<th>Gini</th>
<th>Gini_CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acambay</td>
<td>0.452</td>
<td>0.0253</td>
</tr>
<tr>
<td>Acolman</td>
<td>0.425</td>
<td>0.0191</td>
</tr>
<tr>
<td>Aculco</td>
<td>0.439</td>
<td>0.0218</td>
</tr>
<tr>
<td>Almoloya de Alquisiras</td>
<td>0.462</td>
<td>0.0268</td>
</tr>
<tr>
<td>Almoloya de Juárez</td>
<td>0.476</td>
<td>0.0295</td>
</tr>
<tr>
<td>Almoloya del Río</td>
<td>0.438</td>
<td>0.0309</td>
</tr>
</tbody>
</table>
Visualization

```r
> map_table <- data.frame(Domain = unique(census$domain_id),
                         mun = sort(shp_mex$mun))

> map_plot(object = modelFit, CV = TRUE, map_obj = shp_mex,
          indicator = "Gini", map_dom_id = "mun", map_tab = map_table)
```

![Gini Map](image1)

![Gini CV Map](image2)
Export to excel

1. `> write.excel(modelFit, file = "excel_output.xlsx", indicator = "Gini", CV = TRUE)`
Conclusion

- Official statistics are interested in disaggregated indicators
- SAE methods enable the estimation of disaggregated indicators and several packages in R provide these methods
- Non-linear indicators like the Gini coefficient or the At-risk-of-poverty rate are of special interest
- Package `sae` and package `emdi` enable the estimation of these indicators
- As shown, package `emdi` provides an overall package for the user from the estimation to the visualization and export of results
References


R package version 1.0., URL: https://CRAN.R-project.org/package=hbsae

R package version 0.2.3., URL: https://CRAN.R-project.org/package=JoSAE

R package version 1.0-1., URL: https://CRAN.R-project.org/package=BayesSAE
References

  *R package version 1.0.*, [URL](https://CRAN.R-project.org/package=saery)

  *R package version 0.1-1.*, [URL](https://CRAN.R-project.org/package=sae2)

  *R package version 1.1.0.*

  *The R Journal 7*(1), 81-98.
References


