mme: An R package for small area estimation with multinomial mixed models

M.E. López-Vizcaíno\textsuperscript{1}, M.J. Lombardía\textsuperscript{2} and D. Morales \textsuperscript{3}

\textsuperscript{1}Instituto Galego de Estatística
\textsuperscript{2}Universidade da Coruña
\textsuperscript{3}Universidad Miguel Hernández de Elche
In Spain, like in other European countries, the estimation of some socioeconomic indicators (employed, unemployed, poverty,...) is made by means of surveys that most municipalities and another local areas are not represented in the sample and many of them are present with a very small sample size.

In this situation the sample size could be enlarged: delays in obtaining results and the impact of non-sampling errors.

The increase of the sample size is not always advisable and even sometimes unfeasible from an economic point of view.

Frequently, auxiliary variables exists that are correlated with the variable of interest and several estimators can make use of auxiliary information.

This situation may be treated by using small area estimation techniques.
Small area estimation techniques can be divided into design-based methods and model-based methods. The model-based methods make inference by taking into account the underlying model. The estimators based on these methods give to practitioners an idea of the data generation process.

Mixed models are suitable for small area estimation due its flexibility to make an effective combination of different sources of information and its capacity to describe the various sources of error.

These models incorporate random area effects that explain the additional variability that is not explained by the fixed part of the model.
Packages in R for small area estimation.

- **JoSAE.** Implements the unit level EBLUP and GREG estimators (11/10/2011)
- **rsae.** Computes robust basic unit-level and area-level SAE models (8/01/2012).
- **hbsae** Small area estimation based on the basic unit-level and area-level models. The small area estimates are computed in a hierarchical Bayesian way (5/09/2012).
- **sae.** Model based estimators include EBLUP based on a Fay-Herriot model and the EBLUP based on a unit level nested error model
- **BayesSAE.** Provides a variety of functions to deal with several specific small area- level models in Bayesian context (28/10/2013).
- **maSAE.** an S4 implementation of the unbiased extension of the model-assisted synthetic-regression estimator proposed by Mandallaz (2013), Mandallaz et al. (2013) and Mandallaz (2014) (28/04/2014).
Objective

Present an R package that implements four multinomial area level mixed models for small area estimation.

- **Model 0** is based on the area level multinomial mixed model with common random effects for the categories of the response variable (Molina et al (2007))\(^1\).

- **Model 1** is based on the area level multinomial mixed model with independent random effects for each category of the response variable (López-Vizcaíno et al, 2013)\(^2\).

---


Objective

Present an R package that implements four multinomial area level mixed models for small area estimation.

- **Model 2** take advantage from the availability of survey data from different time periods and use a multinomial model with independent random effects for each category of the response variable and with independent time and domain random effects (López-Vizcaíno et al, 2014).

- **Model 3** is similar to Model 2, but with correlated time and domain random effects (López-Vizcaíno et al, 2014).

---

The multinomial logit mixed model

- Use indexes $k = 1, \ldots, (q - 1)$; $d = 1, \ldots, D$ and $t = 1, \ldots, T$ for the categories of the target variable, for the $D$ domains and for the $T$ periods of time.

- Let $u_{1,dk}$ and $u_{2,dtk}$ be the random effects associated to category $k$, domain $d$ in time $t$, independents.

- We assume that the response vectors $y_{dt}$ conditioned to $u_{1,d}$ and $u_{2,dt}$, are independents with multinomial distributions

$$y_{dt} \mid u_{1,d}, u_{2,dt} \sim \mathcal{M}(\nu_{dt}, p_{1,dt}, \ldots, p_{q-1,dt}), \quad d = 1, \ldots, D, \quad t = 1, \ldots, T.$$
The multinomial logit mixed model

The model

\[ \eta_{dtk} = \log \frac{p_{dtk}}{p_{dtq}} = x_{dtk} \beta_k + u_{1,dk} + u_{2,dtk} \]

where

- \( d = 1, \ldots, D \), \( t = 1, \ldots, T \), \( k = 1, \ldots, q - 1 \)
- \( x_{dtk} = \text{col}'_{1 \leq r \leq l_k} (x_{dtkr}) \)
- \( \beta_k = \text{col}_{1 \leq r \leq l_k} (\beta_{kr}) \)

\[ p_{dtk} = \frac{\exp\{\eta_{dtk}\}}{1 + \sum_{\ell=1}^{q-1} \exp\{\eta_{dt\ell}\}} \]

is the probability of the multinomial category \( k \) in the domain \( d \) and in the time \( t \).
The multinomial logit mixed model

- **Model 0** is the model for one time period, $u_{2,dtk} = 0$ and common random effects $u_{1,dtk} = u_d$.

- **Model 1** is the model for one time period and independent random effects, $u_{2,dtk} = 0$ and $u_{1,dtk} = u_{dk}$.

- **Model 2** Include independent time and domain random effects.

- **Model 3** Include correlated time and domain random effects (AR(1)).
Model-based small area estimation

The problem is to estimate the domain total

\[ m_{dt} = \hat{N}_{dt} p_{dt}, \]

- \( d=1,\ldots,D; \ t=1,\ldots,T. \)

- \( \hat{N}_d \) is an estimated population size that can be obtained from the unit-level survey data. In the application to real data, we take \( \hat{N}_{dt} = \hat{N}^{dir}_{dt} \)

- We estimate \( m_{dt} \) by means of

\[ \hat{m}_{dt} = \hat{N}_{dt} \hat{p}_{dt} \]
MSE estimation

**Estimator 1:** Analytical estimation (Prasad and Rao (1990))\(^1\).

We approximate the mean square error of \( \hat{m}_d \) by means of

\[
\text{MSE}(\hat{m}_{dtk}) \approx \mathcal{G}_1(\sigma) + \mathcal{G}_2(\sigma) + \mathcal{G}_3(\sigma),
\]

The proposed analytic mean square error estimator

\[
\text{mse}(\hat{m}_{dtk}) = \mathcal{G}_1(\hat{\sigma}) + \mathcal{G}_2(\hat{\sigma}) + 2\mathcal{G}_3(\hat{\sigma}).
\]

**Estimator 2:** Parametric bootstrap (González-Manteiga et al. (2008))\(^2\).

\[
\text{mse}^*_{dtk} = \frac{1}{B} \sum_{b=1}^{B} (\hat{m}^*_{dtk} - m^*_{dtk})^2.
\]

---


# mme package. Principal functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>data.mme</td>
<td>Based on the input data this function generates some matrices that are required in subsequent calculations and the initial values for the fitting algorithm.</td>
</tr>
<tr>
<td>fitmodel0</td>
<td>Function used to fit the Model 0</td>
</tr>
<tr>
<td>fitmodel1</td>
<td>Function used to fit the Model 1</td>
</tr>
<tr>
<td>fitmodel2</td>
<td>Function used to fit Model 2</td>
</tr>
<tr>
<td>fitmodel3</td>
<td>Function used to fit Model 3</td>
</tr>
<tr>
<td>msef</td>
<td>This function is used to calculate the analytic MSE for Model 1</td>
</tr>
<tr>
<td>msef.it</td>
<td>This function is used to calculate the analytic MSE for Model 2</td>
</tr>
<tr>
<td>msef.ct</td>
<td>This function is used to calculate the analytic MSE for Model 3</td>
</tr>
<tr>
<td>mseb</td>
<td>Function used to calculate the bias and the MSE for the multinomial mixed effects models using parametric bootstrap</td>
</tr>
</tbody>
</table>
mme package. Input data

- **k**: number of categories of the response variable
- **pp**: vector with the number of the auxiliary variables per category
- **mod**: model 0,1,2, or 3

**Input data set**: area-time-sample size-population size-response variable-auxiliary variables

<table>
<thead>
<tr>
<th>county</th>
<th>time</th>
<th>n</th>
<th>N</th>
<th>emp</th>
<th>unemp</th>
<th>inac</th>
<th>SS</th>
<th>REG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>23</td>
<td>3969.77</td>
<td>15</td>
<td>3</td>
<td>5</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>23</td>
<td>3989.08</td>
<td>15</td>
<td>3</td>
<td>5</td>
<td>0.47</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>23</td>
<td>3892.38</td>
<td>16</td>
<td>2</td>
<td>5</td>
<td>0.47</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>23</td>
<td>3940.28</td>
<td>16</td>
<td>1</td>
<td>6</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>21</td>
<td>3537.07</td>
<td>15</td>
<td>1</td>
<td>5</td>
<td>0.45</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>20</td>
<td>3477.65</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0.33</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>21</td>
<td>3690.75</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>0.34</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>21</td>
<td>3612.63</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>0.35</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>20</td>
<td>3462.78</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0.34</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>3382.49</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>0.34</td>
<td>0.07</td>
</tr>
</tbody>
</table>
mme package. Syntax

> k=3 #number of categories of the response variable
> pp=c(1,1) #vector with the number of auxiliary variables
> mod=3 #Model 3

> data=read.csv2("data.csv")
> datar=data.mme(data,k,pp,mod)

> #Model fit
> result=model(datar$d,datar$t,pp,datar$Xk,datar$X,datar$Z,
> datar$y[,1:(k-1)],datar$n,datar$N,mod)

> #Bootstrap parametric MSE
> B=500 #Bootstrap iterations
> mse.pboot=mseb(pp,datar$Xk,datar$X,datar$Z,datar$n,
> datar$N,result,B,mod)
Objective

Estimate the total of employed and unemployed people and the unemployment rates per sex in the counties of Galicia.
**Sample information:**
SLFS of Galicia from the **third quarter of 2009 to the fourth quarter of 2011**. Information available on individual level.

**Domains of interest:**
The 51 counties of Galicia crossed with sex for each period, \( D = 102 \) domains \( P_{dt} \) partitioned in the subsets

\[
\begin{array}{|c|c|}
\hline
P_{dt1} & \text{employed} \\
P_{dt2} & \text{unemployed} \\
P_{dt3} & \text{inactive} \\
\hline
\end{array}
\]

**Target population parameters:**
The totals of employed and unemployed people and the unemployed rate.

\[ Y_{dtk} = \sum_{j \in P_{dt}} y_{dtkj}, \quad R_{dt} = \frac{Y_{d2}}{Y_{dt1} + Y_{dt2}}, \quad k = 1, 2 \]
**Data description**

**Auxiliary variables**

- **SEXAGE**: Combinations of sex and age groups, with 6 values. SEX is coded 1 for men and 2 for women and AGE is categorized in 3 groups with codes 1 for 16-24, 2 for 25-54 and 3 for $\geq 55$. The codes 1, 2, ..., 6 are used for the pairs of *sex-age* (1, 1), (1, 2), ..., (2, 3).

- **STUD**: This variable describes the achieved education level, with values 1-3 for the illiterate and the primary, the secondary and the higher education level respectively.

- **SS**: This variable indicates if an individual is registered or not in the national insurance contribution system.

- **REG**: This variable indicates if an individual is registered or not as unemployed in the administrative register of employment claimants.
Data description

Figure: Log-rates of employed and unemployed over inactive people versus proportions of people in the national insurance contribution system (left) and registered as unemployed (right), respectively.
In the fourth quarter of 2011 the distribution of the sample sizes per domains in the SLFS of Galicia has the quantiles

\[ q_{\text{min}} = 13, q_1 = 54, q_2 = 97, q_3 = 153, q_{\text{max}} = 1554. \]

This means that the direct estimators are not reliable.
Application to real data

Data.

<table>
<thead>
<tr>
<th>countie</th>
<th>time</th>
<th>n</th>
<th>N</th>
<th>emp</th>
<th>unemp</th>
<th>inac</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>17</td>
<td>2844.76</td>
<td>11</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>17</td>
<td>3057.11</td>
<td>9</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>17</td>
<td>2803.11</td>
<td>9</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>17</td>
<td>2824.67</td>
<td>10</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>17</td>
<td>3004.77</td>
<td>9</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>23</td>
<td>3969.77</td>
<td>15</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>employed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEXAGE1</td>
</tr>
<tr>
<td>0.10</td>
</tr>
<tr>
<td>0.10</td>
</tr>
<tr>
<td>0.10</td>
</tr>
<tr>
<td>0.10</td>
</tr>
<tr>
<td>0.10</td>
</tr>
<tr>
<td>0.12</td>
</tr>
</tbody>
</table>
Application to real data. Cont

Data.

<table>
<thead>
<tr>
<th>SEXAGE1</th>
<th>SEXAGE2</th>
<th>SEXAGE3</th>
<th>SEXAGE4</th>
<th>SEXAGE5</th>
<th>STUD1</th>
<th>REG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.63</td>
<td>0.27</td>
<td>0.00</td>
<td>0.00</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>0.10</td>
<td>0.64</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>0.10</td>
<td>0.63</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>0.10</td>
<td>0.63</td>
<td>0.27</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>0.10</td>
<td>0.64</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>0.12</td>
<td>0.57</td>
<td>0.31</td>
<td>0.00</td>
<td>0.00</td>
<td>0.15</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Application to real data

mme

> k=3  #number of categories of the response variable
> pp=c(7,7)  #vector with the number of auxiliary variables
> mod=3  #Model 3
>
> datos=read.csv2("datos.csv")
> datar=data.mme(datos,k,pp,mod)
>
> #Model fit
> result=model(datar$d,datar$t,pp,datar$Xk,datar$X,datar$Z,
> datar$y[,1:(k-1)],datar$n,datar$N,mod)
>
> names(result)
[1] "Estimated.probabilities" "u1" "u2" "mean"
[5] "warning1" "Fisher.information.matrix.beta"
[9] "phi.Stddev.p.value" "warning2" "rho"
[10]"rho.Stddev.p.value"
## Application to real data

**mme. Fixed effects**

```r
> result
Multinomial mixed effects model Call: Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.Error</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.457</td>
<td>0.200</td>
<td>0.000000</td>
</tr>
<tr>
<td>SEXAGE1</td>
<td>1.029</td>
<td>0.383</td>
<td>0.007262</td>
</tr>
<tr>
<td>SEXAGE2</td>
<td>2.078</td>
<td>0.214</td>
<td>0.000000</td>
</tr>
<tr>
<td>SEXAGE3</td>
<td>0.145</td>
<td>0.167</td>
<td>0.385993</td>
</tr>
<tr>
<td>SEXAGE4</td>
<td>0.459</td>
<td>0.423</td>
<td>0.277387</td>
</tr>
<tr>
<td>SEXAGE5</td>
<td>1.716</td>
<td>0.271</td>
<td>0.000000</td>
</tr>
<tr>
<td>STUD1</td>
<td>-0.778</td>
<td>0.168</td>
<td>0.000004</td>
</tr>
<tr>
<td>SS</td>
<td>1.467</td>
<td>0.259</td>
<td>0.000000</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.684</td>
<td>0.282</td>
<td>0.000000</td>
</tr>
<tr>
<td>SEXAGE1</td>
<td>1.965</td>
<td>0.652</td>
<td>0.002583</td>
</tr>
<tr>
<td>SEXAGE2</td>
<td>2.136</td>
<td>0.340</td>
<td>0.000000</td>
</tr>
<tr>
<td>SEXAGE3</td>
<td>-0.484</td>
<td>0.241</td>
<td>0.044500</td>
</tr>
<tr>
<td>SEXAGE4</td>
<td>1.091</td>
<td>0.760</td>
<td>0.151094</td>
</tr>
<tr>
<td>SEXAGE5</td>
<td>1.357</td>
<td>0.413</td>
<td>0.001018</td>
</tr>
<tr>
<td>STUD1</td>
<td>-0.565</td>
<td>0.254</td>
<td>0.026366</td>
</tr>
<tr>
<td>REG</td>
<td>11.420</td>
<td>0.800</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
```
Application to real data

```
> result
Random effects
       Estimate  Std.dev   p.value
[1,]  0.0242  0.00521 3.52e-06
[2,]  0.0810  0.01748 3.54e-06
[3,]  0.0126  0.00194 0.00e+00
[4,]  0.0970  0.01030 0.00e+00

Correlation random effects
       Estimate  Std.Error   p.value
[1,]  0.581    0.0992 0.000000
[2,]  0.290    0.0837 0.000533
>
```
Application to real data

Direct and model based estimators. Women

```r
> result$mean
> dir=(cbind(datos$ocu, datos$par)/datos$n)*datos$N
```

![Plot showing direct estimator versus model-based estimator for employed and unemployed groups.](image)

Figure: Direct estimator versus model-based estimator.
Figure: Direct and model-based estimates of totals of employed men (left) and women (right) for counties with small sample size in the fourth quarter of 2011.
Application to real data

Figure: Direct and model-based estimates of unemployment rates for men (left) and women (right) for counties with small sample size in the fourth quarter of 2011.
Figure: Model 3 estimates of men unemployment rates in Galician counties in IV/2011 (left) and of variations between men unemployment rates from IV/2009 to IV/2011 (right).
Figure: Model 3 estimates of women unemployment rates in Galician counties in IV/2011 (left) and of variations between women unemployment rates from IV/2009 to IV/2011 (right).
RMSE. Parametric bootstrap

```r
> B=500 #Bootstrap iterations
> mod=3
> mse.pboot=mseb(pp,datar$Xk,datar$X,datar$Z,datar$n,
> datar$N,result,B,mod)
>
> names(mse.pboot)
[1] "bias.pboot" "mse.pboot" "rmse.pboot"
>
> mse.pboot$rmse.pboot
```
Application to real data

Figure: RMSE.

RMSE employed women – IV/2011

RMSE unemployment rate women – IV/2011
Application to real data

Figure: Significative variations between men (left) and women (right) unemployment rates from IV/2009 to IV/2011.
The mme package is designed to make life easier for people who work with multinomial models in small area estimation.

The inclusion of time effects allows to obtain estimates in a more accurate and stable form.

This package include bootstrap estimators as a good alternative to the Prasad-Rao methodology.

The obtained model-based estimates for the models are compared with the direct ones. They have lower mean squared errors.
THIS IS ALL.......THANK YOU!!!