mme: a package for small area estimation with multinomial mixed models

E López-Vizcaíno, M J Lombardía Cortiña, D Morales González

Abstract

The **mme** package for R (R Development Core Team 2010) implements three multinomial area level mixed models for small area estimation. The first model is the area level multinomial mixed model with independent random effects for each category of the response variable (López-Vizcaíno, Lombardía, and Morales 2013a). The second model takes advantage from the availability of survey data from different time periods and uses a multinomial model with independent random effects for each category of the response variable and with independent time and domain random effects. The third model is similar to the second one, but with correlated time and domain random effects (López-Vizcaíno, Lombardía, and Morales 2013b). In all the models the package uses two approaches to estimate the mean square error (MSE), first through an analytical expression and second by bootstrap techniques.

Keywords: small area, R package, multinomial mixed models..

1. Overview

Small area estimation problems appear when the domain simple sizes are small and direct estimates are not precise. In the small area estimation context, an estimator of a parameter in a given domain is direct if it is based only on the sample data of the specific domain. A drawback of these estimators is that they cannot be calculated when there is no sample observations in an area of interest.

Generally 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 are useful because they give to practitioners an idea of how the data generation process is and how the different sources of information are incorporated. Mixed models are suitable for small area estimation due to its flexibility to make an effective combination of different sources of information and to 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.

The objective of this manuscript is to present a R package that implements three multinomial area level mixed models for small area estimation. The first model is the area level multinomial mixed model with independent random effects for each category of the response variable (López-Vizcaíno *et al.* 2013a). The second model takes advantage from the availability of survey data from different time periods and uses a multinomial model with independent random effects for each category of the response variable and with independent time and

domain random effects. The third model is similar to the second one, but with correlated time and domain random effects (López-Vizcaíno *et al.* 2013b). In all the models the package use two approaches to estimate the mean square error (MSE), first through an analytical expression and second by bootstrap techniques.

2. Models

Let us start by giving some notation and assumptions. Let us 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 time periods respectively. Let $u_{1,dk}$ and $u_{2,dkt}$ be the random effects associated to the domain d and the category k and to the domain d, the category k and the time instant t respectively. In the third model (Model 3) we write the random effects in the form

$$\boldsymbol{u}_1 = \operatorname*{col}_{1 \le d \le D}(\boldsymbol{u}_{1,d}), \quad \boldsymbol{u}_{1,d} = \operatorname*{col}_{1 \le k \le q-1}(u_{1,dk}), \quad \boldsymbol{u}_2 = \operatorname*{col}_{1 \le d \le D}(\boldsymbol{u}_{2,d})$$

$$\boldsymbol{u}_{2,d} = \operatorname*{col}_{1 \le k \le q-1}(\boldsymbol{u}_{2,dk}), \quad \boldsymbol{u}_{2,dk} = \operatorname*{col}_{1 \le t \le T}(u_{2,dkt}), \quad \boldsymbol{u}_{2,dt} = \operatorname*{col}_{1 \le k \le q-1}(u_{2,dkt}),$$

and we suppose that

1. $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are independent,

2. $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{V}_{u_1})$, where $\boldsymbol{V}_{u_1} = \operatorname*{diag}_{1 \le d \le D}(\operatorname*{diag}_{1 \le k \le q-1}(\varphi_{1k}))$, $k = 1, \ldots, q-1$.

3. $\boldsymbol{u}_{2,dk} \sim N(\boldsymbol{0}, \boldsymbol{V}_{u_{2,dk}})$, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$, are independent with covariance matrix AR(1), i.e. $\boldsymbol{V}_{u_{2,dk}} = \varphi_{2k}\Omega_d(\phi_k)$ and

$$\Omega_d(\phi_k) = \Omega_{d,k} = \frac{1}{1 - \phi_k^2} \begin{pmatrix} 1 & \phi_k & \cdots & \phi_k^{T-2} & \phi_k^{T-1} \\ \phi_k & 1 & \ddots & & \phi_k^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \phi_k^{T-2} & & \ddots & 1 & \phi_k \\ \phi_k^{T-1} & \phi_k^{T-2} & \cdots & \phi_k & 1 \end{pmatrix}_{T \times T}.$$

It holds that $\boldsymbol{V}_u = \operatorname{var}(\boldsymbol{u}) = \operatorname{diag}(\boldsymbol{V}_{u_1}, \boldsymbol{V}_{u_2})$, where $\boldsymbol{V}_{u_2} = \operatorname{var}(\boldsymbol{u}_2) = \operatorname*{diag}_{1 \le d \le D}(\operatorname*{diag}_{1 \le k \le q-1}(\boldsymbol{V}_{u_{2,dk}}))$. We also assume that the response vectors $\boldsymbol{y}_{dt} = \operatorname*{col}_{1 \le k \le q-1}(y_{dkt})$, conditioned to $\boldsymbol{u}_{1,d}$ and $\boldsymbol{u}_{2,dt}$, are independent with multinomial distributions

$$\boldsymbol{y}_{dt}|\boldsymbol{u}_{1,d}, \boldsymbol{u}_{2,dt} \sim M(\nu_{dt}, p_{d1t}, \ldots, p_{dq-1t}), \ d = 1, \ldots, D, \ t = 1, \ldots, T. \tag{1}$$

where the ν_{dt}'s are known integer numbers. The covariance matrix of $\boldsymbol{y}_{dt}$ conditioned to $\boldsymbol{u}_{1,d}$ and $\boldsymbol{u}_{2,dt}$ is $\operatorname{var}(\boldsymbol{y}_{dt}|\boldsymbol{u}_{1,d}, \boldsymbol{u}_{2,dt}) = \boldsymbol{W}_{dt} = \nu_{dt}[\operatorname{diag}(\boldsymbol{p}_{dt}) - \boldsymbol{p}_{dt}\boldsymbol{p}'_{dt}]$, where $\boldsymbol{p}_{dt} = \operatorname*{col}_{1 \le k \le q-1}(p_{dkt})$ and $\operatorname{diag}(\boldsymbol{p}_{dt}) = \operatorname*{diag}_{1 \le k \le q-1}(p_{dkt})$. For the natural parameters $\eta_{dkt} = \log \frac{p_{dkt}}{p_{dqt}}$, we assume the model

$$\eta_{dkt} = \boldsymbol{x}_{dkt}\boldsymbol{\beta}_k + u_{1,dk} + u_{2,dkt}, \quad d = 1, \ldots, D, \ k = 1, \ldots, q-1, \ t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{x}_{dkt} = \operatorname*{col'}_{1 \leq r \leq p_r}(x_{dktr})$, $\boldsymbol{\beta}_k = \operatorname*{col}_{1 \leq r \leq p_k}(\beta_{kr})$ and $p = \sum_{k=1}^{q-1} p_k$.

We also consider two simpler models. Model 2 is the restriction of Model 3 to $\phi_1 = \ldots = \phi_{q-1} = 0$. Model 1 is obtained by restricting Model 2 to one time period ($T = 1$) and by considering only the random effect $\boldsymbol{u}_1$. This is the model studied by López-Vizcaíno *et al.* (2013a). For the sake of brevity we skip formulas for Models 1-2. In matrix notation, Model 3 is

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_1', \boldsymbol{Z}_2')'$, $\boldsymbol{\eta} = \operatorname*{col}_{1 \leq d \leq D}(\boldsymbol{\eta}_d)$, $\boldsymbol{X} = \operatorname*{col}_{1 \leq d \leq D}(\boldsymbol{X}_d)$, $\boldsymbol{Z}_1 = \operatorname*{diag}_{1 \leq d \leq D}(\boldsymbol{Z}_{1d})$, $\boldsymbol{Z}_2 = \operatorname*{diag}_{1 \leq d \leq D}(\boldsymbol{Z}_{2d})$,

$$\boldsymbol{\eta}_d = \operatorname*{col}_{1 \leq k \leq q-1}(\operatorname*{col}_{1 \leq t \leq T}(\eta_{dkt})), \quad \boldsymbol{X}_d = \operatorname*{diag}_{1 \leq k \leq q-1}(\operatorname*{col}_{1 \leq t \leq T}(\boldsymbol{x}_{dkt})), \quad \boldsymbol{\beta} = \operatorname*{col}_{1 \leq k \leq q-1}(\boldsymbol{\beta}_k),$$

$$\boldsymbol{Z}_{1d} = \operatorname*{diag}_{1 \leq k \leq q-1}(\boldsymbol{1}_T), \quad \boldsymbol{Z}_{2d} = \operatorname*{diag}_{1 \leq k \leq q-1}(\operatorname*{diag}_{1 \leq t \leq T}(1)) = \boldsymbol{I}_{T(q-1)}, \quad \boldsymbol{1}_T = \operatorname*{col}_{1 \leq t \leq T}(1).$$

To fit the model we combine the PQL method, introduced by Breslow and Clayton (1996) for estimating and predicting the β_{kr}'s, the $u_{1,dk}$'s and the $u_{2,dkt}$'s, with the REML method for estimating the variance components φ_{1k}, φ_{2k} and ϕ_k, $k = 1, \ldots, q - 1$. The presented method is based on a normal approximation to the joint probability distribution of the vector $(\boldsymbol{y}, \boldsymbol{u})$. The combined algorithm was first introduced by Schall (1991) and later used by Saei and Chambers (2003), Molina, Saei, and Lombardía (2007) and Herrador, Morales, Esteban, Sánchez, Santamaría, Marhuenda, Pérez, and Molina (2009) in applications of generalized linear mixed models to small area estimation problems. We adapt the combined algorithm to Model 3. The algorithm has two parts. In the first part the algorithm updates the values of $\boldsymbol{\beta}$, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. In the second part it updates the variance components.

For the estimation of the mean squared error (MSE) of model-based small area estimators we adapt the resampling approaches appearing in González-Manteiga, Lombardía, Molina, Morales, and Santamaría (2008) to introduce a parametric bootstrap procedure. We also give an approximation to the MSE based on a Taylor linearization. By applying the ideas of (Prasad and Rao 1990) to the linearized model, the MSE is approximated and an estimator of the given approximation is derived.

3. The package mme

In the mme package we introduced a range of new functions that may be of interest to those conducting applied research. The nine principal new functions are summarized in Table 1.

In what follows we provide illustrative examples of the use of the functions describe in Table 1. Many of these functions rely on numerical integration and can be computationally demanding.

4. Example to fit model 1

The following code provides and example to fit the model 1. It is necessary to use a data frame with this variables: area indicator, time indicator, sample, population, categories of the response variable and covariates of each category of the response variable. The package requires two imput parameters: *pp* is a vector with the number of auxiliary variables in each

Function	Description	Reference
data.mme	Based on the input data this function generates some matrices that are required in subsequent calculations and the initial values for the fitting algorithm	López-Vizcaíno *et al.* (2013a)
fitmodel1	This function fits the multinomial mixed model with one independent random effect per category of the response variable (Model 1)	López-Vizcaíno *et al.* (2013a)
fitmodel2	This function fits the multinomial mixed model with two independent random effects for each category of the response variable: one domain random effect and another independent time and domain random effect (Model 2)	López-Vizcaíno *et al.* (2013b)
fitmodel3	This function fits the multinomial mixed model with two independent random effects for each category of the response variable: one domain random effect and another correlated time and domain random effect (Model 3)	López-Vizcaíno *et al.* (2013b)
model	This function chooses one of the three models	López-Vizcaíno *et al.* (2013a) and López-Vizcaíno *et al.* (2013b)
msef	This function calculates the analytic MSE for Model 1	López-Vizcaíno *et al.* (2013a)
msef.it	This function calculates the analytic MSE for Model 2	López-Vizcaíno *et al.* (2013a)
msef.ct	This function calculates the analytic MSE for Model 3	López-Vizcaíno *et al.* (2013b)
mseb	This function calculates the bias and the MSE for the multinomial mixed effects models using parametric bootstrap	López-Vizcaíno *et al.* (2013a) and López-Vizcaíno *et al.* (2013b)

Table 1: New mme functions.

category and k is the number of categories of the response variable. The example uses a data frame with 50 small areas and with 10 periods. However, this example only works with the last period. The response variable has three categories ($k = 3$), and we use one covariate for each category, then $pp = c(1, 1)$. The last three columns of the data frame contain the direct estimators of the categories of the response variable.

```
R> library(mme)
R> datos=as.data.frame(datos)
R> names(datos)


 [1] "area"        "time"        "sample"      "population" "y1"
 [6] "y2"          "y3"          "x1"          "x2"          "y11"
[11] "y22"         "y33"


R> datos1=subset(datos,datos$time==10)
R> dat=datos1[,1:9]
R> k=3 #number of categories of the response variable
```

```
R> pp=c(1,1) #vector with the number of auxiliary variables in each category
R> mod=1 #Model 1
R> #Needed matrix and initial values
R> datar=data.mme(dat,k,pp,mod)
R> #Model fit
R> result=model(datar$d,datar$t,pp,datar$Xk,datar$X,datar$Z,datar$initial,
+ datar$y[,1:(k-1)],datar$n,datar$N,mod)
R> result

Multinomial mixed effects model

Call:

Coefficients
          Estimate Std.Error p.value
Intercept    2.523      2.24   0.259
x1          -1.895      1.66   0.253
Intercept   -0.695      3.03   0.819
x2           0.455      1.75   0.796


Random effects
     Estimate Std.Error p.value
[1,]     1.05     0.241       0
[2,]     2.56     0.557       0

R> #Fixed effects
R> result$beta.Stddev.p.value


          Estimate Std.Error p.value
Intercept    2.523      2.24   0.259
x1          -1.895      1.66   0.253
Intercept   -0.695      3.03   0.819
x2           0.455      1.76   0.796

R> #Random effects
R> result$phi.Stddev.p.value


     Estimate Std.Error  p.value
[1,]     1.05     0.241 1.40e-05
[2,]     2.56     0.557 4.22e-06

R> #Direct estimators
R> dir1=datos1$y11
R> dir2=datos1$y22
R>
```

The following code will generate Figure 1 that plots direct estimators versus model estimators.

```
R> #Plot direct estimator versus model estimator
R> dos.ver<-matrix(1:2,1,2)
R> layout(dos.ver)
R> plot(log(dir1),log(result$mean[,1]),main="Small area estimator Y1",
+ xlab="Direct estimate", ylab="model estimate",font.main=2,cex.main=1.5,
+ cex.lab=1.3)
R> abline(a=0,b=1)
R> plot(log(dir2),log(result$mean[,2]),main="Small area estimator Y2",
+ xlab="Direct estimate", ylab="model estimate",font.main=2,cex.main=1.5,
+ cex.lab=1.3)
R> abline(a=0,b=1)
R>
```
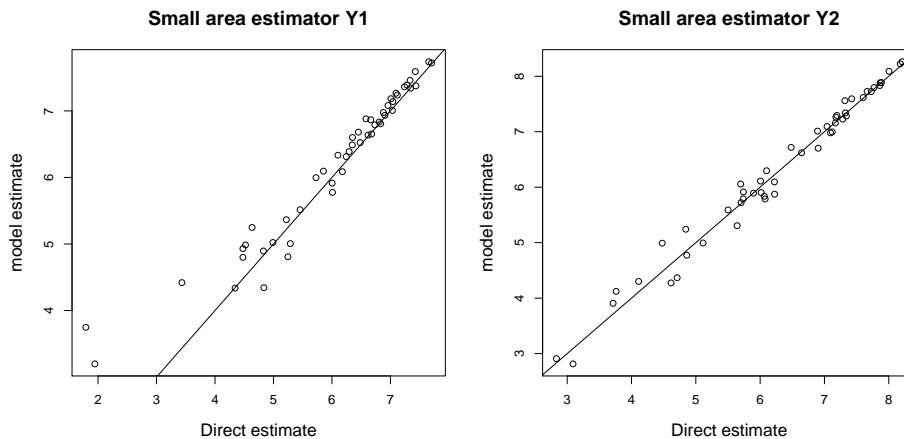

Figure 1: Model estimates versus direct estimates.

```
R> #Model estimator
R> datos1$yest1=result$mean[,1]
R> datos1$yest2=result$mean[,2]
```

The following code generates Figure 2 that plots direct estimators and model estimators sorted by sample size.

```
R> #Plot direct estimators and model estimators sorted by sample size
R> dos.ver<-matrix(1:2,1,2)
R> layout(dos.ver)
R> a=datos1[order(datos1[,3]),]
R> g_range <- range(0,45)
R> plot(a$y11/1000,type="b", col="blue",axes=FALSE, ann=FALSE)
R> lines(a$yest1/1000,type="b",pch=4, lty=2, col="red")
R> title(xlab="Sample size")
R> axis(1,at=c(1,10,20,30,40,50),lab=c(a$sample[1],a$sample[10],
```

```
+ a$sample[20],a$sample[30],a$sample[40],a$sample[50]))
R> axis(2, las=1, at=1*0:g_range[2])
R> legend("topleft", c("Direct","Model"), cex=1, col=c("blue","red"),
+    lty=1:2,pch=c(1,4), bty="n")
R> title(main="Small area estimator Y1", font.main=1.2,cex.main=1)
R> plot(a$y22/1000,type="b",col="blue",axes=FALSE, ann=FALSE)
R> lines(a$yest2/1000,type="b",pch=4, lty=2, col="red")
R> title(xlab="Sample size")
R> axis(1,at=c(1,10,20,30,40,50),lab=c(a$sample[1],a$sample[10],
+ a$sample[20],a$sample[30],a$sample[40],a$sample[50]))
R> axis(2, las=1, at=1*0:g_range[2])
R> legend("topleft", c("Direct","Model"), cex=1, col=c("blue","red"),
+    lty=1:2,pch=c(1,4), bty="n")
R> title(main="Small area estimator Y2", font.main=1.2,cex.main=1)
```
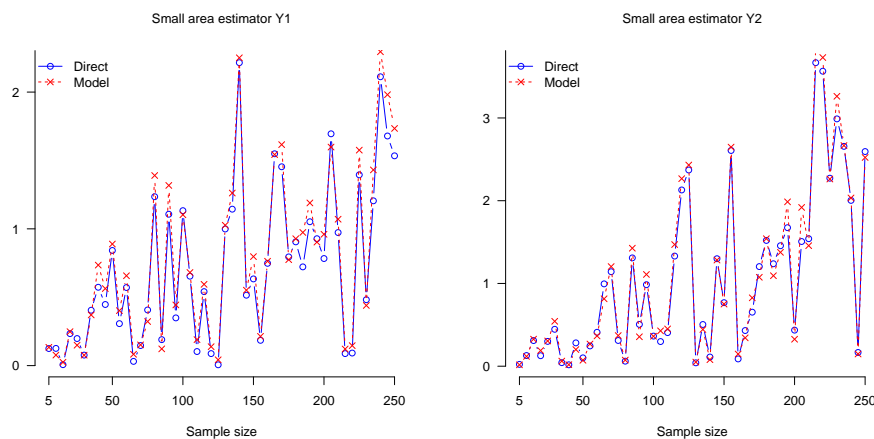

Figure 2: Model estimator and direct estimator sorted by sample size.

The following code calculates the parametric bootstrap BIAS and MSE for the model-based estimators.

```
R> ##Bootstrap parametric BIAS and MSE
R>
R> B=10    #Bootstrap iterations
R> ss=12345 #SEED
R> set.seed(ss)
R> mse.pboot=mseb(pp,datar$Xk,datar$X,datar$Z,datar$n,datar$N,result,B,mod)
R> cv=mse.pboot[[3]]
R>
```

The following code generates Figure 3 that plots the root mean squared error (RMSE) of the model-based estimates.

```
R> dos.ver<-matrix(1:2,1,2)
R> layout(dos.ver)
R> g_range <- range(0,45)
R> plot(cv[,1],type="b", col="blue",axes=FALSE, ann=FALSE)
R> title(xlab="Sample size")
R> axis(1,at=c(1,10,20,30,40,50),lab=c(a$sample[1],a$sample[10],
+ a$sample[20],a$sample[30],a$sample[40],a$sample[50]))
R> axis(2, las=1, at=10*0:g_range[2])
R> title(main="RMSE for the estimator of Y1", font.main=1.2,cex.main=1)
R> g_range <- range(0,45)
R> plot(cv[,2],type="b",col="blue",axes=FALSE, ann=FALSE)
R> title(xlab="Sample size")
R> axis(1,at=c(1,10,20,30,40,50),lab=c(a$sample[1],a$sample[10],
+ a$sample[20],a$sample[30],a$sample[40],a$sample[50]))
R> axis(2, las=1, at=10*0:g_range[2])
R> title(main="RMSE for the estimator of Y2", font.main=1.2,cex.main=1)
R>
```
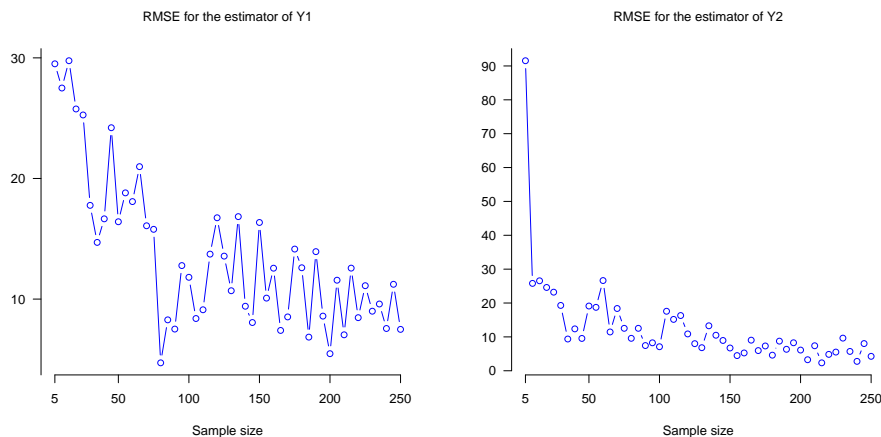

Figure 3: RMSE of model-based estimates

References

Breslow N, Clayton D (1996). "Aproximate inference in generalized linear mixed models." *Journal of the American Statistical Association*, **88**, 9–25.

González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2008). "Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model." *Computational Statistics and Data Analysis*, **52**, 5242–5252.

Herrador M, Morales D, Esteban MD, Sánchez A, Santamaría L, Marhuenda Y, Pérez A,

Molina I (2009). "Estimadores de áreas pequeñas basados en modelos para la Encuesta de Población Activa." *Estadística Española*, **51**(170), 133–137.

López-Vizcaíno M, Lombardía M, Morales D (2013a). "Multinomial-based small area estimation of labour force indicators." *Statistical Modelling*, **13**, 153–178.

López-Vizcaíno M, Lombardía M, Morales D (2013b). "Small area estimation of labour force indicator under a multinomial mixed model with correlated time and area effects." *Submited for review.*

Molina I, Saei A, Lombardía MJ (2007). "Small area estimates of labour force participation under multinomial logit mixed model." *The Journal of the Royal Statistical Society, series A*, **170**, 975–1000.

Prasad NGN, Rao JNK (1990). "The Estimation of the Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association*, **85**(409), 163–171.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Saei A, Chambers R (2003). "Small area estimation under linear an generalized linear mixed models with time and area effects." *Technical report.*

Schall R (1991). "Estimation in Generalized Linear Models with Random Effects." *Biometrika*, **78**, 719–727.

Affiliation:

Esther López Vizcaíno
Instituto Galego de Estatística
Complexo Administrativo San Lázaro
15703 Santiago de Compostela
E-mail: mestherlv32@gmail.com