Package 'JoSAE'

January 20, 2025

Type Package

Title Unit-Level and Area-Level Small Area Estimation

Version 0.3.0

Date 2018-05-31

Author Johannes Breidenbach

Maintainer Johannes Breidenbach <job@nibio.no>

Description Implementation of some unit and area level EBLUP estimators as well as the estimators of their MSE also under heteroscedasticity. The package further documents the publications Breidenbach and Astrup (2012) <DOI:10.1007/s10342-012-0596-7>, Breidenbach et al. (2016) <DOI:10.1016/j.rse.2015.07.026> and Breidenbach et al. (2018 in press). The vignette further explains the use of the implemented functions.

License GPL-2

LazyLoad yes

Depends nlme

Suggests lattice,xtable

NeedsCompilation no

Repository CRAN

Date/Publication 2018-05-06 09:05:03 UTC

Contents

1

| JoSAE-package | *Provides functions for some small area estimators and their mean squared errors.* |

Description

This package implements unit-level (Battese et al. 1988) and area-level EBLUP (Fay and Herriot 1979), and GREG (Sarndal 1984) estimators as well as their variance/MSE estimators. It also contains data and a vignette that explain its use. Heteroscedasticity can be considered.

Details

The aim in the analysis of sample surveys is frequently to derive estimates of subpopulation characteristics. Often, the sample available for the subpopulation is, however, too small to allow a reliable estimate. If an auxiliary variable exists that is correlated with the variable of interest, small area estimation (SAE) provides methods to solve the problem (Rao 2003, Rao and Molina 2015).

The purpose of this package is primarily to document the functions used in the publications Breidenbach and Astrup (2012) and Breidenbach et al. (2018). The data used in Breidenbach and Astrup (2012) are provided. A subset of the data used in Breidenbach et al. (2018) is also provided for testing the functions. The vignette further documents the publication Breidenbach et al. (2015).

You might wonder why this package is called JoSAE. Well, first of all, JoSAE sounds good (if pronounced like the female name). The other reason was that the packages SAE and SAE2 already exist (Gomez-Rubio, 2008). They are, however, not available on CRAN and unmaintained (as of July 2011). They also do not seem to implement the variance estimators that we needed. So I just combined SAE with the first part of my name.

Note

All the implemented functions/estimators are described in Rao (2003). This package merely makes the use of the estimators easier for the users that are not keen on programming. Especially the EBLUP variance estimator would require some effort.

Today, there are several well programmed SAE packages available on CRAN that also provide the functions described here. This was not the case when the first version of the package was uploaded. This package is therefore mostly to document the publications mentioned above. With respect to heteroskedasticity, this package also implements functions that, as far as I can see, are not available in other packages (as of May 2018).

Author(s)

Johannes Breidenbach

Maintainer: Johannes Breidenbach <job@nibio.no>

References

Battese, G. E., Harter, R. M. & Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data Journal of the American Statistical Association, 83, 28-36

Breidenbach, J. and Astrup, R. (2012), Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research, 131:1255-1267.

Breidenbach, J., Ronald E. McRoberts, Astrup, R. (2016), Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. Remote Sensing of Environment, 173, 274-281. https://doi.org/10.1016/j.rse.2015.07.026

Breidenbach, J., Rahlf, J., Magnussen, S., Astrup, R. (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment. In press.

Fay, R. E., 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 (366a), 269-277.

Gomez-Rubio (2008), Tutorial on small area estimation, UseR conference 2008, August 12-14, Technische Universitat Dortmund, Germany.

Rao, J.N.K. (2003), Small area estimation. Wiley.

Rao, J. N., Molina, I., (2015). Small area estimation, 2nd Edition. John Wiley & Sons.

Sarndal, C. (1984), Design-consistent versus model-dependent estimation for small domains Journal of the American Statistical Association, JSTOR, 624-631

Schoch, T. (2011), rsae: Robust Small Area Estimation. R package version 0.1-3.

See Also

eblup.mse.f.wrap, JoSAE.sample.data, JoSAE.domain.data, sae.al.f, sae.ul.f

Examples

```
#mean auxiliary variables for the populations in the domains
data(JoSAE.domain.data)
#data for the sampled elements
data(JoSAE.sample.data)
plot(biomass.ha~mean.canopy.ht,JoSAE.sample.data)

## use the original wrapper function

    #lme model
summary(fit.lme <- lme(biomass.ha ~ mean.canopy.ht, data=JoSAE.sample.data
                       , random=~1|domain.ID))

    #domain data need to have the same column names as sample data or vice versa
d.data <- JoSAE.domain.data
names(d.data)[3] <- "mean.canopy.ht"

result <- eblup.mse.f.wrap(domain.data = d.data, lme.obj = fit.lme)
result
```

```
##END: use the original wrapper function


## the same with a newer function that can consider heteroskedasticity
res <-
    sae.ul.f(samp.data=JoSAE.sample.data,
             population.data=d.data,
#assuming homoskedasticity
             k.ij=rep(1, nrow(JoSAE.sample.data)),
             formula=biomass.ha ~ mean.canopy.ht,
             domain.col="domain.ID",
             sample.id.col="sample.ID",
             neg.sfrac=TRUE)
res$est$est

##END: the same with a newer function that can consider heteroskedasticity
```

---

eblup.mse.f.wrap   *Functions to calculate the variance of an EBLUP estimate.*

Description

Functions to calculate the EBLUP MSE (=variance). The wrap function calls all the other functions
and calculates EBLUP, GREG, SRS, and Synthetic estimates for domain means as well as the
variances of the EBLUP, GREG and SRS estimates.

Usage

```
eblup.mse.f.wrap(domain.data, lme.obj, debug=F, ...)
eblup.mse.f.gamma.i(lme.obj, n.i, ...)
eblup.mse.f.c1(lme.obj, n.i, gamma.i, ...)
eblup.mse.f.c2.ai(lme.obj, n.i, gamma.i, X.i, ...)
eblup.mse.f.c2(gamma.i, X.i, X.bar.i, sum.A.i, ...)
eblup.mse.f.c3.asyvarcovarmat(lme.obj, n.i, ...)
eblup.mse.f.c3(lme.obj, asympt.var.covar, n.i, ...)
eblup.mse.f.c3.star(lme.obj, asympt.var.covar, n.i, mean.resid.i, ...)
```

Arguments

domain.data	data set with the mean of the auxiliary variables for every domain including the domain ID. Names of the variables must be the same as in the unit-level sample data that were used to fit the lme model.
lme.obj	a linear mixed-effects model generated with [lme](#)
n.i	the number of samples within domain i
gamma.i	the gamma_i value resulting from the gamma.i method of this function
X.i	the design matrix of sampled elements in domain i

X.bar.i	mean of the populations elements design matrix in domain i
sum.A.i	sum of the domains A_i matrices resulting from eblup.mse.f.c2.ai
asympt.var.covar	
	the asymptotic variance-covariance matrix of the mixed-effects model resulting from eblup.mse.f.c3.asyvarcovarmat
mean.resid.i	the mean residual of the fixed-part of the linear mixed-effects model in domain i (i.e., use level=0 in predict.lme)
...	forward attributes to other functions. Not used so far.
debug	details are printed if true - Only used by the wrapper function

Details

Most users will probably only use the convenient wrap function. Nonetheless, all components of the EBLUP variance can also be calculated separately.

Value

A compontent of the EBLUP variance (aka mean squared error). Which component depends on the function used.

The wrap function returns a data frame with many entrys for every domain: The domain-level predictor variables obtained from the domain.data, the mean of the predictor variables and response (aka the sample mean or SRS estimate) observed at the samples, the number of samples (n.i.sample), the mean residuals of a lm (fitted using the fixed part of the lme) and the lme (mean.resid.lm, mean.resid.lme), the synthetic (Synth), EBLUP, and GREG estimates of the mean of the variable of interest, the gamma_i value, the variance of the means for the sample and GREG estimates (.var.mean), the components of the EBLUP variance (c1-c3star), the results of the first and the second method (cf. Rao 2003) to derive the EBLUP variance (EBLUP.var.1, EBLUP.var.2), and the standard errors derived from the variances (.se).

Note

Currently, only random intercept mixed-effects models with homogeneous variance structure are supported.

Author(s)

Johannes Breidenbach

References

Breidenbach and Astrup (2012), Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research, 131:1255-1267.

See Also

JoSAE-package for more examples

Examples

```
library(nlme)
data(JoSAE.sample.data)
#fit a lme
summary(fit.lme <- lme(biomass.ha ~ mean.canopy.ht, data=JoSAE.sample.data
                        , random=~1|domain.ID))
#calculate the first component of the EBLUP variance for a domain with 5 samples
eblup.mse.f.c1(fit.lme, 5, 0.2)
```

JoSAE.domain.data	*Dataframe containing the domains population and number of elements and mean of the auxiliary variable*

Description

Auxiliary variable: Mean canopy height derived from a photogrammetric canopy height model

Usage

```
data(JoSAE.domain.data)
```

Format

A data frame with 14 observations on the following 3 variables.

domain.ID a numeric vector

N.i a numeric vector - number of population elements

mean.canopy.ht.bar a numeric vector - mean of the auxiliary variable

Source

Breidenbach, J. and Astrup, R. (2012), Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research, 131:1255-1267.

See Also

[JoSAE-package](#) for more examples

Examples

```
data(JoSAE.domain.data)

str(JoSAE.domain.data)
```

JoSAE.sample.data | *Sample plots of the Norwegian National Forest Inventory (NNFI) with a variable of interest and an auxiliary variable*

Description

Above ground forest biomass over all tree species is the variable of interest. Mean canopy height derived from a photogrammetric canopy height model of 20~cm geometric and 10~cm radiometric resolution is the auxiliary variable.

Usage

```
data(JoSAE.sample.data)
```

Format

A data frame with 145 observations on the following 4 variables.

sample.ID a numeric vector

domain.ID a numeric vector

biomass.ha a numeric vector of the variable of interest

mean.canopy.ht a numeric vector of the auxiliary variable

Source

Breidenbach, J. and Astrup, R. (2012), Small area estimation of forest attributes in the Norwegian National Forest Inventory. European Journal of Forest Research, 131:1255-1267.

See Also

[JoSAE-package](#) for more examples

Examples

```
data(JoSAE.sample.data)
## maybe str(JoSAE.sample.data) ; plot(JoSAE.sample.data) ...
plot(biomass.ha~mean.canopy.ht,JoSAE.sample.data)
```

---

landsat   *LANDSAT data: Prediction of County Crop Areas Using Survey and Satellite Data*

Description

The `landsat` data.frame is a compilation (by Battese et al., 1988) of survey and satellite data. It consists of data on segments (primary sampling unit; 1 segement =approx= 250 hectares) under corn and soybeans for 12 counties in north-central Iowa; see Details, below.

The landsat data.frame was made available by Tobias Schoch with the R package rsae. Since rsae was archived as of R 3.0.2, the data and this description was copied from rsae 0.1-4 in the archives.

Usage

```
data(landsat)
```

Format

A data frame with 37 observations on the following 10 variables.

SegmentsInCounty a numeric vector; no. of segments per county

SegementID a numeric vector; sample segment identifier (per county)

HACorn a numeric vector; hectares of corn for each sample segment (as reported in the June 1978 Enumerative Survey)

HASoybeans a numeric vector; hectares of soybeans for each sample segment (as reported in the June 1978 Enumerative Survey)

PixelsCorn a numeric vector; no. of pixels classified as corn for each sample segment (LANDSAT readings)

PixelsSoybeans a numeric vector; no. of pixels classified as soybeans for each sample segment (LANDSAT readings)

MeanPixelsCorn a numeric vector; county mean number of pixels classified as corn

MeanPixelsSoybeans a numeric vector; county mean number of pixels classified as soybeans

outlier a logical vector; flags observation no. 33 as outlier

CountyName a factor with levels (i.e., county names) `Cerro Gordo` `Hamilton` `Worth` `Humboldt` `Franklin` `Pocahontas` `Winnebago` `Wright` `Webster` `Hancock` `Kossuth` `Hardin`

Details

The `landsat` data is a compilation (by Battese et al., 1988) of the LANDSAT satellite data from the U.S. Department of Agriculture (USDA) and the 1978 June Enumerative Survey.

Survey data: The survey data on the areas under corn and soybeans (reported in hectares) in the 37 segments of the 12 counties (north-central Iowa) have been determined by USDA Statistical Reporting Service staff, who interviewed farm operators. A segment is about 250 hectares.

Satellite data: For the LANDSAT satellite data, information is recorded as "pixels". The USDA has been engaged in research toward transforming satellite information into good estimates of crop areas at the individual pixel and segments level. A pixel is about 0.45 hectares. The satellite (LANDSAT) readings were obtained during August and September 1978.

Data for more than one sample segment are available for several counties (i.e, unbalanced data).

Observations No. 33 has been flagged as outlier (cf., Battese et al. (1988, p. 28).

Source

The data landsat is from Table 1 of Battese et al. (1988, p. 29).

Schoch (2011) rsae: Robust Small Area Estimation, R package version 0.1-4: http://cran.r-project.org/src/contrib/Archive/rsa
4.tar.gz

References

Battese, G.E, R.M. Harter, and W.A. Fuller (1988): An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association* 83, pp. 28–36.

Examples

```
data(landsat)
```

---

nfi	*Sample plots of the Norwegian National Forest Inventory (NFI) with a variable of interest and an auxiliary variable*

Description

A total of 131 sample plots of the Norwegian National Forest Inventory (NFI) with timber volume interpolated and extrapolated to the year 2011 as the variable of interest and mean vegetation height derived from image matching as the auxiliary variable.

Usage

```
data(nfi)
```

Format

A data frame with 131 observations on the following 2 variables.

vol.2011 numeric vector of the variable of interest

Elev.Mean numeric vector of the auxiliary variable

Source

Breidenbach, J., Ronald E. McRoberts, Astrup, R. (2015), Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. Remote Sensing of Environment. In press.

See Also

[JoSAE-package](#) and Vignette for more examples

Examples

```
data(nfi)
plot(vol.2011~Elev.Mean, nfi)
```

| sae.al.f | *Area-level small area estimation under heteroscedasticity or homoscedasticity.* |

Description

Area-level small area estimation, possibly under heteroscedasticity. Assumes SRS within domains.

Usage

```
sae.al.f(domain.id, n.i, psi.i, formula, data, b.i, type = "RE", verbose
= F, iter = 100, a.conv = 0.001,...)
```

Arguments

domain.id	Vector of unique domain IDs corresponding to data.
n.i	Vector of number of samples within each domain i, corresponding to data.
psi.i	Vector of variances of the direct estimator within each domain i, corresponding to data.
formula	Formula for fixed effects.
data	Domain-level data.
b.i	Parameter for modeling heteroscedasticity.
type	Method for estimating variance parameters. Currently, only option is "RE" for REML.
verbose	Boolean. Should more info be printed?
iter	Max number of iterations. Default is 100.
a.conv	Threshold for accepting convergency. Default is 0.001
...	Parameters forwarded to other functions.

Details

SRS (simple random sampling) is assumed within each domain. If the direct estimator is not based on SRS, RV components of the results do not make sense.

Value

results
: Area-level estimates and associated standard errors (SE). FH Fay-Harriot estimate (the standard area-level estimator), RV Rivest-Vandal estimate, considering uncertainty in variance parameters. `bias.sig.sq.v` Bias in random effect variance. `synth` Synthetic estimate. `trans.resid.v.i` Transformed residuals. The other parameters are details and are named as in the reference.

beta.hat
: Estimated fixed-effects parameters.

des.mat
: Design matrix.

Other list elements have descriptive names.

Author(s)

Johannes Breidenbach

References

Breidenbach et al. (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment.

Examples

```
library(nlme)
#sample data
data(ulal.sub.samp.dat)
#domain means of x
data(ulal.sub.dom.dat)

#The easiest way of getting the data into the right format
#is to fit a unit-level model first.
res <-
    sae.ul.f(samp.data=ulal.sub.samp.dat,
             population.data=ulal.sub.dom.dat,
             k.ij=ulal.sub.samp.dat[,"k.ij.one"],
             formula=w.VMPRHA ~ elev.mean + elev.mean.sq,
             domain.col="stand.ID",
             sample.id.col="plot.ID",
             neg.sfrac=TRUE)

#data for AL
dat.al <- merge(res$data$samp.agg.X.pop[,c("domain.id","n.i","w.VMPRHA.ybar.i","elev.mean.X.pop")],
                res$est$se[,c("domain.id","se.srs")])

#area-level SAE under homoscedasticity
res <- sae.al.f(
    domain.id=dat.al$domain.id , n.i=dat.al$n.i , psi.i=dat.al$se.srs^2
```

```
    , formula=w.VMPRHA.ybar.i ~ elev.mean.X.pop , data=dat.al
    , b.i=rep(1, nrow(dat.al))
    , type="RE")

#area-level SAE heteroskedasticity
bi.par <- 0.39#select b.i parameter
#helper function
range01 <- function(x, ...){(x - min(x, ...)) /
                                (max(x, ...) - min(x, ...))}
res <- sae.al.f(
    domain.id=dat.al$domain.id , n.i=dat.al$n.i , psi.i=dat.al$se.srs^2
    , formula=w.VMPRHA.ybar.i ~ elev.mean.X.pop, data=dat.al
    , b.i=range01(dat.al$elev.mean.X.pop)+bi.par
    , type="RE")
```

---

sae.ul.f   *Unit-level small area estimation under heteroscedasticity.*

Description

The only function that should be used is `sae.ul.f` which wraps the data preparation (`ul.data.prep.f`), parameter estimation (`ul.reml.f`), and EBLUP and MSE estimation (`ul.est.f`) functions. The other functions are helpers that do not need to be called by the user directly.

Usage

```
sae.ul.f(...)
ul.data.prep.f(k.ij, samp.data, population.data, formula, domain.col,
sample.id.col, neg.sfrac, pop.r, sum.i.k.ij.sq.r, N.i, ...)
ul.reml.f(samp.data, formula, samp.agg.X.pop, y.name, X.names, ...)
ul.est.f(samp.data, samp.agg.X.pop, X.names, y.name, beta.hat,
cov.beta.hat, sig.sq.e, sig.sq.v, V.bar.ee, V.bar.vv, V.bar.ve,
neg.sfrac, resid = F, ...)
```

Arguments

`samp.data`	Data frame of unit-level sample data of the response and explanatory variables. Transformations of x have to be calculated before hand, not using formula.
`k.ij`	Vector of the same length and corresponding to `samp.data` containing the k_ij values for modeling heteroscedasticity. A vector of 1's if heteroscedasticity is not considered. k_ij is squared and used as the input to the `varFix` variance function in `lme`.
`population.data`	
	Data frame of domain-level means of the explanatory variables. Names have to match to sample data.
`formula`	The fixed effects formula to be used in the mixed model using `lme`.

domain.col	Character string identifying the column name in sample.data and population.data containing unique domain IDs.
sample.id.col	Character string identifying the column name in sample.data containing unique sample IDs.
neg.sfrac	Boolean. Are sampling fractions negligible?
pop.r	Only required if neg.sfrac is FALSE. Same as population.data but calculated with the sample elements removed.
sum.i.k.ij.sq.r	
	Only required if neg.sfrac is FALSE. Same as the values in the population.k.column but calculated with the sample elements removed.
N.i	Only required if neg.sfrac is FALSE. Number of elements within domain i.
samp.agg.X.pop	Data frame containing the aggregated sample data and domain-level explanatory variables.
y.name	Character string identifying the column name of the response.
X.names	Vector of character strings identifying the column names of the explanatory variables.
beta.hat	Vector of estimated fixed effects.
cov.beta.hat	Covariance matrix of fixed effects.
sig.sq.e	Residual variance.
sig.sq.v	Random effect variance.
V.bar.ee	Assymptotic variance of the residual variance.
V.bar.vv	Assymptotic variance of the random effect variance.
V.bar.ve	Assymptotic covariance of the residual and random effect variance.
resid	Boolean. Should residuals be returned?
...	Parameters forwarded to other functions.

Details

These functions can also be used for eblups without heteroskedasticity as the the older functions around eblup.mse.f.wrap.

Value

est	List of point estimates and standard errors based on different estimators. est: Point estimates of EBLUP, SRS, GREG, Survey regression (svreg), and EBLUP synthetic estimators. Also contains random effects. se: Standard errors. EBLUP global was used in the reference publication. resids: Different forms of raw and transformed residuals.
var.pars	List of estimated variance parameters. Among others the lme object fitted to the data. See list of arguments above.
data	List of data frames in the format required by the other functions.

Author(s)

Johannes Breidenbach

References

Breidenbach et al. (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment.

Examples

```
library(nlme)
#sample data
data(ulal.sub.samp.dat)
#domain means of x
data(ulal.sub.dom.dat)

#eblup under homoskedasticity
res <-
    sae.ul.f(samp.data=ulal.sub.samp.dat,
             population.data=ulal.sub.dom.dat,
             k.ij=ulal.sub.samp.dat[,"k.ij.one"],
             formula=w.VMPRHA ~ elev.mean + elev.mean.sq,
             domain.col="stand.ID",
             sample.id.col="plot.ID",
             neg.sfrac=TRUE)

#eblup under heteroskedasticity
res <-
    sae.ul.f(samp.data=ulal.sub.samp.dat,
             population.data=ulal.sub.dom.dat,
             k.ij=ulal.sub.samp.dat[,"k.ij.em.0.48"],
             formula=w.VMPRHA ~ elev.mean + elev.mean.sq,
             domain.col="stand.ID",
             sample.id.col="plot.ID",
             neg.sfrac=TRUE)
```

---

ulal.sub.dom.dat   *Domain-level data.*

Description

Domain-level means of the explanatory variables obtained from digital aerial photogrammetry (i.e., means of all elements not just the means of the samples). The domains are 15 forest stands in the municipality of Stokke, Norway.

Usage

```
data("ulal.sub.dom.dat")
```

Format

A data frame with 15 observations on the following 5 variables.

`stand.ID` a factor with unique stand IDs.

`elev.mean` a numeric vector. Mean height obtained from digital aerial photogrammetry.

`elev.mean.sq` a numeric vector. Mean height squared obtained from digital aerial photogrammetry.

`k.i.one` a numeric vector of ones.

`sum.i.k.ij.sq.em.0.48.r` a numeric vector. Domain-level sums of the k_ij values of all elements that were not sampled. Needed if fpc cannot be neglected in EBLUP estimation.

`n.pixels` Number of elements with the domain i.

Source

Breidenbach et al. (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment.

Examples

```
data(ulal.sub.dom.dat)
```

---

ulal.sub.samp.dat *Unit-level sample data.*

Description

73 sample plots clustered in 15 stands in the municipality of Stokke, Norway.

Usage

```
data("ulal.sub.samp.dat")
```

Format

A data frame with 73 observations on the following 7 variables.

`plot.ID` a character vector with unique plot IDs.

`w.VMPRHA` a numeric vector. Timber volume per hectare (m3/ha).

`stand.ID` a factor with unique stand IDs.

`elev.mean` a numeric vector. Mean height obtained from digital aerial photogrammetry.

`elev.mean.sq` a numeric vector. Mean height squared obtained from digital aerial photogrammetry.

`k.ij.one` a numeric vector of ones (used in cases where heteroscedasticity can be ignored)

`k.ij.em.0.48` a numeric vector. `elev.mean` to the power of 0.48 for modeling heteroscedasticity.

Source

Breidenbach et al. (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sensing of Environment.

Examples

```
data(ulal.sub.samp.dat)
```

Index