




sptotal: an R package for predicting totals and weighted sums from spatial data

Matt Higham¹, Jay Ver Hoef², Bryce Frank³, and Michael Dumelle⁴

¹ St. Lawrence University ² National Oceanic and Atmospheric Administration ³ Bureau of Land Management ⁴ United States Environmental Protection Agency

DOI:

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Submitted:

Published:

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

In ecological monitoring surveys of an animal population, a plant population, or an environmental resource, predicting the total abundance, the mean, or some other quantity in a finite region is often of interest. However, because of time and money constraints, sampling the entire region is often unfeasible. The purpose of the `sptotal` R package is to provide software that gives a prediction for a quantity of interest, such as a total, and an associated standard error for the prediction. The predictor, referred to as the Finite-Population-Block-Kriging (FPBK) predictor in the literature (J. M. Ver Hoef, 2008), incorporates possible spatial correlation in the data and also incorporates an appropriate variance reduction for sampling from a finite population.

In the remainder of the paper, we give an overview of both the background of the method and of the `sptotal` package.

Statement of Need

The primary purpose of `sptotal` is to provide an implementation of the Finite Population Block Kriging (FPBK) methods developed in J. Ver Hoef (2002) and J. M. Ver Hoef (2008). While we refer the interested reader to those sources for the full development of the FPBK predictor, we provide a very short overview of the setting in which the predictor can be used.

Suppose that we have a response variable $Y(\mathbf{s}_i)$, $i = 1, 2, \dots, N$, where the vector \mathbf{s}_i contains the coordinates for the i^{th} spatial location and N is a finite number of spatial locations. Then \mathbf{y} , a vector of the $Y(\mathbf{s}_i)$, can be modeled with a spatial linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1)$$

where \mathbf{X} is a design matrix for the fixed effects and $\boldsymbol{\beta}$ is a parameter vector of fixed effects. The random errors, $\boldsymbol{\epsilon}(\mathbf{s}_i)$, have a mean of $\mathbf{0}$ and a covariance of

$$\text{var}(\boldsymbol{\epsilon}) = \tau^2 \mathbf{I} + \sigma^2 \mathbf{R}, \quad (2)$$

where τ^2 is the spatial independent error variance (commonly called the nugget), \mathbf{I} is the identity matrix, σ^2 is the spatial dependent error variance (commonly called the partial sill), and \mathbf{R} is a spatial correlation matrix. A common model used to generate \mathbf{R} is the exponential correlation function (Cressie, 2015).

Our goal is to predict some linear function of the response, $f(\mathbf{y}) = \mathbf{b}'\mathbf{y}$, where \mathbf{b} is a vector of weights. A common vector of weights is a vector of 1's so that the resulting

prediction is for the total abundance across all sites. In a finite population setting, if we are interested in predicting the realized total and we sampled all N spatial locations, then we would simply add up the realized values of $Y(\mathbf{s}_i)$ across all N locations. However, in many practical settings, we do not observe all N locations and instead sample a subset of these locations.

If only some of the values in \mathbf{y} are observed, then the `sptotal` package can be used to find the the Best Linear Unbiased Predictor (BLUP) for $\mathbf{b}'\mathbf{y}$, referred to as the FPBK predictor, along with its prediction variance. When the number of sites sampled is equal to the total number of sites in the region, we know the realized total exactly and the prediction variance is equal to 0.

Package Methods

Before discussing comparable methods and R packages, we show how the main functions in `sptotal` can be used on a real data set to predict total abundance. We use the `AKmoose_df` data in the `sptotal` package, provided by the Alaska Department of Fish and Game.

```
library(sptotal)
data("AKmoose_df")
```

The data contains a response variable `total`, x-coordinate centroid variable `x`, y-coordinate centroid variable `y`, and covariates `elev_mean` (the elevation) and `strat` (a stratification variable). There are a total of 860 rows of unique spatial locations. Locations that were not surveyed have an NA value for `total`.

The two primary functions in `sptotal` are `slmfit()`, which fits a spatial linear model, and `predict.slmfit()`, which predicts a quantity of interest (such as a mean or total) using a fitted `slmfit` object. `slmfit()` has required arguments `formula`, `data`, `xcoordcol`, and `ycoordcol`. If `data` is a simple features object from the `sf` (E. J. Pebesma et al., 2018) package, then `xcoordcol` and `ycoordcol` are not required. The `CorModel` argument is the correlation model used for the errors.

```
moose_mod <- slmfit(total ~ elev_mean + strat, data = AKmoose_df,
                   xcoordcol = "x", ycoordcol = "y",
                   CorModel = "Exponential")
summary(moose_mod)
```

```
##
## Call:
## total ~ elev_mean + strat
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.695 -3.768 -1.304  1.111 35.816
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.700203   2.128817   0.329  0.74254
## elev_mean    0.004479   0.006620   0.677  0.49944
## stratM      2.586271   0.868335   2.978  0.00323 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Covariance Parameters:
```

```
##                               Exponential Model
## Nugget                       29.71177
## Partial Sill                  8.04658
## Range                         33.65766
##
## Generalized R-squared: 0.03966569
```

With the `summary()` generic, we obtain output similar to the summary output of a linear model fit with `lm()`, as well as a table of fitted covariance parameter estimates. Next, we use `predict()` to obtain a prediction for the total abundance across all spatial locations, along with a standard error for the prediction. By default, `predict()` gives a prediction for total abundance, though the default can be modified by specifying a column of prediction weights for the vector `b` with the `wtscol` argument.

```
predict(moose_mod)
```

```
## Prediction Info:
##      Prediction      SE 90% LB 90% UB
## total      1610 413.2  930.1   2290
##      Numb. Sites Sampled Total Numb. Sites Total Observed Average Density
## total                218           860           742           3.404
```

The output of printed `predict()` gives a table of prediction information, including the **Prediction** (a total abundance of 1610 moose, in this example), the **SE** (Standard Error) of the prediction, and bounds for a prediction interval (with a nominal level of 90% by default). Additionally, some summary information about the data set used is given.

`sptotal` also provides many helper generic functions for spatial linear models. The structure of the arguments and of the output of these generics often mirrors that of the generics used for base R linear models fit with `lm()`. Examples (applied to the `moose_mod` object) include `AIC(moose_mod)`, `coef(moose_mod)`, `fitted(moose_mod)`, `plot(moose_mod)`, and `residuals(moose_mod)`.

Comparable Methods and Related Work

Methods that can also be used to predict a total, mean, or other quantity in a finite population include design-based methods. Design-based methods make inferences based on how the sample was selected. One example of a design-based sampling method for spatial data is the Generalized Random Tessellation Stratified (GRTS) spatially balanced sampling algorithm (Stevens Jr & Olsen, 2004). If a GRTS sample is taken, then an analysis using a local neighborhood variance estimator can be used to obtain a prediction for a population total or mean with a variance that has a finite population adjustment. Dumelle, Higham, Ver Hoef, Olsen, & Madsen (2022) show that FPBK outperforms the GRTS design-based analysis in simulations.

Statistical learning methods can also be applied to spatial data to obtain a prediction for a finite population total or mean. For example, k-nearest-neighbors [Fix (1985); knn] is a statistical learning algorithm popular in forestry applications that makes predictions at unobserved locations from the values of the closest observed locations. However, quantifying uncertainty in a prediction resulting from knn is much more challenging. Additionally, J. M. Ver Hoef & Temesgen (2013) show that FPBK outperforms knn in many settings.

Note that there are many spatial packages in R that can be used to predict values at unobserved locations, including `gstat` (E. Pebesma, Graeler, & Pebesma, 2015), `geoR` (Ribeiro Jr, Diggle, Ribeiro Jr, & Suggests, 2007), and `spmodel` (INSERT CITATION), among other packages. What `sptotal` contributes is the ability to obtain a prediction

variance that incorporates a variance reduction when sampling from a finite number of sampling units.

Past and Ongoing Research Projects

Dumelle et al. (2022) used the `sptotal` package to compare model-based and design-based approaches for analysis of spatial data. Currently, a `Shiny` app is in development at the Alaska Department of Fish and Game that uses `sptotal` to predict abundance from moose surveys conducted in Alaska.

References

- Cressie, N. (2015). *Statistics for spatial data - revised edition*. John Wiley & Sons.
- Dumelle, M., Higham, M., Ver Hoef, J. M., Olsen, A. R., & Madsen, L. (2022). A comparison of design-based and model-based approaches for finite population spatial sampling and inference. *Methods in Ecology and Evolution*, 13(9), 2018–2029.
- Fix, E. (1985). *Discriminatory analysis: Nonparametric discrimination, consistency properties* (Vol. 1). USAF school of Aviation Medicine.
- Pebesma, E. J. et al. (2018). Simple features for r: Standardized support for spatial vector data. *R J.*, 10(1), 439.
- Pebesma, E., Graeler, B., & Pebesma, M. E. (2015). Package “gstat.” *Comprehensive R Archive Network (CRAN)*, 1–0.
- Ribeiro Jr, P. J., Diggle, P. J., Ribeiro Jr, M. P. J., & Suggests, M. (2007). The geoR package. *R news*, 1(2), 14–18.
- Stevens Jr, D. L., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American statistical Association*, 99(465), 262–278.
- Ver Hoef, J. (2002). Sampling and geostatistics for spatial data. *Ecoscience*, 9(2), 152–161.
- Ver Hoef, J. M. (2008). Spatial methods for plot-based sampling of wildlife populations. *Environmental and Ecological Statistics*, 15(1), 3–13.
- Ver Hoef, J. M., & Temesgen, H. (2013). A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. *PloS one*, 8(3), e59129.