

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


Matt Higham<sup>1</sup>, Jay Ver Hoef<sup>2</sup>, Bryce Frank<sup>3</sup>, and Michael Dumelle<sup>4</sup> 

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

DOI: [10.21105/joss.05363](https://doi.org/10.21105/joss.05363)

## Software

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

Editor: Fabian Scheipl 

## Reviewers:

- [@Athene-ai](#)
- [@garretrc](#)

Submitted: 17 March 2023

Published: 24 May 2023

## License

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

## Summary

In ecological or environmental surveys, it is often desired to predict the mean or total of a variable in some finite region. 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

sptotal provides an implementation of the Finite Population Block Kriging (FPBK) methods developed in J. Ver Hoef ([2002](#)) and J. M. Ver Hoef ([2008](#)). Next we provide a short overview of FPBK.

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

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

where  $\mathbf{X}$  is a design matrix for the fixed effects and  $\beta$  is a parameter vector of fixed effects. The vector of random errors follows a multivariate normal distribution with a mean vector of  $\mathbf{0}$  and a covariance of

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

where  $\sigma^2$  is the spatial dependent error variance (commonly called the partial sill),  $\tau^2$  is the spatial independent error variance (commonly called the nugget), and  $\mathbf{I}$  is the identity matrix. The  $i^{th}$  row and  $j^{th}$  column of the  $N \times N$  spatial correlation matrix  $\mathbf{R}$  contains the correlation between the random error of the  $i^{th}$  spatial location,  $\epsilon_i$ , and the random error of the  $j^{th}$  spatial location,  $\epsilon_j$ . A common model used to generate  $\mathbf{R}$  is the exponential correlation function ([Cressie, 2015](#)).

FPBK predicts some linear function of the response,  $f(\mathbf{y}) = \mathbf{b}'\mathbf{y}$ , where  $\mathbf{b}$  is an  $N$ -length column 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. 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.

The primary functions in the `sptotal` package are described in the following section. In short, the FPBK method is implemented in `sptotal`'s `predict()` generic function, which is used on a spatial model that is fit with `sptotal::slmfit()`.

## 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 of moose in a region of Alaska. 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 uses FPBK to predict 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` (Pebesma & others, 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 implement FPBK and 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

Design-based analysis and k-nearest neighbors ([Fix, 1985](#)) are two approaches that can be used to compute a mean or total in a finite population. Dumelle, Higham, Ver Hoef, Olsen, & Madsen ([2022](#)) provide an overview of design-based spatial analysis and FPBK, showing that FPBK often outperforms the design-based analysis. J. M. Ver Hoef & Temesgen ([2013](#)) show that FPBK often outperforms k-nearest-neighbors and highlight that quantifying uncertainty is much more challenging with k-nearest-neighbors.

Note that there are many spatial packages in R that can be used to predict values at unobserved locations, including `gstat` ([Pebesma, 2004](#)), `geoR` ([Ribeiro Jr, Diggle, Schlather, Bivand, & Ripley, 2020](#)), and `spmode` ([Dumelle, Higham, & Ver Hoef, 2023](#)), among others. What `sptotal` contributes is the ability to obtain the appropriate variance of a linear combination of predicted values 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.

## Disclaimer

The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## References

- Cressie, N. (2015). *Statistics for spatial data - revised edition*. John Wiley & Sons.
- Dumelle, M., Higham, M., & Ver Hoef, J. M. (2023). sptotal: Spatial statistical modeling and prediction in R. *PLOS ONE*, 18(3), e0282524. doi:[10.1371/journal.pone.0282524](https://doi.org/10.1371/journal.pone.0282524)
- 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. doi:[10.1111/2041-210X.13919](https://doi.org/10.1111/2041-210X.13919)
- Fix, E. (1985). *Discriminatory analysis: Nonparametric discrimination, consistency properties* (Vol. 1). USAF school of Aviation Medicine. doi:[10.1037/e471672008-001](https://doi.org/10.1037/e471672008-001)
- Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers & geosciences*, 30(7), 683–691. doi:[10.1016/j.cageo.2004.03.012](https://doi.org/10.1016/j.cageo.2004.03.012)
- Pebesma, E. J., & others. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1), 439. doi:[10.32614/rj-2018-009](https://doi.org/10.32614/rj-2018-009)
- Ribeiro Jr, P. J., Diggle, P. J., Schlather, M., Bivand, R., & Ripley, B. (2020). *geoR: Analysis of geostatistical data*. Retrieved from <https://CRAN.R-project.org/package=geoR>
- Ver Hoef, J. (2002). Sampling and geostatistics for spatial data. *Ecoscience*, 9(2), 152–161. doi:[10.1080/11956860.2002.11682701](https://doi.org/10.1080/11956860.2002.11682701)
- Ver Hoef, J. M. (2008). Spatial methods for plot-based sampling of wildlife populations. *Environmental and Ecological Statistics*, 15(1), 3–13. doi:[10.1007/s10651-007-0035-y](https://doi.org/10.1007/s10651-007-0035-y)
- 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. doi:[10.1371/journal.pone.0059129](https://doi.org/10.1371/journal.pone.0059129)