**Title:** remotePARTS: statistical analysis of very large spatial and spatiotemporal datasets

**Abstract**

1. Many spatiotemporal environmental datasets exhibit both temporal and spatial autocorrelation. Although statistical methods are available to account for temporal and spatial autocorrelation, these methods struggle to analyze the large remote-sensing datasets that consist of maps containing millions of pixels, with each pixel containing a time series of data.
2. PARTS (Partitioned Autoregressive Time Series) analysis can be used to conduct map-scale estimation and test hypotheses that are formulated as regressions. Here, we present remotePARTS, a software package for the R statistical programming language that contains the tools to conduct PARTS analyses. To demonstrate the applicability of PARTS to a variety of statistical and ecological problems, we conducted a set of simulation studies with the remotePARTS software.
3. We found that PARTS is a robust and accurate statistical approach for testing a variety of hypotheses. remotePARTS performed well in testing hypotheses about the effects of spatial variables, temporal variables, and spatiotemporal variables on spatial and spatiotemporal responses.
4. These results demonstrate that remotePARTS solves many of the challenges of using big spatiotemporal data to understand ecological, biogeographical, and environmental problems at global scales.

**Introduction**

Many environmental problems involve time and space. How does a species use different habitat types, and how does habitat use change over decades? Are fires becoming more prevalent or larger in some regions and not others? Are changes in plant phenology, driven by increasing global temperatures, more pronounced in some regions, and are these phenological changes increasing through time? These example questions all illustrate the importance of understanding spatiotemporal systems from a statistical perspective. Ecologists and environmental scientists are constantly addressing such questions, yet few analyses appropriately account for both spatial and temporal autocorrelation (refs). Tobler’s first law of geography (ref 1969) states that nearby entities are more similar than distant ones. This is often true not only in space, but also in time: nearby locations are exposed to similar environmental conditions and those conditions tend to change slowly over time. Methods that fail to account for spatiotemporal autocorrelation can both falsely identify patterns that do not exist and overlook patterns that do ().

Many of the questions that researchers ask using spatiotemporal datasets can be caset in the form of a regression with a response (dependent) variable (e.g., habitat types, plant phenology) and multiple explanatory (independent) variables (e.g., time, latitude, land-cover classification). A simple and flexible regression model that contains spatiotemporal autocorrelation is

*yi*(*t*) = **X***i*(*t*)*Bi* + *i*(*t*) (Eq. 1)

*i*(*t*) = r*i**i*(*t*–1) + *i*(*t*)

*i*(*t*) ~ N(0, 2****(**D**))

where *yi*(*t*) is the response variable for location *i* (*i* = 1,..., *n*) at time *t* (*t* = 1, ..., *T*). We assume there are *k* explanatory variables contained in the 1 x *k* vector **X***i*(*t*). These explanatory variables may change through time; they may differ among locations but be temporally unchanging; they may consist of zeros and ones that differ among locations to give categorical variables corresponding to temporally invariant contrasts among locations; or they may equal one to give an intercept. The random error *i*(*t*) is a temporal autoregressive process of order 1, AR(1), in which the innovations *i*(*t*) are spatially autocorrelated innovations given by the covariance matrix 2****(**D**). The covariance matrix 2****(**D**) contains the covariance between *i*(*t*) and *j*(*t*) from locations *i* and *j* that depends on the distance *dij* between them contained in the matrix **D**.

The model in equation (1) is linear regression in which the response and explanatory variables can vary in both space and time, and the random error *i*(*t*) contains spatiotemporal autocorrelation. Thus, it has the flexibility and generality of simple linear used broadly outside the realm of spatiotemporal dynamics. It encompasses polynomial regression, and appropriate transforms of response and explanatory variables justify its use for non-Gaussian data. Furthermore, classical statistical methods such as Generalized Least Squares (GLS), Maximum Likelihood (ML), and Restricted Maximum Likelihood (REML) can be used for modeling fitting, giving rise to parameter estimates and hypothesis tests. Nonetheless, in application to big data which are increasingly common in ecology and evolutionary biology, classical methods are not computationally feasible. They are limited by the need to invert the covariance matrix of *i*(*t*), whose computational burden scales with the cube of the number of elements it contains, (*n*\**T*)3. For even a small remote-sensing study, for example, there may be 100,000 pixels containing data for 30 years, and inverting a dense, 3,000,000 x 3,000,000 matrix might take a lifetime with existing computer resources.

**remotePARTS**

PARTS (Partitioned Autoregressive Time Series) is a two-step approach to statistical inference for spatiotemporal datasets that can account for spatiotemporal autocorrelation (Ives et al 2021a). The first step consists of a time-series regression analysis which effectively collapses the temporal dimension into a single parameter of interest. For example, if there were a specific variable of interest, say *xj,i*(*t*) in equation (1), then *n* regressions would be performed for each of the *i* (*i* = 1, ..., *n*) locations, and coefficients *bj,i* estimated. The second step uses GLS to regress the coefficients *bj,i* onto temporally invariant explanatory variables that differ among locations. In this way, both spatial and temporal variation are incorporated into the model but calculated separately. Although this reduces the full spatiotemporal model to a spatial model, for large datasets this is still numerically challenging. PARTS addresses this problem by subsetting the spatial dataset into random partitions, estimating parameters from each partition, and performing a single test on the collection of results. As a consequence of partitioning, the computational burden for PARTS scales linearly with *N*. The statistical result making this possible is computing the covariance between the test statistics calculated from each partition so that an overall test can be computed {Ives, 2021 #12495;Ives, 2022 #12315}.

Our R package, remotePARTS (github ref), provides the tools for implementing PARTS with any spatial or spatiotemporal dataset (Table 1). Two functions are provided for time-series analyses for the first step of PARTS; fitCLS() and fitAR(), respectively, use conditional least squares (CLS) and regression with AR(1) autocorrelated errors fit using REML. The companion functions fitCLS\_map() and fitAR\_map() apply these time-series methods to all pixels in a map. Users can also implement their own time-series analyses in place of those provided in remotePARTS. For the second step of PARTS, fitGLS() performs a single GLS for the full dataset, whereas fitGLS\_partition() analyzes partitions that can be created with the function sample\_partition(). In most applications to spatial data, the spatial autocorrelation should be fit with a "nugget" to allow for local (spatially uncorrelated) variation, which is estimated during fitting with fitGLS() and fitGLS\_partition() (making these methods technically Estimated Generalized Least Squares, although we have dropped the "Estimated" as is commonly done). Spatial autocorrelation can be given different functional forms specified by fitV(). Parameters for spatial autocorrelation can either be obtained from the residuals of the time-series analyses using fitSpatialcor() or fit during the spatial GLS using fitGLS\_opt(); the latter is necessary when performing analyses on purely spatial data. These seven functions provide users with access to the entirety of the PARTS method. The package also contains additional tools for more options, fine-scale control over methods, and additional functionality (ref to vignette).

**Relationship to other methods**

Methods that can be used to analyze large spatiotemporal datasets have arisen both from time-series analyses and from spatial analyses. Methods developed to analyze multiple time series {e.g.`, \Tsay, 2014 #11756;Holmes, 2012 #11757;Ives, 2003 #801;Harvey, 1989 #1208} can be extended to the case of multiple time series on a map by specifying spatial correlations between them. Nonetheless, these methods are designed for data in which the temporal dimension (*T*) is large relative to the spatial dimension (*n*). In contrast, spatiotemporal methods arising from spatial methods such as kriging are better suited for data with large spatial dimension, such as pixels on a map. Numerous approximations have been developed to make it possible to fit equation (1) simultaneously in both temporal and spatial dimensions {Wikle, 2019 #11755;Finley, 2012 #11751;Kang, 2010 #11735}. With existing R packages INLA {Krainski, 2019 #11761} and FRK {Zammit-Mangion, 2018 #11760}, it is possible to analyze quite large datasets. As an extreme, the EUSTACE project aims to estimate daily weather data since 1850 for the globe at a resolution of 0.25 degrees, which involves estimating roughly 1011 values, although numerous simplifications, and lots of computing power, are needed {Rayner, 2020 #12540;Lindgren, 2022 #12541}.

remotePARTS differs from other approaches for analyzing spatiotemporal data in both its primary goal and simplicity. The spatiotemporal methods arising from spatial statistics focus primarily on smoothing, interpolation, and extrapolation to points in space and time for which data have not been collected. An archetypal example is estimating the global distribution of CO2 concentrations using data consisting of 100,000-300,000 point samples per day from the NASA OCO2 satellite {Zammit-Mangion, 2021 #12542/`, Figs 21`, 22}; the statistical problem is to interpolate across space and time from samples taken during repeated passes along a polar orbit that covers the globe roughly every 16 days. In contrast, remotePARTS is designed for regression problems such as whether the rate of greening (NDVI) inferred from satellite images over the last 30 years has been greater in one land-cover class than another. This question can be posed as a regression in the form of equation (1) by letting *yi*(*t*) be the annual average NDVI for year *t* in pixel *i*, and including an interaction term q1 *t*\**xi* in which *xi* is a (0,1) factor denoting whether pixel *i* is in land-cover class 1:

*yi*(*t*) = *b*0,*i* + *b*1*xi* + (q0 + q1*xi*)*t* + *i*(*t*) (Eq. 2)

*i*(*t*) = r*i**i*(*t*–1) + *i*(*t*)

*i*(*t*) ~ N(0, 2****(**D**))

The statistical test for whether greening is occurring more rapidly in land-cover class *x* = 1 is based on the coefficient q1. Because remotePARTS focuses only on the regression coefficients, it does not give predictions beyond those inferred from the estimates of the regression coefficients (e.g., that the rate of greening depends on q1 *xi*).

The goal of only estimating regression coefficients simplifies the analyses for remotePARTS. To fit the model given by equation (2), it is necessary to estimate all parameters simultaneously for the entire dataset, and this requires assumptions about parameters other than q1. For example, NDVI is affected by precipitation, which is lower around 30 degrees latitude than either closer to the Equator or poles, and is affected by numerous other variables such as elevation that have strong autocorrelation. Therefore, the intercept *b*0,*i* should not be treated as constant among all pixels, because *b*0,*i* itself is spatially autocorrelated in a way that cannot be accounted for by the spatiotemporal autocorrelation *i*(*t*). Similarly the temporal autocorrelation in *i*(*t*) given by r*i* might differ among pixels {see \Ives, 2021 #12495}, making it necessary to incorporate spatial autocorrelation in the strength of temporal autocorrelation. remotePARTS greatly simplifies this problem by fitting time series for each pixel separately, thereby reducing the model in equation (2) to

*yi*(*t*) = *c*0,*i* + *c*1,*i t* + *i*(*t*) (Eq. 3)

*c*1,*i* = q0 + q1*xi* + g*i*

g*i* ~ N(0, g2**g**(**D**))

where the spatial model is a regression of the coefficients from the pixel-level time-series analyses, *c*1,*i*, against *xi*. The costs of this approach are that (i) analyzing each time series separately as if they were independent does not leverage information from surrounding pixels to give better (e.g., true maximum likelihood) estimates, and (ii) information about the spatiotemporal dynamics is "thrown away" because only one parameter from the time-series analyses, *c*1,*i*, are retained. The advantages of this approach, however, come from not having to specify the full spatiotemporal model, which makes remotePARTS robust against mis-specification of the full model and computationally easier.

**Simulation study of remotePARTS**

We performed six simulation studies to investigate the performance of remotePARTS (Tables 2, 3). Two studies (Table 2; i, iii) simulated data with the same model used to fit the data, thereby giving information about the accuracy of the parameter estimators. Three studies (ii, iv, v) simulated data with a model different from that used to fit the data in order to investigate the robustness of remotePARTS to model mis-specification. Three studies (i-iii) addressed only spatial data, while three studies (iv-vi) addressed spatiotemporal data. Although remotePARTS was designed primarily for spatiotemporal datasets, the studies using only spatial data both demonstrate its ability to analyze spatial data and give useful illustrations of its performance characteristics. The sixth study (Table 3) compared the performance of remotePARTS against a "gold standard" statistical model identical to the simulation model and fit with REML. We present the simulation studies along with their results below, building from the simplest to most complex, and at each step we only describe changes from the previous study.

*i. Spatial data*

To investigate the effects of spatial extent and spatial autocorrelation on the performance of remotePARTS (Table 2, case i), we simulated data on a square grid containing 1042, 1442, 2002, or 2802 locations (pixels) consisting of two classes (*xi* = 0 or 1) in a 44 checker-board pattern (Figure 1). Spatial variation given by the random error d*i* was Gaussian, with variance s2 = 1 and correlations among locations *i* and *j* given by exp(-*dij*/*r*) where *dij* is the distance between *i* and *j*, and *r* is the "range" parameter that scales the extent of spatial autocorrelation. Distances *dij* were scaled to make the maximum distance between locations equal to one. When varying grid size we used *r* = ??, while for the grid with 1042 locations, we performed simulations with *r* = 0, 0.05, and 0.25. When fitting simulation data, we estimated not only the regression coefficients but also range *r* and nugget using fitGLS\_opt().

In the simulations, the estimates for the effects of the classes *xi* = 0 and 1 on *yi* (q0 = 0 and q1 = 0.2) were unbiased (Table 2, case i). With increasing spatial extent and decreasing spatial autocorrelation, the estimates of the coefficients became more precise (lower standard deviations). Finally, for the case of the grid with 1042 locations, we performed the simulations 500 times to make it possible to assess type I error rates: as hoped for, in roughly 5% of the simulations, the hypotheses that q0 = 0 and q1 = 0.2 were rejected at the significance level of alpha = 0.05.

*ii. Spatial data with non-Gaussian errors*

Case ii involved simulations similar to case i, except data were simulated in which the random errors were given by a *t*-distribution with 3 degrees of freedom. The *t*3 distribution has fatter tails (positive kurtosis) than a normal distribution. Applying the same estimation model as in case i for a grid with 1042 locations, there was no bias in the estimates, and the precision was similar to that when simulated with a Gaussian distribution (case i). Furthermore, type I error rates were not inflated, showing that type I error rates are robust to non-Gaussian random errors.

*iii. Spatial data with latent spatial autocorrelation*

A common challenge when analyzing large spatial datasets is the confounding effects of unmeasured variables. We simulated spatial data with a latent variable *zi* as a 2-dimensional sine wave,

(Eq. 4)

where *k* and *l* are the vertical and horizontal positions of location *i*, and *N* is the number of cycles on the grid (Fig. 1b-d). When variation in the latent variable *z* was either coarser (Fig. 1b, *N* = 1) or finer (Fig. 1d, *N* = 9) than the spatial variation in classes *x*, the model was able to estimate the coefficients q0 and q1 with little bias and precision similar to that found without the latent variable (Table 2, compare case iii with i and ii). However, when scale of variation in the latent variable *z* was similar to that for classes *x* (Fig. 1c, *N* = 4), estimates of q0 and q1 were biased. This shows the unsurprising result that if the variable under analysis covaries with an unmeasured latent variable, the estimates for the effects of the measured variable will be confounded.

4. Spatiotemporal data

To simulate spatiotemporal data, we assumed that classes *xi* affect not only the mean value of *yi*(*t*) at location *i*, but also the change in *yi*(*t*) as a linear function of time. We set the goal of the analysis to estimate the time trends associated with classes *x* = 0 and 1, coefficients q0 and q1 (Table 2 case iv). We simulated data for *T* = 30 time points on a grid with 1042 points, with the random error given by equation (1) in which there is both autocorrelation through time (r = ?) and in space (2****(**D**)). We estimated q0 and q1 using the full remotePARTS two-part procedure, estimating *r* from the residuals of the time-series analyses using fitCor() and the nugget during the spatial analysis using fitGLS\_partition().

Estimates of the time trends q0 = 0 and q1 = 1/30 were unbiased (Table 2, case iv), and precision decreased with increasing spatial autocorrelation (*r* = 0, 0.05, and 0.25). Nonetheless, the was no apparent inflation of type I errors. Thus, even though remotePARTS reduced the temporal dimension of the spatiotemporal data for the spatial analysis, estimation was still biased and type I errors appropriate.

5. Spatiotemporal data with latent spatiotemporal autocorrelation

In case v we added a latent variable *ui*(*t*) to the simulation model, with the goal of assessing the robustness of remotePARTS to mis-specification of the spatiotemporal patterns in the data. We assumed that *ui*(*t*) was a spatiotemporal random variable having the same form as e*i*(*t*) (Table 2, case v) but with different temporal (r*u* = 0 or 0.4) and spatial (*ru* = 0 or 0.4) autocorrelation. [I'm not positive what the results are beyond the comments in your original Table, so I'm leaving this blank for you to fill.] These results show that remotePARTS is robust to spatiotemporal latent variables in the processes generating data.

*vi. Comparison with full spatiotemporal GLMM model*

The two-step strategy of remotePARTS, performing time-series analyses on separate time series and then analyzing coefficients from the time series with a spatial model, discards information. Therefore, remotePARTS might be expected to have low statistical power to detect associations with explanatory variables inferred from regression coefficients. Because we expected this loss of information to have the greatest effect on statistical power for small datasets, and to speed computations, we performed 1000 simulations on a 8x8 grid for 30 time points using the model given by equation (2). We then fit the simulated datasets with remotePARTS using fitGLS(), rather than fitGLS\_partition(), because the small dataset did not need to be partitioned. For comparison, we fit the same datasets with a GLMM having exactly the same form as the model used to simulate the datasets; for fitting, we modified pglmm() in the R package `phyr` {Li, 2020 #12305} to include the spatiotemporal random error e*i*(*t*) {Ives, 2010 #3948} and used REML fitting.

remotePARTS and the GLMM showed almost identical results, with both showing little bias except for large simulation values of q1 (Table 3). The type I error rates (when the true value of q1 = 0) were slightly low for both methods, implying that the approximated *P*-values given by the methods were slightly too high. Surprisingly, the power of remotePARTS (the ability to reject the null hypothesis that q1 = 0 when in fact it is false) was similar between methods, with both methods rejecting the null hypothesis in ~92% of simulated datasets when q1 = 0.75.

**Discussion**

remotePARTS provides a robust method for performing regression analyses using very large spatial and spatiotemporal datasets. The robustness, flexibility, and computational speed of the method comes from focusing on the regression problem. For regression, it is possible to perform spatiotemporal analyses by first separately fitting time series and analyzing the fitted time-series parameters in a spatial model. Furthermore, the spatial analyses can be partitioned, with test statistics computed separately from all partitions then being stitched together using information about their covariance. In comparison with a full spatiotemporal GLMM, remotePARTS has good statistical power to identify statistically significant coefficients, even for small datasets. Although remotePARTS is not designed to predict values of the response variable and therefore cannot be used for smoothing, interpolation or extrapolation, it nonetheless makes it possible to investigate relationships among variables in very large spatiotemporal datasets.

**Acknowledgements**

This work was supported by NASA-AIST [80NSSC20K0282] to ARI, Volker C. Radeloff, Fangfang Wang, and Jun Zhu, and NSF [DEB-1556208] to ARI.

**References**

Finley, A.O., Banerjee, S. & Gelfand, A.E. (2012) Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes. *Journal of Geographical Systems,* **14,** 29-47.

Harvey, A.C. (1989) *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, U.K.

Holmes, E.E., Ward, E.J. & Wills, K. (2012) MARSS: Multivariate autoregressive state-space models for analyzing tme-series data. *The R Journal,* **4,** 11-19.

Ives, A.R., Abbott, K.C. & Ziebarth, N.L. (2010) Analysis of ecological time series with ARMA(p,q) models. *Ecology,* **91,** 858-871.

Ives, A.R., Dennis, B., Cottingham, K.L. & Carpenter, S.R. (2003) Estimating community stability and ecological interactions from time-series data. *Ecological Monographs,* **73,** 301-330.

Ives, A.R., Zhu, L., Wang, F., Zhu, J., Morrow, C.J. & Radeloff, V.C. (2021) Statistical inference for trends in spatiotemporal data. *Remote Sensing of Environment,* **266,** 112678.

Ives, A.R., Zhu, L., Wang, F., Zhu, J., Morrow, C.J. & Radeloff, V.C. (2022) Statistical tests for non-independent partitions of large autocorrelated datasets. *MethodsX,* **9**.

Kang, E.L., Cressie, N. & Shi, T. (2010) Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics-Revue Canadienne De Statistique,* **38,** 271-289.

Krainski, E.T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. & Rue, H. (2019) *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC Press/Taylor and Francis Group.

Li, D., Dinnage, R., Nell, L.A., Helmus, M.R. & Ives, A.R. (2020) phyr: An r package for phylogenetic species-distribution modelling in ecological communities. *Methods in Ecology and Evolution,* **11,** 1455-1463.

Lindgren, F., Bolin, D. & Rue, H. (2022) The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics***,** 100599.

Rayner, N.A., Auchmann, R., Bessembinder, J., Brönnimann, S., Brugnara, Y., Capponi, F., Carrea, L., Dodd, E.M.A., Ghent, D., Good, E., Høyer, J.L., Kennedy, J.J., Kent, E.C., Killick, R.E., van der Linden, P., Lindgren, F., Madsen, K.S., Merchant, C.J., Mitchelson, J.R., Morice, C.P., Nielsen-Englyst, P., Ortiz, P.F., Remedios, J.J., van der Schrier, G., Squintu, A.A., Stephens, A., Thorne, P.W., Tonboe, R.T., Trent, T., Veal, K.L., Waterfall, A.M., Winfield, K., Winn, J. & Woolway, R.I. (2020) The EUSTACE Project: Delivering Global, Daily Information on Surface Air Temperature. *Bulletin of the American Meteorological Society,* **101,** E1924-E1947.

Tsay, R.S. (2014) *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, Inc., New Jersey, USA.

Wikle, C.K., Zammit-Mangion, A. & Cressie, N. (2019) *Spatio-temporal statistics with R*. Taylor and Francis Group, Boca Raton, FL.

Zammit-Mangion, A. & Cressie, N. (2018) FRK: An R package for spatial and spatio-temporal prediction with large datasets. *arXiv*.

Zammit-Mangion, A. & Cressie, N. (2021) FRK: An R Package for Spatial and Spatio-Temporal Prediction with Large Datasets. *Journal of Statistical Software,* **98,** 1 - 48.

**Figures and Tables**

A picture containing crossword puzzle, text, indoor, tiled

Description automatically generated

Figure 1: Distributional pattern of land-cover classes .

**Chart

Description automatically generated**Chart

Description automatically generated**Chart

Description automatically generated**

Figure 2 Fixed spatial variation given by 2D sin wave. The wave was generated with 1 (left), 4 (middle) or 9 (right) cycles per map.