**Title:** remotePARTS: a robust toolset for statistical analysis of spatial and spatiotemporal datasets

**Abstract**

1. “Set the context for and purpose of the work”: Many ecological systems exhibit spatial and temporal variation. With this variation comes spatiotemporal autocorrelation. Any statistical toolset meant to address these problems must be able to adequately account for this autocorrelation.
2. “Indicate the approach and methods”: PARTS (Partitioned Autoregressive Time Series analysis) can be used to conduct map-scale hypothesis tests of spatiotemporal systems in a statistically valid way. remotePARTS is 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. “Outline main results”: We found that PARTS is a robust and accurate statistical approach for testing a variety of hypotheses. We found that PARTS performed well at testing for effects of spatial variables, temporal variables, and spatiotemporal variables on spatial and spatiotemporal responses.
4. “Identify conclusions and wider implications”: These results demonstrate that remotePARTS and the PARTS method can be applied to any problem in which spatial and temporal variation occur. We suggest that community ecology, population dynamics, and genetics are among the many ecological disciplines that might benefit from the PARTs method.

**Introduction**

Many ecological problems involve space and time. How does a species or population use habitat types and how has that utilization changed through the decade? Are fires becoming more prevalent in certain regions than others? Are changes in plant phenology, driven by increasing global temperatures, more pronounced in some areas and are these effects compounding through time? These example questions all illustrate the importance of understanding spatiotemporal systems from a statistical perspective. Ecologists and Earth scientists are constantly addressing such questions, but very few 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 in both dimensions of space and time; nearby locations are exposed to similar environmental conditions and those conditions tend to change slowly over time. It is critical that the methods used to answer questions in spatiotemporal systems account for this spatiotemporal autocorrelation.

In its simplest form, a spatiotemporal variable at time is generated by some process given by *equation (1)*. In this formulation, the value of at each of *n* locations is a function of time , which is a *q*-length vector of timepoints. is a design matrix with *k* columns, corresponding to each independent spatial variable, and *n* rows. is a *k*-length vector of spatial effects. is a *k*-length vector of interaction effects between each spatial variable and time. The stochastic component contains the spatiotemporal autocorrelation. This error term can be thought of as the result of a temporally autoregressive (AR) process with spatially correlated innovations. Correlations among these innovations, then, are a function of the distances between locations. Many methods of analyzing spatiotemporal datasets do not adequately account for both types of autocorrelation (refs).

|  |  |
| --- | --- |
|  | (1) |

PARTS (Partitioned Autoregressive Time Series) is a two-part approach to statistical inference for spatiotemporal datasets that can account for spatiotemporal autocorrelation (Ives et al 2021a). The first part of the method collapses temporal variation at each location via time trend regression analysis. The correlations among the regression residuals and distances among each pixel are used to estimate spatial correlation structure. The second part uses generalized least squares (GLS) to regresses the collapsed temporal variation onto the independent variables while accounting for the spatial correlation structure. In this way, both spatial and temporal variation are utilized by the model. For many applications where spatiotemporal variation is relevant, such as remote sensing, datasets are extremely large. In these cases, computational complexity prevents PARTS from being applied to the full dataset at once. The solution to this problem is partitioning. By breaking the map into smaller pieces, estimating parameters from each piece, and performing a single test on the collection of results, PARTS can handle datasets that would be otherwise infeasible. For example, to estimate from *equation (1)* for a large map, the first step is to estimate a time trend for each location , as in *equation (2)*. Next, estimate the covariance matrix from the correlations () and distances () among each location. The dataset is then randomly subset into *p* partitions and   is estimated, from the GLS in *equation (3)*, for each partition. Finally, results for all partitions are combined into a single test for the entire dataset. PARTS performs well under these assumptions (Ives et al 2021), but further evaluation is needed to determine its usefulness under various conditions.

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

Our R package, remotePARTS (github ref), provides the tools for implementing PARTS with any spatial or spatiotemporal datasets (Table 1). Two functions are provided for collapsing temporal variation; fitCLS() or fitAR() will conduct time series analyses using either conditional least squares (CLS) or AR REML, respectively. External temporal trend analyses can also be used in place of those provided. The next two functions provide tools to estimate spatial correlation structure from the data. fitSpatialcor() estimates the spatial correlations among time series residuals and fitV() fits a covariance matrix from these correlations and distances between points. GLS functionality resides in fitGLS(), which applies the GLS step on a full dataset while the alternative fitGLS.partition() conducts the step on partitioned datasets. The function sample.partition() is also included to help users create these partitions from their data. Both fitGLS() and fitGLS.partition() provide map-scale hypothesis test results to answer ecological questions. These 7 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).

Table 1: Main tools contained within remotePARTS

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PARTS step** | **remotePARTS function** | **Description** |
| Part 1 | Temporal collapsing | fitCLS(), fitAR() | Fit time-series regression to a location and return estimate of time trend coefficient |
| Estimate spatial parameters  (from residuals) | fitSpatialcor() | Find ML estimates of spatial parameters by comparing residual correlations with distances among points |
| Fit covariance matrix | fitV() | Fit a covariance matrix from distances among points and spatial parameter estimates. |
| Part 2 | Fit GLS (small datasets) | fitGLS() | Fit GLS to data, given covariance parameters |
| Partition dataset (large datasets) | sample.partition() | Produce a random partition matrix containing indices to locations in an *n*-length dataset. |
| Fit partitioned GLS (large datasets) | fitGLS.partition() | Fit GLS to partitioned data, given covariance parameters |
| Other | Combined covariance estimation and GLS | optimize\_GLS() | Estimate spatial parameters from data rather than residuals, fit covariance matrix, and fit GLS. Primarily used as alternate method for spatial parameter estimation in certain contexts. |

Here, we aim to evaluate the performance of PARTS and remotePARTS by conducting simulation studies designed to push the method to its limits. This is primarily to validate whether PARTS can be extended beyond its perceived limitations. Our goals are to: 1) Evaluate the suitability of PARTS for testing spatial hypotheses, absent temporal variation, under both Gaussian and non-Gaussian error structures; 2) evaluate the accuracy of PARTS when data are generated with an unknown fixed spatial autocorrelation component; 3) evaluate the accuracy of PARTS to detect temporal trends with varying ranges of spatial autocorrelation; 4) evaluate the accuracy of PARTS to detect effects of independent spatiotemporal variables on temporal trends – with varying spatial and temporal autocorrelation; and 5) compare PARTS to a ‘gold standard’ model.

**Methods**

To investigate the performance and robustness of PARTS, we conducted a series of simulation studies. Simulations were designed with conditions intended to push the method to its limits and to violate assumptions. Data were simulated using the tools contained in remotePARTS. For most simulations, we generated 104 104-pixel maps with *equation (4)*, which differs from *equation (1)* by the addition of two terms and . The first new term allows for a spatiotemporal independent variable , which is a *nq* design matrix, and its scalar effect . The second new term allows for fixed spatial variation , which is a vector of length *n*, and its scalar effect *.* In these simulations, our independent variable represents land-cover class designations. contains two land classes ( and ) in a 44 checkerboard pattern (Figure 1). Gaussian errors, and , are generated as with spatially correlated innovations , as shown by *equations (5)* and *(6)*. The AR parameter determines the strength of temporal autocorrelation. Spatial correlation structure is given by a tapered covariance function given by *equation (9)*. The parameters of are distances among points, symbolized by distance matrix , and a maximum range of spatial correlation . By altering parameter values, this simulation process allowed us to tailor datasets towards addressing our various goals.

Table 1: Simulation model

|  |  |
| --- | --- |
|  | (4) |
|  | (5) |
|  | (6) |
|  | (7)  (8) |
|  | (9) |

The PARTS method, via remotePARTS, was applied to all simulated datasets to estimate the effect of the various independent variables. To perform the first part of the method, we fit the model given by *equation (10)* to each pixel independently and estimated the parameter of interest. If the parameter of interest was , , or , then we estimated , , and , respectively. For the second part of the method, we split simulated datasets into five random partitions, each with 2000 locations. For these partitions, we then substituted estimated parameters from part one into the GLS, given by *equation (11),* for  . The correlation structure used by the GLS (; *equation (12)*) was estimated with exponential covariance function , given by *equation (13)*. The spatial range parameter was estimated as the maximum likelihood solution to , given distances between pixels , for a random subset of 1000 locations. Partition-level estimates of the interest parameter (represented by in (11)) were then averaged into a map-level estimate. These map-level estimates are reported as the difference between the true parameter value and the estimate (i.e., ) for all sets of simulations. Finally, for a small number of simulation sets, we calculated the Type I error rates. These sets were simulated many more times to improve precision. Significance tests were performed via correlated chi-square tests (ref) of parameter estimates among partitions. This serves as a map-level hypothesis test (Ives et al 2021). Significant tests are reported as a proportion of the total number of simulations in a set. Data were simulated with only the relevant model components for each question. For example, to test spatial estimation properties we generated datasets without temporal components by setting and (Table 2).

Table 2: Fitted model

|  |  |
| --- | --- |
|  | (10) |
|  | (11) |
|  | (12) |
|  | (13) |

Goal (1) was addressed by simulating spatial-only datasets and estimating the effect of land cover . The true effect of land class 0 and 1 was set to 0 ( and 0.2 (), respectively. The impact of map size was evaluated by simulating square maps that were 104, 144, 200, or 280 pixels wide with all else equal. We tested the effect of spatial autocorrelation range by setting to one of 0, 5, or 25% of the map’s width. We also compared the effect of non-Gaussian errors by forcing the spatial errors to follow a studentized t-distribution with three degrees of freedom ( rather than a normal distribution. The distribution has much fatter tails (greater deviation from the mean) than a normal distribution, which could cause poor estimation if strict normality assumptions are required.

Goal (2) was addressed by adding in unmeasured fixed spatial variation and testing the ability to estimate given this variation. The effect of the unmeasured variable was set to be equal in magnitude to the effect of random spatial variation and their combined effect is equivalent to spatial variation in other simulation sets (i.e., ). represents any variable that is spatially autocorrelated for which a researcher fails to account, but which influences the response and potentially overlaps with the predictor variable of interest (i.e., ). As such, was generated as a two-dimensional sin wave across the map with *equation* 14. The number of cycles present in over the course of the map is given by which was set to one of 1, 4, or 9. These 3 cases allow for to overlap with in different ways (Figure 2).

|  |  |
| --- | --- |
|  | (14) |

Goal (3) was addressed by testing the impact of spatial autocorrelation on estimation of the temporal parameter . We simulated data over thirty time points (i.e., ) and allowed the response in land class 1 to increase by 1 over the study’s duration (i.e., ). We also set the spatial autocorrelation range to one of 0, 5, or 25% of the map’s width.

Goal (4) was addressed by adding in the spatiotemporal predictor variable , which can represent climate variables with spatial and temporal dimensions (e.g., temperature, precipitation). The strength of temporal autocorrelation in was set to of 0 or 0.4. The range of temporal autocorrelation was similarly set to one of 0 or 40% of the map’s range.

Goal (5) was addressed by simulating small datasets ( pixels), with spatial and temporal autocorrelation, and fitting both a PARTS model and a ‘gold standard’ model. The ‘gold standard’ is a generalized mixed effects ARMA model, fit using the R package pglmm\_ARMA from the phyr package. The pglmm\_ARMA model contains time-invariant random effects for both the intercept and time trend and spatial correlation among the residuals. [More info about the gold standard method]

**Results**

PARTS accurately estimated parameters of interest in nearly all the simulation studies (Table 2). We found that estimates of spatial parameters were accurate even when no underlying temporal variation occurred, demonstrating that the method can address purely spatial problems as well as spatiotemporal ones. We also found that detection of spatial patterns is independent of map size, with accurate estimates on small and large maps alike. Precision of the estimates, however, did increase with larger maps. PARTS also accurately estimated parameters under both Gaussian and non-Gaussian error structures. This indicates that PARTS is robust to deviations from normality assumptions of regression analyses. The method is also insensitive to the magnitude of spatial autocorrelation; spatial and temporal parameter estimates were accurate at all ranges of spatial autocorrelation. We also show that PARTS performs incredibly well when compared to the ‘gold standard’ (Table 3) and has expected levels of Type I error (Tables 2 and 3).

We were also able to detect the effect of an independent variable that was itself spatiotemporally autocorrelated. PARTS accurately estimated the variable’s impact on the response. These results were insensitive to the magnitude of either spatial or temporal autocorrelation in the independent variable. This indicates that PARTS can be expanded to address problems beyond its original scope and is robust to the additional components.

In the case of an independent spatial variable that is unmeasured by or unknown to the researcher, but which influences the response variable, PARTS did not exceed expectations. When this fixed source of spatial variation was confounded with land class, the model performed poorly. When , The peak of the wave overlapped primarily with , leading to overestimates of this class’s effect. Conversely troughs aligned with , causing PARTS to overestimate this class’s effect. The opposite case was true when . In these cases, troughs aligned with and peaks aligned with causing PARTS to underestimate and overestimate effects, respectively. Because peaks and troughs were more evenly distributed among land classes when , PARTS performed better in this case, though estimates were still statistically different from the true values.

Table 2 Simulation studies (Goals 1-4)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Goal** | **Simulation** | **Constant parameters** | **Cases** | **N Sims** | **Value of interest (mean +/- 95% CI)** | **Conclusions** |
| 1 | Map size | ;  ; ; |  | 200 |  | Accurate estimates for all map sizes.  Precision increases with map size. |
|  | 200 |  |
|  | 200 |  |
|  | 200 |  |
| Spatial-only (Gaussian) | ;  ; ; |  | 500 |  | Accurate estimates for all values.  No inflated Type I Error (0.047).  Parts can be used to detect effects of a spatial variable, absent temporal variation. |
|  | 500 |  |
|  | 500 |  |
| Spatial-only  (Non-Gaussian) | ;  ; ; |  | 1000 |  | Accurate estimates for all values.  No inflated Type I Error (0.047).  PARTS is robust to violation of normality. |
|  | 500 |  |
|  | 500 |  |
| 2 | Unknown fixed spatial variation | ;  ; ;  ; |  | 200 |  | Poor estimation when unmeasured spatial variation is confounded with independent variables.  Spatial variation that is unaccounted for by the model results in poor fit. |
|  | 200 |  |
|  | 200 |  |
| 3 | Temporal predictor | ;  ; ;  ; |  | 200 |  | Accurate estimates for all values.  PARTS is robust to wide ranges of spatial autocorrelation. |
|  | 200 |  |
|  | 200 |  |
| 4 | Spatiotemporal predictor | ;  ; ;  ; |  | 200 |  | Accurate estimates for all combinations of and .  PARTS can be used to test for the effect of a spatiotemporal predictor variable. |
|  | 200 |  |
|  | 200 |  |
|  | 200 |  |

Table 3 Gold standard comparison (Goal 5)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Goal** | **Simulation** | **Constant parameters** | **Cases** | **N Sims** | **Estimate**  **(PARTS)** | **Estimate**  **(pglmm)** | **Rejection**  **(PARTS)** | **Rejection**  **(pglmm)** | **Conclusions** |
| 5 | Power (PARTS vs. Gold Standard) | ;  ; |  | 1000 | -0.001 | -0.002 | 0.042 | 0.033 | PARTS performs extremely well when compared to the ‘gold standard’.  PARTS performs well even on small maps. |
|  | 1000 | 0.239 | 0.242 | 0.172 | 0.157 |
| .50 | 1000 | 0.492 | 0.494 | 0.612 | 0.612 |
| .75 | 1000 | 0.736 | 0.736 | 0.916 | 0.919 |

**Discussion**

PARTS is an extremely robust and statistically rigorous approach to testing spatial and spatiotemporal hypotheses and the R package remotePARTS provides the tools to apply these methods to a wide variety of ecological problems. The method accurately models spatial and spatiotemporal processes and is robust to deviations from assumptions of normality. In addition to accurate parameter estimation, hypothesis testing with remotePARTS is extremely reliable and efficient: it performs as well as a more theoretically appropriate model with less computational overhead. PARTS also demonstrated high detection power without inflated Type I error rates. Overall, these results suggest the potential for broad and diverse applications of remotePARTS, without strict statistical limitations.

The primary limitation of remotePARTS, is shared by regression and other modelling techniques: it is unable to account for unmeasured confounding variation. Confidence intervals of regression coefficients, for example, account for uncertainty attributable to stochastic variation (e.g., sampling error) but cannot assess confounding uncertainty (refs, Knaeble et al. 2020 – Epidemiologic methods). Similarly, remotePARTS and the PARTS method generally, cannot adequately assess confounding variation. As with regression, coefficient and confidence intervals and coefficient estimates will always be skewed by such affects. Therefore, it is crucial for researchers to evaluate their system fully and consider all biotic and abiotic factors that may influence their response of interest for inclusion in the models.

PARTS and its corresponding software have already been used to answer a variety of ecological questions but there are still many more application opportunities. Primarily, PARTS has been used by our collaborators in the field of remote sensing. Ives et al (2021) found that, despite previous studies to the contrary, there is no statistical evidence that NDVI is increasing in any continents but Asia and Europe. Even in Asia and Europe, patterns were only statistically valid for a few land-cover classes. Our group also demonstrated that, while there appears to be patterns of increased greening at northern latitudes in Alaska, the relationship is not statistically sound (Ives et al, remotePARTS vignette). … [Other examples] … These methods could also be applied to community ecology and population dynamics, geography, fisheries and wildlife management, genetics, and any other problem in which spatial and temporal variation occur.

**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.