The level of spatial autocorrelation (SAC) present in model residuals can have a detrimental influence on statistical inferences in geographical ecology. There has been significant progress in understanding the nature and causes of residual SAC and developing new methods to reduce the SAC. However, ecologists have not yet found a suitable answer as to when the level of residual autocorrelation is likely to magnify and whether its shift through spatial regression (i.e. the difference in the residual SAC produced by non-spatial and spatial approaches) can be predicted using a set of readily available variables. In this paper, I reanalyzed the outcomes reported by Bini et al. (2009), who originally compared the differences in the standardized regression coefficients resulting from non-spatial ordinary least squares and various spatially explicit methods. Although these researchers were unable to identify reliable predictors of the coefficient shifts, in the present study, I observed that the level of residual autocorrelation significantly and linearly increased with an increase in the magnitude of the SAC inherently possessed by, especially, response and explanatory variables. Moreover, the amount of shift in the residual SAC varied as a linear function of the response and explanatory variables. These findings imply that, in general, specific conditions exist under which there is an increase in the magnitude of the residual SAC, that such a shift can be predicted by the nature and structure of the data sets involved, and that, after all, the residual SAC varies systematically regardless of the selection of the data type, statistical method and spatial scale. I suggest that the level of the SAC present in the input variables can directly indicate how much improvement (i.e. reduction in the residual SAC) a non-spatial model will experience when a proper spatial approach is employed.

Keywords: biogeography, macroecology, model improvement, spatial modeling, spatial structure, species distribution modeling

Introduction

Spatial autocorrelation (SAC) is one of the most puzzling yet important issues in geographical ecology (Legendre 1993, Beale et al. 2010, Peres-Neto and Legendre 2010, Hawkins 2012). Many researchers have struggled to understand the nature and causes
of SAC in macroecological data and devised new methods to account for the effects of SAC during spatial modeling (Dormann et al. 2007). Although substantial progress has been made with respect to these efforts over the last two decades, a couple of related questions remain that deserve further attention.

It is commonly known that, for a given set of variables in a multiple regression framework, the outcomes of a model that explicitly incorporates SAC differ from those of another model that does not do so, with regard to, for example, partial regression coefficients, the level of residual autocorrelation, coefficient of determination (R²) and Akaike's information criterion (Keitt et al. 2002, Dormann 2007, Kühn 2007, Kissling and Carl 2008). This statement appears simple and self-evident, but its implications are never that simple. First, it is still elusive which of the outcomes from spatial and non-spatial methods we should select with confidence to infer and predict ecological patterns and processes. In a more straightforward manner, are spatial approaches always better and more desirable than their non-spatial counterparts? Second, related to this, there is still a critical paucity of agreement in the literature as to when the differences in these outcomes will be significant or negligible. Numerous studies have reported the differences that occur if SAC is appropriately considered (Fuller and Enquist 2012, Miller 2012, Record et al. 2013, Swanson et al. 2013, Domisch et al. 2019), but little is known about their magnitude.

These points have been widely discussed in the context of the coefficient shifts, which, since 2000, can be referred to as the ‘red herrings’ debate. Lennon (2000) initiated this discourse arguing that, in the conventional ordinary least squares (OLS), the importance of environmental predictors that are strongly autocorrelated over space can be overstated, such that the associated standardized coefficients are inflated and the errors are underestimated. Supporting this idea, several empirical studies have shown that the magnitude of these coefficients substantially decreased after incorporating SAC during spatially explicit modeling (Lichstein et al. 2002, Tognelli and Kelt 2004, Kim 2018). Furthermore, Václavík et al. (2012) proposed that predictor variables with stronger spatial structures should have weaker effects on the patterns of species distribution than less autocorrelated predictors when explicitly modeling SAC (Kim and Shin 2016). This broadly indicates that the differences in the outcomes of spatial and non-spatial approaches may be sensitive to the degree of SAC inherently possessed in the input variables.

In contrast, a series of studies suggest that SAC does not necessarily generate bias in standardized coefficients, that previous analyses without taking SAC into account cannot be considered flawed (Diniz-Filho et al. 2003, Hawkins et al. 2007, but see Beale et al. 2007), and that differences in the outcomes identified between spatial and non-spatial regressions can have multiple causes other than spatial autocorrelation (e.g. multi-collinearity, non-stationarity and scale-dependency) (Diniz-Filho et al. 2007). These views culminated in the study of Bini et al. (2009), who analyzed 97 widely divergent empirical data sets from various parts of the world, finding no reliable predictors of coefficient shifts. They concluded that changes in standardized coefficients depend on the spatial methods used and are largely idiosyncratic. In summary, these discussions imply that ecologists have yet to find the appropriate answers with respect to the nature – especially, the magnitude – of the differences in the model estimates between spatial and non-spatial techniques.

This line of research and debate has also focused on the level of SAC in the residuals (Crase et al. 2012, Gaspard et al. 2019, Lany et al. 2020). Previous studies have stressed that the main statistical problems (i.e. the underestimation of standard errors and inflation of the type I error rate) occur when SAC remains in the model residuals, rather than in the input variables themselves (Anselin 2002, Diniz-Filho et al. 2003, Bini et al. 2009). Moreover, Beale et al. (2007) conducted multiple simulations of spatially structured data sets, demonstrating that the precision of the standardized coefficients produced by the OLS significantly decreased when the residual autocorrelations were strong. If the level of the residual SAC is the real concern, then the following questions deserve investigation:

- Under what circumstances is the magnitude of the residual SAC likely to increase?
- Are there predictor variables that can reliably explain the shift in the residual SAC?
- Overall, can we determine whether there is systematic variation in the residual SAC?

As detailed above, ecologists have failed to propose definitive answers to these questions in the last decade, although there was an initial approach to compare the levels of SAC in the original response variable (bird species richness) and the residuals produced by multiple regression analysis (Hawkins et al. 2003). Significant advances have been minimal since Bini et al. (2009) reported the idiosyncratic and unpredictable nature of coefficient shifts, which Hawkins et al. (2007) had earlier referred to as ‘the worst possible scenario’.

In this study, I suggest that ecologists should switch their attention to the level and pattern of shifts in residual autocorrelation. There is still a slow, but steady, increase in the number of published articles mentioning ‘residual autocorrelation’ in selected journals focusing on macroecology and biogeography (Supporting information). Assuming that this topic will continue to be critical within these disciplines, I address the three questions raised above to generate a predictive framework, such that future researchers can anticipate large or negligible discrepancies in the residual SAC between spatial and non-spatial approaches. Achieving this goal will potentially advance our current fundamental knowledge in geographical ecology one step forward, that is, beyond identifying the influences of SAC, and toward an understanding of how such influences occur and when they tend to magnify. I reanalyze multiple spatial data sets that have been collected in various different parts of the world to obtain general conclusions on the implications that SAC has for macroecological modeling.
Methods

The Bini et al. (2009) data

The data sets analyzed by Bini et al. (2009) provide an excellent opportunity to accomplish the purpose of this study. These data sets are unprecedentedly extensive, covering 97 different regions on Earth. These data sets also do not contain taxonomic or geographic bias and encompass various types of response variables, such as species richness (74 cases), mean body size (14 cases), mean range size (5 cases) and abundance (4 cases). Sample sizes range from 22 to 2403, while the number of explanatory variables varies from 3 to 20. They are not artificially developed or simulated, but consist of real-world data. Bini et al. (2009) were interested in partial regression coefficients, whereas the focus of this paper is on residual autocorrelation. Both papers basically ask whether the differences in model estimates produced by non-spatial OLS and spatial regression can be predicted.

Analysis

In their publication, Bini et al. (2009) developed an Excel spreadsheet. This appendant table contains nearly all the outcomes of their non-spatial and spatial modeling, with all of Moran’s I values calculated using a Gabriel network connection for the first distance class. Reorganizing this existing information, I developed a new spreadsheet composed of three data blocks for this study (Supporting information).

The first data block had nine variables (and 97 sites), consisting of the Moran’s I values for the model residuals produced by non-spatial OLS and the eight spatial methods employed by Bini et al. (2009); SEVM-v1 (spatial eigenvector mapping-variant1), SEVM-v2 (spatial eigenvector mapping-variant2), SEVM-v3 (spatial eigenvector mapping-variant3), LagResp (lagged response), LagPred (lagged predictor), SAR (simultaneous autoregressive), CAR (conditional autoregressive) and MA (moving average) (Supporting information). For detailed explanations of these spatial regressions see Rangel et al. (2006, 2010). This first block represented the level of residual autocorrelation. The second block was created by subtracting the absolute Moran’s I values based on the eight spatial approaches from the absolute Moran’s I values based on the OLS. Hence, this block had eight variables and indicated the shift in the residual autocorrelation. In the second block, a positive value represented a reduction in the magnitude of the residual SAC (i.e. model improvement) due to the consideration of the SAC. To construct the third block, I selected seven out of the eight data characteristics used by Bini et al. (2009) as possible predictors of coefficient shifts: n (number of observations), m (number of explanatory variables), VIF (variance inflation factor), Moran’s Iₜₐ (average Moran’s I value of the explanatory variables), R² (adjusted coefficient of determination of OLS), g₁ (skewness of the residuals of OLS) and F (stationarity). I did not select the residual Moran’s I (i.e. the level of SAC in the residuals of the OLS), which was already included in the first data block. I replaced this data characteristic with a new variable, i.e. Moran’s Iᵥᵥ, indicating the level of SAC inherently possessed in the 97 response variables. This information can be provided by Mauricio Bini upon personal request. As all the data characteristics of the third block were found to be significantly skewed in a preliminary analysis, they were transformed using the appropriate methods (Supporting information).

As an exploratory step, I first performed simple correlation analyses 1) between the eight data characteristics and the level of the residual SAC based on nine regression approaches and 2) between the eight data characteristics and the shift in the residual SAC after eight spatial regressions. Then, as a preliminary analysis detected strong multi-collinearities among the eight data characteristics, I conducted a principal component analysis (PCA) to reduce the data dimensionality. The resulting principal components were included in subsequent regression attempts as explanatory variables to predict the level and shift of the residual autocorrelation. Finally, agreeing with Bini et al. (2009) that the influences of the eight spatial methods on the level and shift of the residual SAC might be correlated, I used two sets of canonical correlation analyses (CCorA) to simultaneously test all methods: 1) between the level of the residual SAC and the eight data characteristics (i.e. data blocks 1 versus 3) and 2) between the shift of the residual SAC and the eight data characteristics (i.e. data blocks 2 versus 3).

Finally, I questioned whether the relationship between the level and shift of the residual SAC and data characteristics tested in this study would also occur in other modeling contexts. I attempted to aggregate the present macroecological data (n = 97) with soil–landform data (n = 39; Kim et al. 2016), water quality–landscape data (n = 86; Miralha and Kim 2018) and socio-economic data (n = 561; Kim and Song 2021), which were collected based on different purposes and sampling designs in previous studies (Supporting information). Among the data characteristics, Moran’s Iᵥᵥ was selected as the representative predictor variable.

Results

All spatial methods reduced the magnitude of the SAC in the residuals based on the OLS with varying degrees (Fig. 1). SAR, LagPred and LagResp were most effective at eliminating and stabilizing the residual SAC, such that their average Moran’s I values were nearly zero and the associated standard deviation values had a small range between 0.10 and 0.12 (Supporting information). In contrast, the residuals of the MA had a low average Moran’s I value (0.11), with the greatest variability among all of the methods considered. The eight data characteristics showed high correlation coefficients with each other in many cases (Supporting information). In particular, half of the 28 correlations were significant at the 0.001 level. These characteristics were divided into three groups based on the outcomes of the PCA
The data characteristics, except g1, proved to be significantly correlated with most SAC in the residuals identified by non-spatial and spatial regressions (Supporting information). On average, Moran’s $I_{DV}$ had the highest correlation coefficient ($r = 0.61, p < 0.001$), but even this characteristic was not significantly correlated with the residual Moran’s $I$ values produced by SAR ($r = -0.03, p = 0.771$). The significance and strength of the correlations substantially increased when the characteristics were correlated with the shifts in the Moran’s $I$ values of the residuals after performing the spatial regressions (Supporting information). All data characteristics, except $R^2$ and g1, exhibited high coefficients that were always significant at the 0.001 level. Again, Moran’s $I_{DV}$ exhibited the greatest average coefficient ($r = 0.73, p < 0.001$).

Similar to these findings, the three PCs significantly explained 1) the Moran’s $I$ values of the residuals yielded by all modeling approaches (except SAR) and 2) the shifts in the Moran’s $I$ values of the residuals (Table 1). The associated $t$ and $R^2$ values of the latter case were greater than those of the former. In addition, in both cases, PC 1 was the most important predictor variable, whereas PC 3 (i.e. g1) was rarely a significant predictor variable.

Based on these results, Moran’s $I_{DV}$ was determined to be the most important data characteristic in this study. Hence, its effect on the level and shift of the residual autocorrelation was visually examined in detail (see the Supporting information for the influence of all the characteristics). Overall, two types of significantly linear relationships were identified. First, the level of SAC in the residuals produced by all regression approaches (except SAR) increased with an increase in the Moran’s $I_{DV}$ (Fig. 2a). Second, the amount of the shift in the residual autocorrelation varied as a function of Moran’s $I_{DV}$ (Fig. 2b). The direction of the shift was often negative, especially when Moran’s $I_{DV}$ was smaller than 0.23. The direction became positive when Moran’s $I_{DV}$ was higher than 0.48.

Figure 1. Box-and-whisker plots of the Moran’s $I$ values of the residuals produced by different regression approaches. For visual convenience, outliers are removed from the graph. (OLS: ordinary least squares; SEVM-v1: spatial eigenvector mapping-variant1; SEVM-v2: spatial eigenvector mapping-variant2; SEVM-v3: spatial eigenvector mapping-variant3; LagResp: lagged response; LagPred: lagged predictor; SAR: simultaneous autoregressive; CAR: conditional autoregressive; MA: moving average.)

(Supporting information). The Moran’s $I_{DV}$, F, n, Moran’s $I_{avg}$ and $R^2$ had the greatest factor loadings for, and thus belonged to, PC 1. The second PC was composed of m and VIF. PC 3 comprised g1. These three PCs explained 77.2% of the total variance of the characteristics.

Table 1. Results ($t$ and $R^2$ values) of the regression analyses using the three principal components as explanatory variables to predict (1) the Moran’s $I$ values of the residuals produced by nine modeling approaches and (2) the shifts in the Moran’s $I$ values of the residuals after eight spatial modeling approaches. The greatest average $t$ values are indicated in bold.

| PC 1 | PC 2 | PC 3 | $R^2$ |
|------|------|------|-------|
| OLS  | 16.3*** | -6.9*** | -1.3 | 0.76*** |
| SEVM-v1 | 11.1*** | -2.6* | -0.3 | 0.57*** |
| SEVM-v2 | 19.1*** | -3.8*** | -1.8 | 0.80*** |
| SEVM-v3 | 12.7*** | -3.1** | -0.2 | 0.64*** |
| LagResp | 7.1*** | -0.1 | 1.1 | 0.33*** |
| LagPred | 4.9*** | 0.4 | 0.7 | 0.18*** |
| SAR | 0.1 | -0.4 | 2.3* | 0.02 |
| CAR | 10.7*** | -4.1*** | -0.9 | 0.58*** |
| MA | 5.6*** | -1.5 | 0.8 | 0.25*** |
| Average | **9.7** | 2.5 | 1.0 | 0.46 |

| PC 1 | PC 2 | PC 3 | $R^2$ |
|------|------|------|-------|
| OLS  | 4.6*** | -6.7*** | 0.3 | 0.40*** |
| SEVM-v1 | 11.3*** | -6.7*** | -0.5 | 0.64*** |
| SEVM-v2 | 15.9*** | -6.3*** | -1.1 | 0.75*** |
| SEVM-v3 | 17.2*** | -6.9*** | -1.9 | 0.78*** |
| LagResp | 13.4*** | -6.1*** | -0.5 | 0.69*** |
| LagPred | 12.2*** | -5.8*** | 0.0 | 0.65*** |
| SAR | 6.9*** | -4.4*** | 0.0 | 0.40*** |
| CAR | 18.6*** | -7.4*** | -1.7 | 0.81*** |
| MA | **12.5** | 6.3 | 0.8 | 0.64 |

OLS: ordinary least squares; SEVM-v1: spatial eigenvector mapping-variant1; SEVM-v2: spatial eigenvector mapping-variant2; SEVM-v3: spatial eigenvector mapping-variant3; LagResp: lagged response; LagPred: lagged predictor; SAR: simultaneous autoregressive; CAR: conditional autoregressive; MA: moving average.

*** Significant at the 0.001 level (two-tailed).
** Significant at the 0.01 level (two-tailed).
* Significant at the 0.05 level (two-tailed).
† The absolute values of $t$ or $R^2$ values of each column were averaged.
Figure 2. The level (a) and shift (b) of the residual autocorrelation as a function of the spatial autocorrelation of the response variables (Moran’s $I_{DV}$). (OLS: ordinary least squares; SEVM-v1: spatial eigenvector mapping-variant1; SEVM-v2: spatial eigenvector mapping-variant2; SEVM-v3: spatial eigenvector mapping-variant3; LagResp: lagged response; LagPred: lagged predictor; SAR: simultaneous autoregressive; CAR: conditional autoregressive; MA: moving average.)
Table 2. Overall results of the canonical correlation analyses, relating (1) eight data characteristics and the level of the residual autocorrelation yielded by nine regression models and (2) eight data characteristics and the shifts in the residual autocorrelation resulting from eight spatial regression models.

| Data groups                                | Canonical variate (CV) | Canonical correlation (R) | Eigenvalue | χ²       |
|--------------------------------------------|------------------------|---------------------------|------------|----------|
| Data characteristics versus level of residual SAC | CV 1: 0.98             | 24.32                     | 505.40***  |
|                                            | CV 2: 0.81             | 1.94                      | 221.94***  |
|                                            | CV 3: 0.66             | 0.76                      | 128.24***  |
|                                            | CV 4: 0.56             | 0.46                      | 78.85***   |
|                                            | CV 5: 0.51             | 0.36                      | 45.05**    |
| Data characteristics versus shift of residual SAC | CV 1: 0.98             | 23.34                     | 422.48***  |
|                                            | CV 2: 0.74             | 1.23                      | 139.96***  |
|                                            | CV 3: 0.54             | 0.41                      | 69.68**    |
|                                            | CV 4: 0.42             | 0.21                      | 39.74*     |

*** Significant at the 0.001 level (two-tailed).
** Significant at the 0.01 level (two-tailed).
* Significant at the 0.05 level (two-tailed).
† Only the CVs that are significant at the 0.05 level are presented.

In addition to Moran’s $I_{sv}$, several other data characteristics, such as $n$, Moran’s $I_{sv}$ and $F$, also had positive and significant relationships with the level and shift of the residual SAC (Supporting information).

The findings of the CCorA helped to identify potential indicators of the level and shift of the residual SAC. Overall, the correlations between the canonical variates (CVs) from two groups of variables were highly significant (Table 2). In the first CCorA predicting the level of the residual SAC (data block 1) using the eight data characteristics (data block 3), each CV 1 extracted 40.3% of the variance in the eight data characteristics and 53.3% of the variance in the Moran’s $I$ values of the residuals. In the second CCorA predicting the shift of the residual SAC (data block 2) using the eight data characteristics (data block 3), each CV 1 extracted 42.5% of the variance in the eight data characteristics and 69.9% of the variance in the Moran’s $I$ values of the residuals. Of greatest importance to this paper, CV 1 summarizing the set of eight data characteristics (data block 3) accounted for 51.2% of the variance in the level of the residual SAC (data block 1) and 67.1% of the variance in the shift in the residual SAC (data block 2). Furthermore, there were close similarities in the loading plots between the two CCorA, each predicting the level and shift of the residual SAC (Fig. 3). On the left column of the figure, Moran’s $I_{sv}$ had the strongest relationship with the first CVs, while the other characteristics (except $g_1$) were also significantly correlated always with the first variables and occasionally with the second variables. On the right column, all spatial methods (except SAR) had significant correlations with the first canonical variates.

Finally, there was a general coherence and order with respect to the extent and pattern of the level and shift of the residual autocorrelation as a positive linear function of Moran’s $I_{sv}$, whether combining the various data sets from the different modeling contexts (Fig. 4) or examining them individually (Supporting information).

Discussion

The results of this study provide potentially new insights into the previous debate on the behavior of model estimates after spatial regression. Overall, in contrast to Bini et al. (2009), who were concerned with the changes in the standardized coefficients, the shift in the residual autocorrelation proved to be systematic (i.e. independent of the data type, statistical method or spatial scale), rather than being idiosyncratic (Hawkins et al. 2007). I note that the various data characteristics tested in this study significantly predicted both the level and shift of the residual SAC produced by various spatial techniques. The findings have several implications for geographical ecology. Answering the three questions raised at the beginning of this paper, I propose that particular circumstances exist in which the magnitude of the residual SAC is likely to increase, that such a shift can be reliably predicted based on the nature and structure of the data sets involved, and that, after all, there is systematic variation in the residual SAC. An exception to these ideas may occur if the Moran’s $I$ values of the residuals are close to zero, with a low variability, as observed in the case of the SAR (Supporting information, Fig. 1). These discussions are bolstered by the fact that the level and shift of the residual autocorrelation varied as a positive linear function of Moran’s $I_{sv}$ in other contexts, such as soil–landform, water quality–landscape and socio-economic modeling (Fig. 4, Supporting information). Hence, the level of SAC present in the various types of response variables directly indicates how much improvement (i.e. reduction of the residual SAC) a non-spatial model will experience if that SAC is appropriately taken into account. I suggest that the predictability of the magnitude of the residual autocorrelation demonstrated in this paper, after more thorough and extensive evaluations using both empirical and simulated data, can be considered an important case for a general theory that explains the effect that SAC has on the outcomes and inferences in geospatial modeling.

Given that the level of SAC remaining in the model residuals is widely known to cause a detrimental effect on statistical inferences (Zhang et al. 2005, Barry and Elith 2006, Segurado et al. 2006, Beale et al. 2010), the present findings – especially, that Moran’s $I_{sv}$ and Moran’s $I_{sv}$ are important predictor variables – can be considered encouraging to ecologists with respect to two aspects. First, we can anticipate and further quantify the magnitude of the residual SAC and its change before performing a spatial
regression simply by examining the internal structure of the input variables. In particular, I suggest that the level of SAC inherently possessed by the response variables can serve as a direct indicator. This is far from the worst possible scenario first proposed by Hawkins et al. (2007) and empirically demonstrated by Bini et al. (2009), in which none of the various data characteristics tested explained the effect of incorporating SAC, such that the model estimates (in their case, standardized coefficients) were difficult to interpret.

Due to such uncertainties, previous studies have even classified spatially explicit techniques as a Pandora’s box, which may generate more confusion than that already exists in the literature (Diniz-Filho et al. 2007). I posit that this is not always the case and provide an optimistic perspective. Spatial methods, as far as the degree and shift of the residual autocorrelation are concerned, yield largely predictable and systematic outcomes regardless of the data type and scale of interest.
Figure 4. Combining the macroecological data with the soil–landform, water quality–landscape and socio-economic data. (Moran's $I_{DV}$: level of spatial autocorrelation in the response variable.)
Second, ecologists do not have to consider the long-standing, dichotomous question as to which of the spatial and non-spatial models is better or more robust (Lichstein et al. 2002, Bahn et al. 2006, Kühn 2007). On one hand, the use of the non-spatial OLS may suffice because conducting spatial regression does not always guarantee a significant improvement to the modeling outcomes. Moreover, previous studies have argued that no SAC remains in the residuals, as long as the OLS includes all of the relevant and correct predictor variables that fully capture the spatial variability of the response variable of interest (Besag et al. 1991, Legendre et al. 2002, but see Beale et al. 2010). In contrast, researchers are generally unable to know a priori the conditions at which the correct variables can be identified, let alone whether they really exist. This is why ecologists are strongly recommended to explicitly verify the presence and influence of SAC in modeling attempts (Dormann 2007, Kim et al. 2016, 2019, Kim 2018, Gaspard et al. 2019). I posit that these opposing perspectives are both acceptable and that the problem essentially does not concern which approach is superior to the other (Bini et al. 2009). Rather, the answer can be obtained from the inherent structure of the data being used in the modeling attempt. For example, if a response variable has a low level of SAC, spatial modeling would not yield a significant difference but remain less parsimonious. The results of this study indicate that spatial methods may not be desirable when the Moran’s $I_{DW}$ is smaller than 0.23 (Fig. 2b), which is likely due to an over-correction effect on the residual SAC (Mauricio Bini personal communication; Diniz-Filho et al. 2005, Chun et al. 2016). In contrast, when analyzing a strongly autocorrelated response variable (especially, when Moran’s $I_{DW} > 0.48$), one should account for the SAC by employing a spatial method. In this latter case, I propose that the benefit (or disadvantage) of (not) accounting for SAC increases linearly with an increase in the level of SAC in the response variable. Synthesizing these discussions, I do not blindly advocate the use of spatial regressions in geographical ecology, nor do I invalidate previous studies that adopted non-spatial approaches (Diniz-Filho et al. 2007). This paper provides a new possibility that ecologists can determine whether they should use spatially explicit methods based on certain threshold values for readily available data characteristics (e.g. Moran’s $I_{DW}$, Moran’s $I_{MC}$ and $F$).

Similar to Bini et al. (2009), I do not support the previous idea that spatial methods, incorporating spatially structured variables (e.g. spatial filters) as additional explanatory variables, differ from the OLS more than methods that aim to model SAC in the residuals (Dormann et al. 2007, Kissling and Carl 2008). Both papers, despite each focusing on different model estimates (i.e. standardized coefficients and residuals, respectively), found that lagged models generated outputs that significantly differed from those of the non-spatial OLS. This study shows that such differences linearly magnify with an increase in the level of SAC in a response variable (Fig. 2b).

Finally, I propose a hypothesis based on the fact that the magnitude of the residual SAC significantly increased in response to the variations in several data characteristics, especially, n, Moran’s $I_{MC}$, Moran’s $I_{DW}$ and $F$ (Supporting information). These characteristics – which all belong to PC 1 (Supporting information) – indicate the size, spatial structure and non-stationarity of a data set. Therefore, when combined, they can further be considered to broadly represent the complexity and uncertainty of the data. I hypothesize that these two properties have the potential to increase the level of SAC in the residuals. The exact meaning of data complexity and uncertainty is yet to be defined clearly. This hypothesis, however, may result in new possibilities for future research. For example, ecologists can better understand why correctly fitted spatial models often show autocorrelated residuals (Beale et al. 2010). I posit that the presence of autocorrelation in the residuals cannot be attributed to a single reason and that resolving the issue of model misspecification does not guarantee the removal of residual SAC. The presence of autocorrelation may not arise solely from model misspecification, but also from the overall complexity and uncertainty of the data. Another open topic for future research is the relationship between residual autocorrelation and model prediction. A spatial method can produce strong autocorrelation in the residuals, but its prediction accuracy might still be high. This situation will increasingly be seen as machine learning approaches advance. Then, a question is whether we should accept a model as robust or not, if it shows both high prediction accuracy and strong residual autocorrelation.

Acknowledgements – Much of the idea presented in this paper has been developed in close communication with Heejun Chang, Yongwan Chun, Guetchine Gaspard, Lorrainey Miralha, Insang Song, and the students of the Physical Geography Lab at Seoul National Univ. Luis Mauricio Bini provided valuable data and advice for this study.

Funding – This study was supported by the National Science Foundation (grant no. 1560907) and the National Research Foundation of Korea (grant no. 2020R1F1A1048427).

Transparent Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.05403>.

Data availability statement

The data analyzed in this study are available in the Supporting information.

References

Anselin, L. 2002. Under the hood: issues in the specification and interpretation of spatial regression models. – Agr. Econ. 27: 247–267.

Bahn, V. et al. 2006. Importance of spatial autocorrelation in modeling bird distributions at a continental scale. – Ecography 29: 835–844.
