PROCEEDINGS B

rspb.royalsocietypublishing.org

Comment

Cite this article: Cardillo M, Bromham L, Greenhill SJ. 2015 Links between language diversity and species richness can be confounded by spatial autocorrelation. *Proc. R. Soc. B* **282**: 20142986.
http://dx.doi.org/10.1098/rspb.2014.2986

Received: 8 December 2014
Accepted: 18 February 2015

Author for correspondence:
Simon J. Greenhill
e-mail: simon.greenhill@anu.edu.au

The accompanying reply can be viewed at http://dx.doi.org/doi:10.1098/rspb.2015.0591.

THE ROYAL SOCIETY PUBLISHING

Links between language diversity and species richness can be confounded by spatial autocorrelation

Marcel Cardillo[1], Lindell Bromham[1] and Simon J. Greenhill[2]

[1]Research School of Biology, and [2]School of Culture, History and Language and ARC Centre of Excellence for the Dynamics of Language, ANU College of Asia and the Pacific, Australian National University, Canberra, Australian Capital Territory 0200 Australia

SJG, 0000-0001-7832-6156

Turvey & Pettorelli [1] present a fascinating study exploring links between biological and linguistic diversity across New Guinea. With the world's highest linguistic diversity (around 900 languages, an average of one language per 1000 km^2 [2]), as well as the high biodiversity characteristic of a large mountainous tropical island, New Guinea is an ideal test case for investigating patterns and drivers of biocultural diversity. Turvey & Pettorelli's finding that numbers of languages and mammal species are correlated across grid cells in New Guinea is consistent with studies in other parts of the world showing similar relationships (e.g. [3]). Globally, languages, like species, show a latitudinal diversity gradient [4], and areas of high language diversity often coincide with hotspots of species diversity [5]. In addition, Turvey & Pettorelli report a surprising negative correlation between the numbers of threatened mammal species and languages considered at risk of extinction. This finding contrasts with previous studies showing that extinction risk in languages and species are positively correlated [5,6].

Turvey & Pettorelli's study is an important contribution to our understanding of the distribution of biocultural diversity, with potential practical implications for conservation. If the spatial distributions of threatened species and threatened languages correspond, then an integrated biocultural management strategy may be possible [6]. On the other hand, such strategies may be less effective if there is a lack of congruence in spatial patterns of diversity. Spatial congruence between total language and mammal diversity could also indicate a functional connection between the two, either a direct causal link, or an indirect link via a third factor that influences both language and species diversity independently. For example, both types of diversity may be enhanced by the same environmental factors [7] if geographical barriers such as mountain ranges, rivers or sea inlets impede gene flow in species as well as human communication [2,8], promoting divergence in both cases. Alternatively, human cultural groups may diversify in response to the diversity of local environments [9,10]. Similarly, spatial congruence in vulnerability to extinction of languages and species may suggest that threatening processes are similar for both human cultures and biodiversity. But Turvey & Pettorelli's negative correlation implies the factors that increase extinction risk in languages are different to those for mammal species.

Statistical tests of association, such as correlations and regressions, are used to detect relationships between variables that are unlikely to arise from random variation, thus implying a functional relationship between the variables. But these tests rely on an assumption of statistical independence between data points. In this case, each grid cell is considered to represent an independent instance of the relationship between language and species diversity. However, this assumption of independence of observations is invalid if either or both forms of diversity are spatially autocorrelated. In fact, it is likely that both language and mammal richness are spatially autocorrelated, because many of the species or languages that occur in a particular grid cell will also occur in neighbouring cells. This means that similarity in richness values for different

Table 1. Spatial autoregressive (SAR) error models for (*a*) log(language richness) and (*b*) log(threatened language richness). Results are shown for two univariate models (mammal richness and mean elevation), and one multivariate model (mammal richness + mean elevation + % land cover per grid cell). For each model, the Akaike information criterion (AIC) for the SAR model and the corresponding non-spatial ordinary least squares (OLS) regression model are given. Asterisk (*) indicates the models that provide a statistically significant fit to the data.

model	predictors	slope	*p*-value	AIC (SAR)	AIC (OLS)
(*a*) log(language richness)					
1	log(mammal richness)	0.10	0.11	326.8	468.5
2	log(mean elevation)	0.0001	0.99	329.1	505.9
3	log(mammal richness)	0.09	0.26	329.0	471.5
	log(mean elevation)	−0.04	0.41		
	% land cover	0.002	0.32		
(*b*) log(threatened language richness)					
1	log(threatened mammal richness)	−0.14	0.08	398.0	527.3
2	log(mean elevation)	−0.12	0.01*	394.2	519.2
3	log(threatened mammal richness)	−0.06	0.46	397.7	519.7
	log(mean elevation)	−0.11	0.04*		
	% land cover	0.002	0.95		

grid cells is at least partly predictable from their spatial proximity alone, which can elevate type 1 errors (false positives) in statistical tests of association, including correlation, regression or other linear models [11].

Furthermore, Turvey & Pettorelli's correlations included coastal grid cells with as little as 25% land area. Because richness increases with area, a grid cell with only 25% land area may have unusually low levels of both species and language richness, which could contribute to a spurious positive association. This is important to investigate, because Turvey & Pettorelli's results could be driven by grid cells containing few languages and low species diversity, with no significant relationship between diversity and risk in grid cells of medium to high language and species diversity.

Here, we investigate whether the spatial associations between mammal richness, language richness and elevation reported by Turvey & Pettorelli are robust to these two potential artefacts. We obtained geographical distributions of mammal species from the Global Mammal Assessment (www.iucn.org) and distributions of the world's languages from the Ethnologue [12]. We extracted all distributions that overlap with the mainland of New Guinea (217 mammal species, 898 languages). We then created a raster grid for New Guinea at a resolution of 0.5° (approx. 50 × 50 km) and calculated the total number of mammal species and languages, and the number of threatened mammal species and languages, within each grid cell. Threatened mammal species and threatened languages were defined using the same criteria as Turvey & Pettorelli [1]. We also calculated mean elevation for each grid cell, using data from the STRM 90 m Digital Elevation Database.

To analyse spatial congruence patterns, we first fitted simple Pearson correlations among log-transformed richness and elevation values, across grid cells, to compare with the results of Turvey & Pettorelli [1]. We then fitted a linear model that predicts log(language richness) from log(mammal richness) and tested for spatial autocorrelation in the model residuals, using Moran's *I*. This test indicated significant spatial structure in the residuals (Moran's $I = 0.05$, $p < 0.0001$), necessitating the use of methods that account for spatial autocorrelation to test for associations between variables. We performed two kinds of test. First, for direct comparability with the Pearson correlations used by Turvey & Pettorelli, we performed correlations with significance tested using Dutilleul's modified *t*-test, which uses an effective sample size computed from the spatial covariance matrix [13]. Second, we explored multivariate models using simultaneous autoregressive (SAR) error models [11]. We chose this method over other kinds of SAR models because its underlying assumption that spatial autocorrelation exists in both predictor and response variables seemed most appropriate for these data. Moran's *I* tests on model residuals confirmed that this method adequately removed the effects of spatial autocorrelation. We used SAR models to test for univariate associations between language and mammal richness, threatened language and threatened mammal richness, and between mean elevation and each richness variable. We then fitted multivariate models predicting language richness from mammal richness, mean elevation and the proportion of land area per grid cell; and threatened language richness from threatened mammal richness, mean elevation and proportion of land area. All geographic information system (GIS) procedures were done using functions in the R packages 'sp', 'rgdal', 'rgeos', 'raster', 'fossil' and 'worldmap'. Dutilleul's modified *t*-tests were implemented in the 'SpatialPack' package, and SAR models in the 'spdep' package.

When we assume independence of observations by using Pearson correlations, we obtain similar results to Turvey & Pettorelli. Mean language richness across the 256 grid cells is 5.96 (range: 0–43, s.d. 5.22), mean mammal species richness is 36.12 (2–103, 28.15), and there is a significantly positive correlation between number of species and languages ($r = 0.4$, $p < 0.0001$), and a significant negative correlation between number of threatened species and threatened languages ($r = -0.16$, $p = 0.01$). But when we correct for the non-independence between grid cells due to spatial autocorrelation using Dutilleul's modified *t*-test, there are no significant correlations between language diversity and mammal species richness ($r = 0.24$, $p = 0.22$), or between number of threatened languages and threatened species ($r = -0.13$, $p = 0.24$).

The same results emerge from the SAR models: there are no significant associations between language richness and mammal richness, elevation or land area per grid cell (table 1). The only significant correlate of number of threatened languages per grid cell is elevation, which remains significant when accounting for mammal species richness and land area in a multivariate model. The SAR models all provide a better fit to the data than the corresponding non-spatial regression, when compared using Akaike's information criterion (AIC) (table 1).

Turvey & Pettorelli's negative relationship between threatened languages and threatened mammal species is largely a result of their different elevational distributions—threatened language diversity is highest on lowlands of the north coast, and threatened mammal diversity is highest in the elevated central regions. Because a standard correlation assumes grid cells are statistically independent, it essentially samples this one distinct difference multiple times, resulting in pseudoreplication and elevating type 1 statistical error. When spatial autocorrelation is taken into account, the negative relationship between species and language threat disappears.

Globally, language and species diversity may show consistent trends (e.g. increasing towards the equator) but within smaller regions, local factors may operate to create finer scale patterns (e.g. mammal diversity is greater at higher elevations, while language diversity is strongly shaped by prehistoric settlement along coastal regions). Studies such as Turvey & Pettorelli's which focus on a particular region are an important addition to global-scale analyses. However, at all scales of analysis, it is critical to test whether the fundamental assumptions of the statistical analysis are met. If spatial autocorrelation is detected in the data, then appropriate methods that allow for the resulting non-independence must be used.

References

1. Turvey ST, Pettorelli N. 2014 Spatial congruence in language and species richness but not threat in the world's top linguistic hotspot. *Proc. R. Soc. B* **281**, 20141644. (doi:10.1098/rspb.2014.1644)
2. Foley WA. 2000 The languages of New Guinea. *Annu. Rev. Anthropol.* **29**, 357–404. (doi:10.1146/annurev.anthro.29.1.357)
3. Moore JL, Manne L, Brooks T, Burgess ND, Davies R, Rahbek C, Williams P, Balmford A. 2002 The distribution of cultural and biological diversity in Africa. *Proc. R. Soc. Lond. B* **269**, 1645–1653. (doi:10.1098/rspb.2002.2075)
4. Mace R, Pagel M. 1995 A latitudinal gradient in the density of human languages in North America. *Proc. R. Soc. Lond. B* **261**, 117–121. (doi:10.1098/rspb.1995.0125)
5. Gorenflo LJ, Romaine S, Mittermeier RA, Walker-Painemilla K. 2012 Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proc. Natl Acad. Sci. USA* **109**, 8032–8037. (doi:10.1073/pnas.1117511109)
6. Amano T, Sandel B, Eager H, Bulteau E, Svenning J-C, Dalsgaard B, Rahbek C, Davies RG, Sutherland WJ. 2014 Global distribution and drivers of language extinction risk. *Proc. R. Soc. B* **281**, 20141574. (doi:10.1098/rspb.2014.1574)
7. Nettle D. 1998 Explaining global patterns of language diversity. *J. Anthropol. Archaeol.* **17**, 354–374. (doi:10.1006/jaar.1998.0328)
8. Axelsen JB, Manrubia S. 2014 River density and landscape roughness are universal determinants of linguistic diversity. *Proc. R. Soc. B* **281**, 20133029. (doi:10.1098/rspb.2013.3029)
9. Harmon D. 1996 Losing species, losing languages: connections between biological and linguistic diversity. *Southwest J. Linguist.* **15**, 89–108.
10. Maffi L. 2005 Linguistic, cultural, and biological diversity. *Annu. Rev. Anthropol.* **34**, 599–617. (doi:10.1146/annurev.anthro.34.081804.120437)
11. Kissling WD, Carl G. 2008 Spatial autocorrelation and the selection of simultaneous autoregressive models. *Glob. Ecol. Biogeogr.* **17**, 59–71. (doi:10.1111/j.1466-8238.2007.00334.x)
12. Lewis MP. 2009 *Ethnologue: languages of the world*. Dallas, TX: SIL International.
13. Dutilleul P. 1993 Modifying the *t* test for assessing the correlation between two spatial processes. *Response Biometrics* **49**, 305–314. (doi:10.2307/2532625)