Cross-national analyses require additional controls to account for the non-independence of

nations

Scott Claessens[1*], Thanos Kyritsis[1], & Quentin D. Atkinson[1,2*]

[1] School of Psychology, University of Auckland, Auckland, New Zealand

[2] School of Anthropology and Museum Ethnography, University of Oxford, Oxford, United Kingdom

*email: scott.claessens@gmail.com; q.atkinson@auckland.ac.nz

Abstract

Cross-national analyses test hypotheses about the drivers of variation in national outcomes. However, since nations are connected in various ways, such as via spatial proximity and shared cultural ancestry, cross-national analyses often violate assumptions of non-independence, inflating false positive rates. Here, we show that, despite being recognised as an important statistical pitfall for over 200 years, cross-national research in economics and psychology still does not sufficiently account for non-independence. In a review of the 100 highest-cited cross-national studies of economic development and values, we find that controls for non-independence are rare. When studies do control for non-independence, our simulations suggest that most commonly used methods are insufficient for reducing false positives in non-independent data. In reanalyses of twelve previous cross-national correlations, half of the estimates are compatible with no association after controlling for non-independence using global proximity matrices. We urge social scientists to sufficiently control for non-independence in cross-national research.

Keywords: spatial non-independence; cultural non-independence; cross-national analyses; simulations; replications; reanalyses

Word count: 5102 words excl. Methods (Methods: 2163 words)

Cross-national analyses require additional controls to account for the non-independence of nations

Introduction

Nations are an important population unit structuring modern human life. The nation in which someone is born has a large effect on what they can expect out of life, including their income level[1], life expectancy[2], mental health[3], subjective well-being[4], and educational attainment[5]. Nations are also among the most important determinants of human cultural variation, with greater cultural similarity within than outside national borders[6].

Given the importance of nations for structuring human behaviour, there is justifiably huge interest in statistical analyses that attempt to predict variation in national outcomes of all kinds. At the time of writing, a search in Web of Science for the term "cross-national" in titles or abstracts returned over 13,000 unique hits. The standard practice for cross-national analyses is to conduct bivariate correlations or multiple regressions with individual data points representing different nations. Such analyses widen the scope of social science beyond Western populations[7,8] and have been used to study, among other topics, the causes of variation in the economic wealth of nations[9–12], global patternings of cultural norms and values[13–16], and the universality and diversity of human behaviour and psychology around the world[17–20].

However, cross-national analyses are complicated by the fact that nations are not statistically independent data points. Unlike independent random samples from a population, nations are related to one another in a number of ways. First, nations that are closer to one another tend to be more similar than distant nations. This phenomenon is known as spatial non-independence[21], and it occurs because nations in close spatial proximity share characteristics due to local cultural diffusion of ideas[22] and regional variation in climate and environment[21]. For example, the neighbouring African nations Zambia and Tanzania have similar levels of terrain ruggedness, which has been used to

25 partially explain their similar levels of economic development[23]. This pattern conforms to

26 Tobler's first law of geography: "everything is related to everything else, but near things

27 are more related than distant things"[24] (p. 236).

28   Second, nations with shared cultural ancestry tend to be more similar than culturally

29 unrelated nations. This is known as cultural phylogenetic non-independence[25–27], and

30 occurs because related nations share cultural traits inherited via descent from a common

31 ancestor. Shared cultural ancestry can result in a form of pseudoreplication, whereby

32 multiple instances of the same trait across nations are merely duplicates of the ancestral

33 original[25,28]. For example, the related island nations Tonga and Tuvalu share similar

34 languages and customs due to cultural inheritance from a common Polynesian population

35 dating back more than 1,000 years[29,30]. More recently, shared ancestry explains similarities

36 in institutions, norms, technologies, and values between colonial settlements and their

37 colonisers (e.g. English, French, Spanish, and Portuguese settlements of the Americas).

38 Importantly, these deep cultural connections between nations often span large geographic

39 distances around the world. Tonga and Tuvalu share cultural traits despite being separated

40 by over 1,500 kilometres of ocean, and South American and European nations remain

41 culturally similar today despite their locations on two separate continents. Shared cultural

42 ancestry must therefore be considered independently of spatial proximity in the study of

43 nations.

44   Spatial and cultural phylogenetic non-independence between nations make

45 cross-national inference challenging. A fundamental assumption of regression analysis is

46 that model residuals should be independently and identically distributed[31]. But without

47 accounting for spatial or cultural non-independence between nations, model residuals can

48 show structure that remains unaccounted for, violating this assumption. Treating nations

49 as independent can thus inflate false positive rates[32], producing spurious relationships

50 between variables that in fact only indirectly covary due to spatial or cultural connections[33]

51 (see Supplementary Figure 1 for an example causal model).

Non-independence is widely acknowledged in fields that routinely deal with spatially or culturally structured data. In ecology and sociology, studies with regional-level data use a variety of autoregressive models designed to account for spatial patternings[34,35]. In anthropology, researchers have recognised cultural non-independence as an important statistical pitfall for over 200 years, with issues of cultural pseudoreplication being identified in early comparative studies of marriage practices across societies[25]. In the twentieth century, anthropologists began to emphasise that human societies do not develop independently, but rather exist in a globally interconnected world system linked by shared history and cultural ancestry[36]. In order to minimise the confounding effects of this non-independence in comparisons of human societies, researchers compiled the Standard Cross-Cultural Sample of 186 cultures which deliberately avoids sampling closely related cultures[28], though it is difficult to completely remove spatial and cultural dependencies[37,38] and the smaller sample of cultures reduces statistical power. Today, anthropologists borrow phylogenetic comparative methods from evolutionary biology, such as phylogenetic least squares regression[39], when comparing societies, treating culturally related societies in the same way as biologists treat genetically related species (e.g.[40,41]).

At the national level, recent reanalyses have revealed that several cross-national relationships reported in economics and psychology do not hold when controlling for non-independence between nations. One working paper replicated 25 analyses of persistence in economics, in which modern national outcomes are regressed against historical characteristics of those nations, and found that over half of the relationships were attenuated when controlling for spatial non-independence[42]. Another replication study found that many of the widely publicised relationships between national-level pathogen prevalence and political institutions and attitudes fail to hold when controlling for various kinds of non-independence[43]. These reanalyses, and others[44–46], raise the question: how widespread a concern is non-independence in studies of national-level outcomes?

To address this question, we consider national-level variables of general interest across

79 the social sciences: economic development and cultural values. These variables are

80 frequently included as both outcomes and predictors in cross-national studies in economics

81 and psychology[9–16]. First, we demonstrate that economic development and cultural values

82 are spatially and culturally non-independent across nations, emphasising the need to

83 control for non-independence. Second, we review the 100 highest-cited cross-national

84 studies of economic development and cultural values and determine baseline rates of

85 controlling for non-independence in the literature. Third, we run simulations to determine

86 whether common methods of dealing with non-independence in the literature sufficiently

87 reduce false positive rates. Fourth, we reanalyse twelve previous cross-national analyses of

88 economic development and cultural values from our literature review, incorporating global

89 geographic and linguistic proximity matrices as controls for spatial and cultural

90 non-independence.

Results

National-level economic development and cultural values are spatially and culturally non-independent

94 In order to motivate our research question, it is important to first quantify the degree

95 of spatial and cultural non-independence for economic development and cultural values

96 around the world. If these variables are independent or only weakly non-independent, then

97 the issue might be safe to ignore. However, if they are more strongly non-independent, then

98 there is a possibility that non-independence could be confounding cross-national inferences.

99 To this end, we used Bayesian multilevel models to simultaneously estimate

100 geographic and cultural phylogenetic signal for a range of economic development and

101 cultural values variables. For economic development, we focused on the Human

102 Development Index[47], gross domestic product per capita, annual gross domestic product

103 per capita growth, and the Gini index of income inequality. For cultural values, we focused

104 on two primary dimensions of cultural values from the World Values Survey, traditional

105 vs. secular values and survival vs. self-expression values[16], as well as cultural tightness[14]

106 and individualism[15]. These variables are not intended to be a comprehensive list of all

107 national-level variables included in cross-national research, but rather an illustrative set of

108 variables that are widely used in the literature.

109 For all of these variables, we found that a substantial proportion of national-level

110 variation was explained by spatial proximity and/or shared cultural ancestry between

111 nations (Figure 1; see Supplementary Table 1 for numerical results). Signal estimates were

112 often strong, with spatial proximity and shared cultural ancestry frequently explaining over

113 half of the national-level variation. For spatial proximity, Bayes Factors indicated strong

114 evidence that the geographic signal estimates differed from zero for all economic

115 development variables and traditional values. However, the evidence was only equivocal for

116 survival values and individualism, and strong evidence was found that the geographic

117 signal estimate for tightness was equal to zero. For shared cultural ancestry, Bayes Factors

118 indicated strong evidence that the cultural phylogenetic signal estimates differed from zero

119 for all economic development and cultural values variables except for gross domestic

120 product per capita growth, for which the evidence was equivocal. These findings emphasise

121 the need to account for spatial and cultural phylogenetic non-independence in

122 cross-national analyses of economic development and cultural values.

Previous cross-national analyses have not sufficiently accounted for non-independence

125 Given that economic development and cultural values show evidence of geographic

126 and cultural phylogenetic signal, have cross-national analyses sufficiently accounted for this

127 non-independence? To assess this, we searched the published literature for articles that

128 combined the search terms "economic development" or "values" with the search terms

129 "cross-national", "cross-cultural", or "cross-country". We removed articles that did not
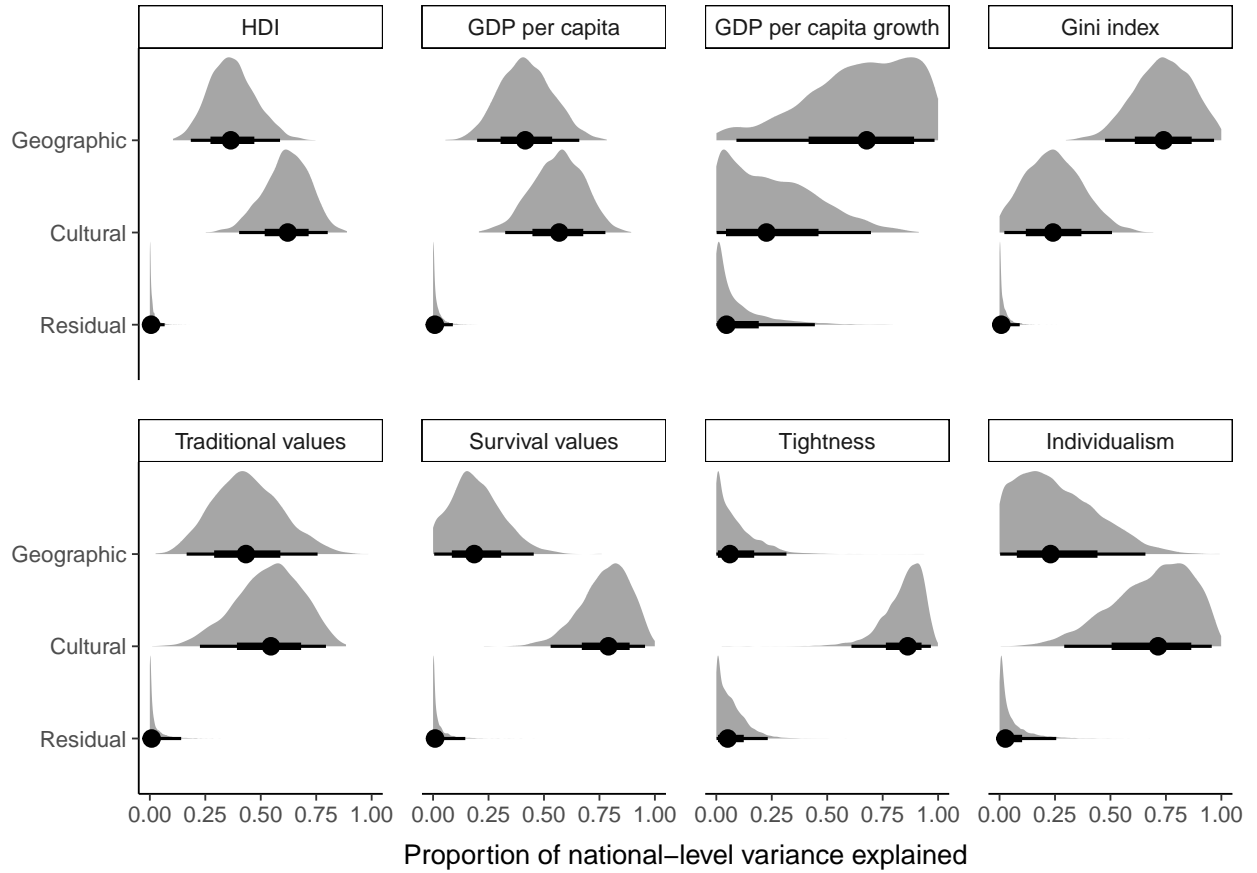
Figure 1. Posterior estimates of geographic and cultural phylogenetic signal for a range of economic development and cultural values variables. Geographic and cultural phylogenetic signal are operationalised as the proportion of national-level variance explained by geographic and linguistic proximity matrices. Grey ridges are full posterior distributions, points are posterior median values, and black lines are equal-tailed 50% and 95% credible intervals. Number of observations from these models are as follows: HDI (n = 1449), GDP per capita (n = 10289), GDP per capita growth (n = 9755), Gini index (n = 1826), traditional values (n = 277676), survival values (n = 277676), tightness (n = 57), and individualism (n = 67). HDI = Human Development Index; GDP = gross domestic product.

report original research, were not relevant to economic development or cultural values, or

did not report at least one cross-national analysis. We then retained the 100 articles (50 for

economic development, 50 for cultural values) with the highest annual rate of citations (see

Supplementary Data 1). For each of these highly-cited articles, we exhaustively recorded

every cross-national analysis reported in the main text (n = 4,308), identifying in each case

whether or not the analysis attempted to control for spatial, cultural, or any other form of

non-independence between nations (see Methods for detailed search criteria and coding

decisions).

The results of our literature review show that most published articles containing

cross-national analyses make no attempt to account for statistical non-independence.

Figure 2a plots the proportion of articles that contain at least one cross-national analysis

accounting for non-independence. We find that 42% of economic development articles

contain at least one attempt to control for non-independence (two-tailed 95% bootstrap

confidence interval [0.30 0.54]), while this proportion decreases to only 8% for cultural

values articles (two-tailed 95% bCI [0.02 0.16]). Both kinds of article are most likely to use

regional fixed effects (e.g. continent fixed effects) to account for non-independence, but

some articles also include controls for spatial distance (e.g. latitude) and shared cultural

history (e.g. colony status). These proportions are even lower when focusing on the full

sample of 4,308 analyses: only 5% (equal-tailed 95% credible interval [0.02 0.13]) of

individual economic development analyses and 1% (equal-tailed 95% CI [0.00 0.02]) of

individual cultural values analyses are estimated to control for non-independence

(Supplementary Figure 2).

While our review contains articles from journals with a range of impact factors, our

estimates could be biased downwards by analyses published in lower impact outlets with

more relaxed standards for issues like non-independence. It is also possible that, since our

literature review goes back as far as 1993, our estimates are being biased downwards by

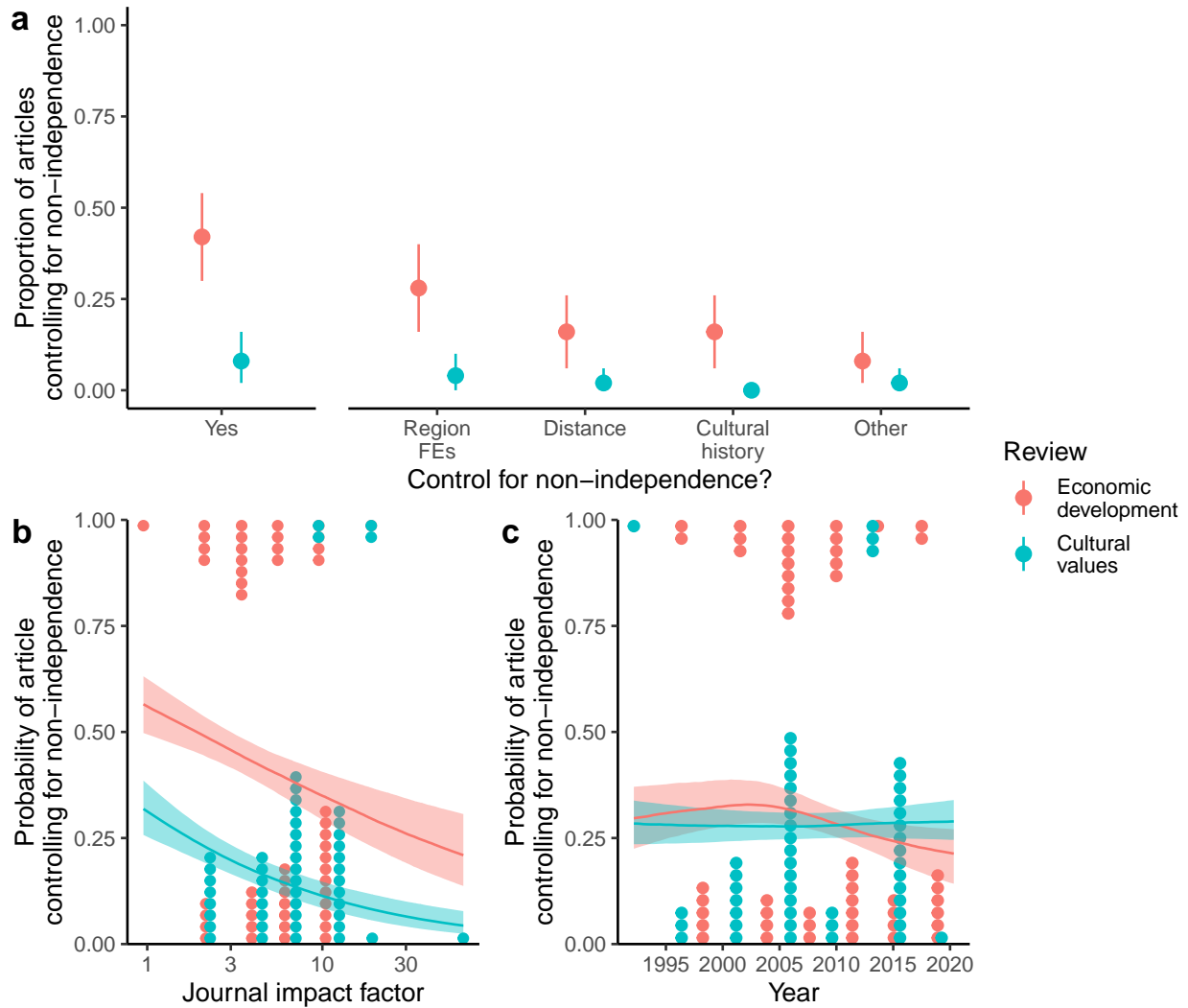earlier studies, and that controls for non-independence have increased over time with

Figure 2. Results from literature review of 100 highly-cited cross-national studies of economic development (red) and cultural values (blue). (a) Proportion of articles containing at least one analysis accounting for non-independence, overall and split by common methods of controlling for non-independence. Points represent raw proportions of articles and ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples). (b) The association between journal impact factor and the probability that an article contains at least one analysis accounting for non-independence. (c) Estimated trend over time for the probability that an article contains at least one analysis accounting for non-independence. Lines and shaded areas are posterior median regression lines and equal-tailed 50% credible intervals from Bayesian logistic regression models (n = 100 observations). Dots represent raw counts of individual articles that did (top) or did not (bottom) account for non-independence. Region FEs = region fixed effects.

157 methodological advancements and greater awareness of the issue. To test these possible

158 explanations for our low estimates, we fitted logistic regression models to the data from the

159 review, including log journal impact factor and publication year as separate predictors.

160 Interestingly, we found that, if anything, studies from higher impact journals were less

161 likely to include at least one control for non-independence than studies from lower impact

162 journals, both for studies of economic development (b = -0.38, equal-tailed 95% CI [-0.87

163 0.09]) and for studies of cultural values (b = -0.56, equal-tailed 95% CI [-1.10 -0.05]; Figure

164 2b). Moreover, splines revealed no relationship between publication year and the

165 probability of including at least one control for non-independence, both for studies of

166 economic development (b = -0.13, equal-tailed 95% CI [-1.10 0.82]) and for studies of

167 cultural values (b = -0.02, equal-tailed 95% CI [-1.00 0.96]; Figure 2c).

168 **Common methods of controlling for non-independence are insufficient for**

169 **reducing false positive rates in non-independent data**

170   Our literature review revealed that most cross-national analyses in the literature do

171 not control for spatial or cultural phylogenetic non-independence. When they do, they tend

172 to include controls like latitude and regional fixed effects. Do these methods sufficiently

173 account for statistical non-independence?

174   To compare the efficacy of different methods in the literature, we conducted a

175 simulation study. We simulated national-level datasets (n = 236 nations) with varying

176 degrees of spatial or cultural phylogenetic autocorrelation (i.e. non-independence) for

177 outcome and predictor variables, but with no direct causal relationship between the

178 variables. We then fitted naive regressions without controls to these datasets, as well as

179 regression models with controls for latitude, longitude, and continent fixed effects. Despite

180 not being identified in our literature review, we also included other methods that are often

181 used in the literature to account for non-independence. Additional spatial controls included

182 the mean of the predictor variable within a surrounding 2000km radius (e.g.[48]) and Conley

183   standard errors[49,50] based on geographic distances between nations (e.g.[19,48]). Additional

184   cultural controls included fixed effects for the language families of the majority-spoken

185   languages in each nation (e.g.[51]) and Conley standard errors based on genetic distances

186   between nations (e.g.[19,48]). These fixed effect approaches attempt to account for

187   non-independence by holding geographic location constant (latitude, longitude), discarding

188   between-region variation and exploiting only local variation (continent fixed effects, mean

189   of surrounding 2000km), or correcting standard errors for autocorrelation post-hoc while

190   leaving model coefficients unchanged (Conley standard errors).

191   Beyond fixed effect approaches, we also fitted Bayesian random effects regressions

192   that explicitly model spatial and/or cultural phylogenetic non-independence by allowing

193   nations to covary according to geographic and/or linguistic proximity matrices. Geographic

194   proximity between nations is calculated from inverse distances between longitude and

195   latitude coordinates. Linguistic proximity between nations is calculated from a global

196   phylogenetic tree that represents hierarchical relationships of genealogical descent for all

197   languages in the world. For each pair of nations, we calculated inverse phylogenetic

198   distances (i.e. number of branches separating two taxa) between all languages spoken in

199   that nation pair and produced an average linguistic proximity score weighted by the

200   percentages of speakers within those nations. To include the resulting geographic and

201   linguistic proximity matrices in our models, we included a Gaussian process[52,53] over

202   latitude and longitude values and/or assumed that nation random intercepts were

203   correlated in proportion to their linguistic proximity[54]. These random effects approaches

204   attempt to account for non-independence by modelling the covariance between nations that

205   is induced by their geographic or linguistic connections.

206   Figures 3 and 4 plot the estimated false positive rates from our simulation study, split

207   by different methods and different degrees of spatial or cultural phylogenetic

208   autocorrelation (see Supplementary Tables 2 and 3 for numerical results and

209   Supplementary Figures 3 and 4 for full distributions of effect sizes under strong

autocorrelation). Across all model types, false positive rates were measured as the

proportion of models that estimated a slope with a two-tailed 95% confidence / credible

interval excluding zero, falsely inferring a relationship when none is present. For reference,

weak autocorrelation in our simulation is comparable to the geographic signal for survival

values in Figure 1 (i.e. 20% of the national-level variance is explained by

non-independence), while moderate and strong levels of autocorrelation are comparable to

the cultural phylogenetic signal for traditional and survival values, respectively (i.e. 50%

and 80% of the national-level variance is explained by non-independence).

Our simulation study revealed that with at least moderate degrees of spatial or

cultural phylogenetic autocorrelation for both outcome and predictor variables, naive

regression models produce false positive rates above chance levels. This false positive rate

increases as the degree of autocorrelation increases. With strong spatial autocorrelation for

both outcomes and predictors, false positive rates reach as high as 77%. We find a slightly

lower false positive rate under strong cultural phylogenetic autocorrelation, though this

false positive rate is still greater than expected by chance (36%).

Most methods common in the literature do not reduce these high false positive rates.

With strong spatial autocorrelation for both outcome and predictor variables, false positive

rates remain above 50% when controlling for latitude, longitude, and language family fixed

effects (Figure 3). Similarly, Conley standard error corrections based on spatial and genetic

distances do not reduce false positive rates below 40% under strong spatial autocorrelation.

The most effective fixed effects methods are continent fixed effects, which continue to

produce a false positive rate of 35% under strong spatial autocorrelation, and controlling

for the mean of the predictor variable within a 2000km radius, which eliminates false

positives under even strong spatial autocorrelation (6%). However, additional simulations

revealed that these reductions in false positive rates come at the cost of lower statistical

power (see Supplementary Figures 5 - 7). In additional simulations where the true

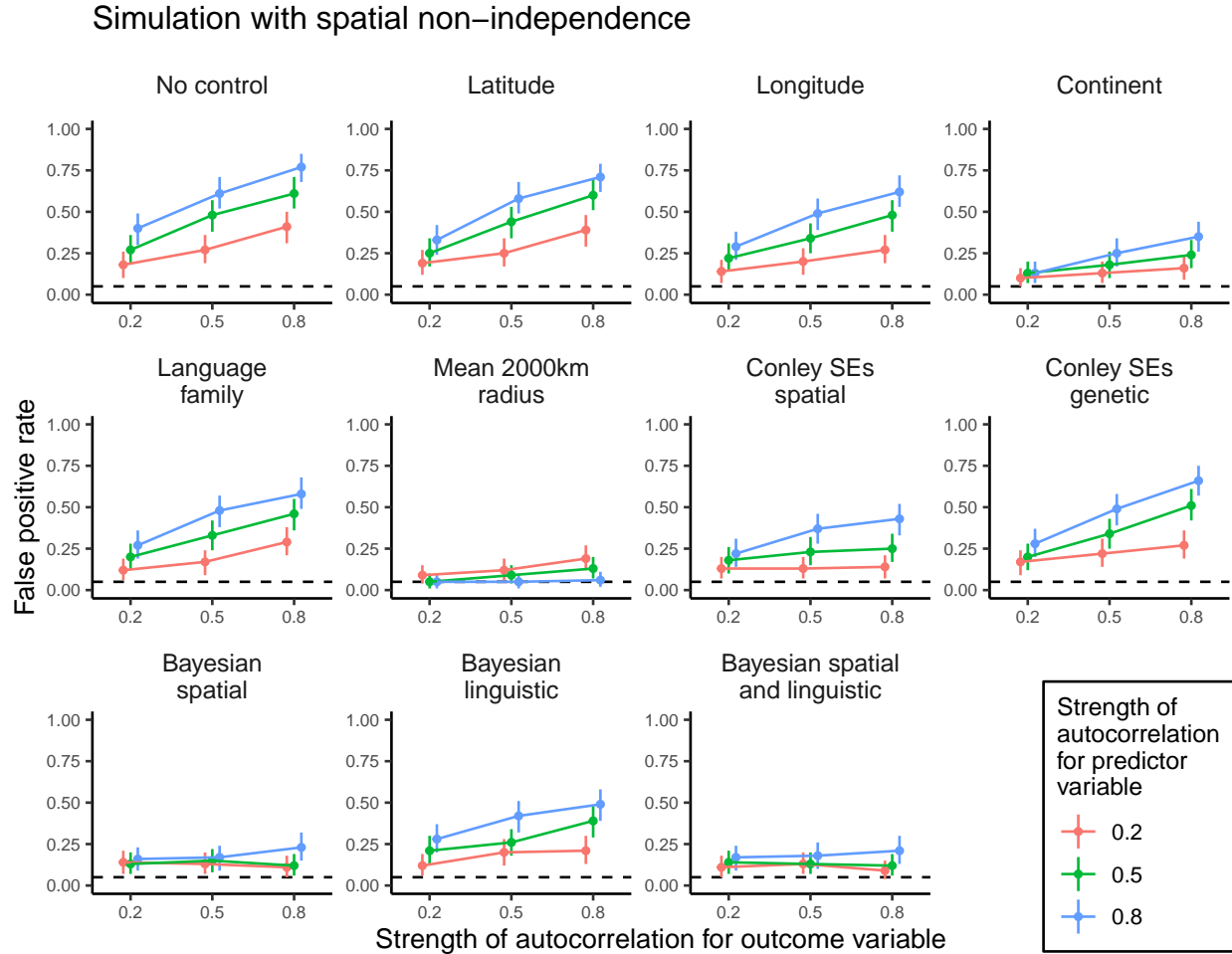relationship between the predictor and outcome variable was known, power analyses

Figure 3. False positive rates for different methods of controlling for spatial non-independence in our simulation study. For simulated outcome and predictor variables, we systematically varied the strength of spatial autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination and fitted different models to each dataset. False positive rates were operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of false positive models, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate the 5% false positive rate that is expected due to chance. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.
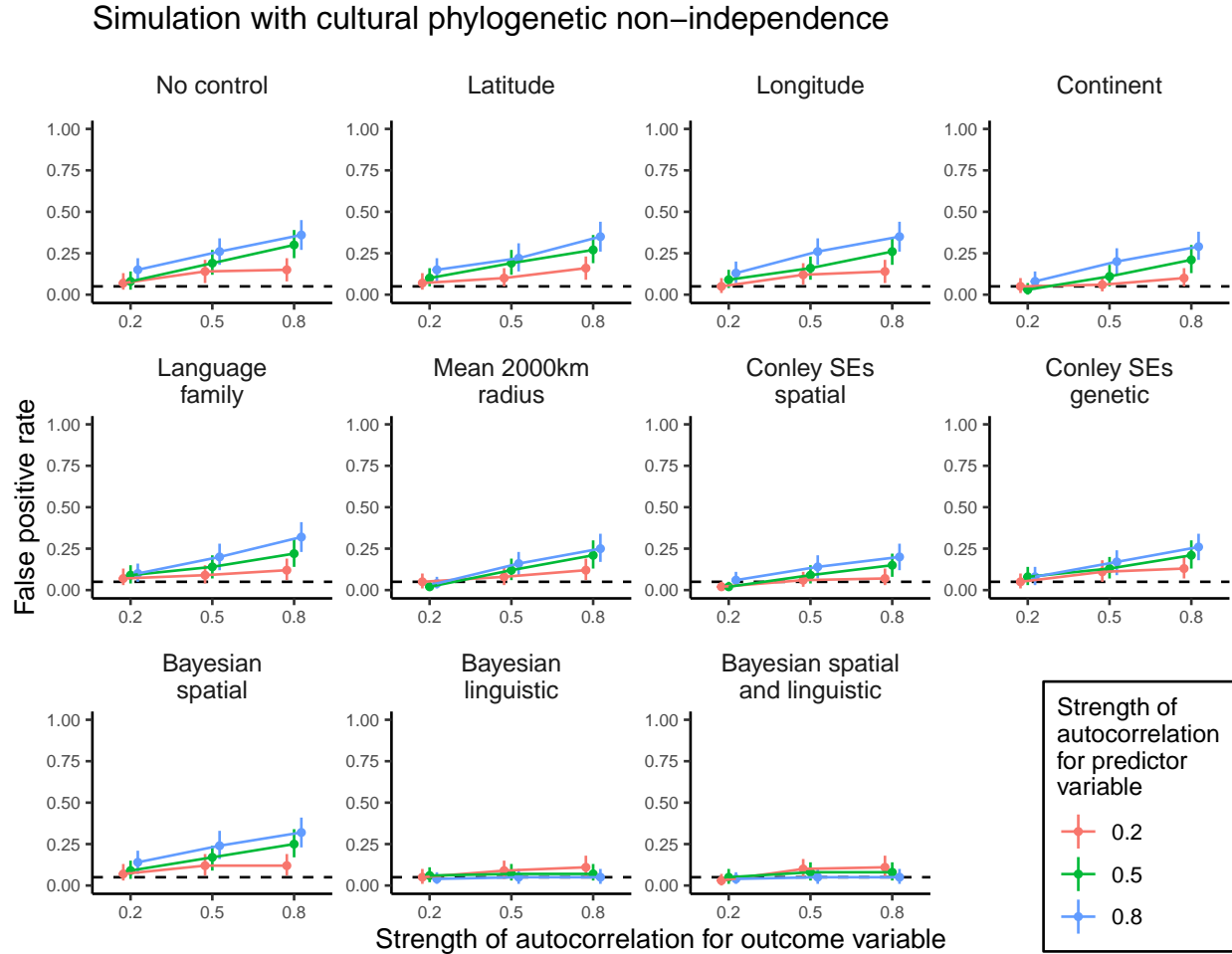
Simulation with cultural phylogenetic non–independence

Figure 4. False positive rates for different methods of controlling for cultural phylogenetic non-independence in our simulation study. For simulated outcome and predictor variables, we systematically varied the strength of cultural phylogenetic autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination and fitted different models to each dataset. False positive rates were operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of false positive models, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate the 5% false positive rate that is expected due to chance. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.

237 showed that both continent fixed effects and the 2000km radius control had less than 80%

238 power to detect moderate true correlations (r = 0.3) under strong spatial autocorrelation.

239   By contrast, Bayesian spatial Gaussian process regression with longitude and latitude

240 strikes a balance between reducing false positives and retaining high statistical power to

241 detect true effects. This approach reduces false positives to 15% under moderate spatial

242 autocorrelation and 23% under strong spatial autocorrelation. Random effects models that

243 additionally account for linguistic proximity between nations perform equally well, though

244 models with only linguistic covariance continue to produce false positives. False positives

245 are not completely eliminated with these random effects models. Nevertheless, these

246 methods have at least 80% power to detect moderate (r = 0.3) and large (r = 0.5) true

247 correlations between variables under all levels of spatial autocorrelation (see

248 Supplementary Figures 5 - 7).

249   In our simulation of cultural phylogenetic non-independence, we find that none of the

250 fixed effects methods reduce false positive rates (Figure 4). Controls for latitude, longitude,

251 continent fixed effects, the mean of the predictor variable in a 2000km radius, and Conley

252 standard error corrections based on spatial and genetic distances do little to change false

253 positive rates. Even language family fixed effects continue to produce a false positive rate

254 of 32% under strong cultural phylogenetic autocorrelation. By contrast, models with

255 random effects covarying according to linguistic proximity completely eliminate false

256 positives across all degrees of cultural phylogenetic autocorrelation. This approach is also

257 the only method that is able to detect large true correlations (r = 0.5) with at least 80%

258 power (see Supplementary Figures 8 - 10). Random effects models that additionally

259 account for geographic proximity between nations perform equally well, though models

260 with only a spatial Gaussian process continue to produce false positives.

Key findings in the literature are not robust to reanalysis with more rigorous methods

Our literature review and simulation study have shown that controls for non-independence are rare in cross-national studies of economic development and cultural values, and when studies do attempt to control for non-independence, the methods typically used are unable to sufficiently reduce false positive rates. This raises the worrying possibility that the cross-national literature in economics and psychology is populated with spurious relationships.

To determine how widespread this issue of spurious cross-national relationships might be, we reanalysed a subset of twelve previous cross-national analyses from our literature review, controlling for spatial and cultural phylogenetic non-independence using global geographic and linguistic proximity matrices. Out of the 100 papers included in our literature review, primary or secondary data were publicly available for 47 papers. We attempted to replicate key statistically significant cross-national correlations from these papers — mostly initial bivariate regression specifications without covariates (see Methods) — and stopped when we had sampled a set of twelve analyses for which we were able to replicate the original result. The final set included six analyses from our economic development review[55–60] and six from our cultural values review[13,14,16,61–63] that had available data and were able to be replicated. We pre-registered this set before running any control models (https://osf.io/uywx8/). We controlled for non-independence by including (1) a Gaussian process allowing nation random intercepts to covary according to a geographic proximity matrix, and/or (2) nation random intercepts that covaried according to a linguistic proximity matrix (see Supplementary Methods for full models).

Figure 5 visualises the results of our reanalysis (see Supplementary Table 4 for numerical results). Cross-national correlation effect sizes tended to reduce when controlling for statistical non-independence between nations, sometimes by as much as half of the

original effect size. Overall, after controlling for non-independence, six out of twelve

cross-national associations had 95% credible intervals that included zero. For the economic

development analyses, four out of six cross-national relationships had 95% credible intervals

including zero when controlling for spatial non-independence. For the cultural values

analyses, two out of six cross-national relationships had 95% credible intervals including

zero when controlling for cultural phylogenetic non-independence. Supplementary Figure

11 shows these cross-national correlations plotted against the raw data.

To understand why some cross-national correlations were attenuated by controls for

non-independence while others were robust, we further explored our fitted models for

evidence of spatial and cultural autocorrelation. For each outcome variable, our Gaussian

process models provided varying estimates of how quickly spatial autocorrelation declined

with distance (Supplementary Figure 12). For example, in Skidmore and Toya[60] gross

domestic product growth was only moderately spatially autocorrelated at 1,000 km

distance (posterior median spatial autocorrelation at 1,000 km = 0.42, equal-tailed 95% CI

[0.07 0.90]), whereas in Inglehart and Baker[16] traditional values were strongly spatially

autocorrelated at the same distance (posterior median spatial autocorrelation at 1,000 km

= 0.96, 95% CI [0.81 0.99]). We also found varying estimates of cultural phylogenetic

signal (Supplementary Figure 13), with some outcome variables expressing low signal

(e.g. confidence in institutions[63]; posterior median = 0.07, 95% CI [0.00 0.53]) and others

expressing high signal (e.g. female labour force participation[62]; posterior median = 0.89,

95% CI [0.63 0.98]). Exploratory regressions provided suggestive evidence that stronger

estimates of spatial autocorrelation or cultural phylogenetic signal resulted in a more

pronounced reduction in the effect size when controlling for non-independence between

nations (Supplementary Figure 14). However, these negative slopes for spatial

autocorrelation (b = -0.19, 95% CI [-1.33 1.04]) and cultural phylogenetic signal (b = -0.28,

95% CI [-1.26 0.70]) were both very uncertain, due to the small number of analyses and the

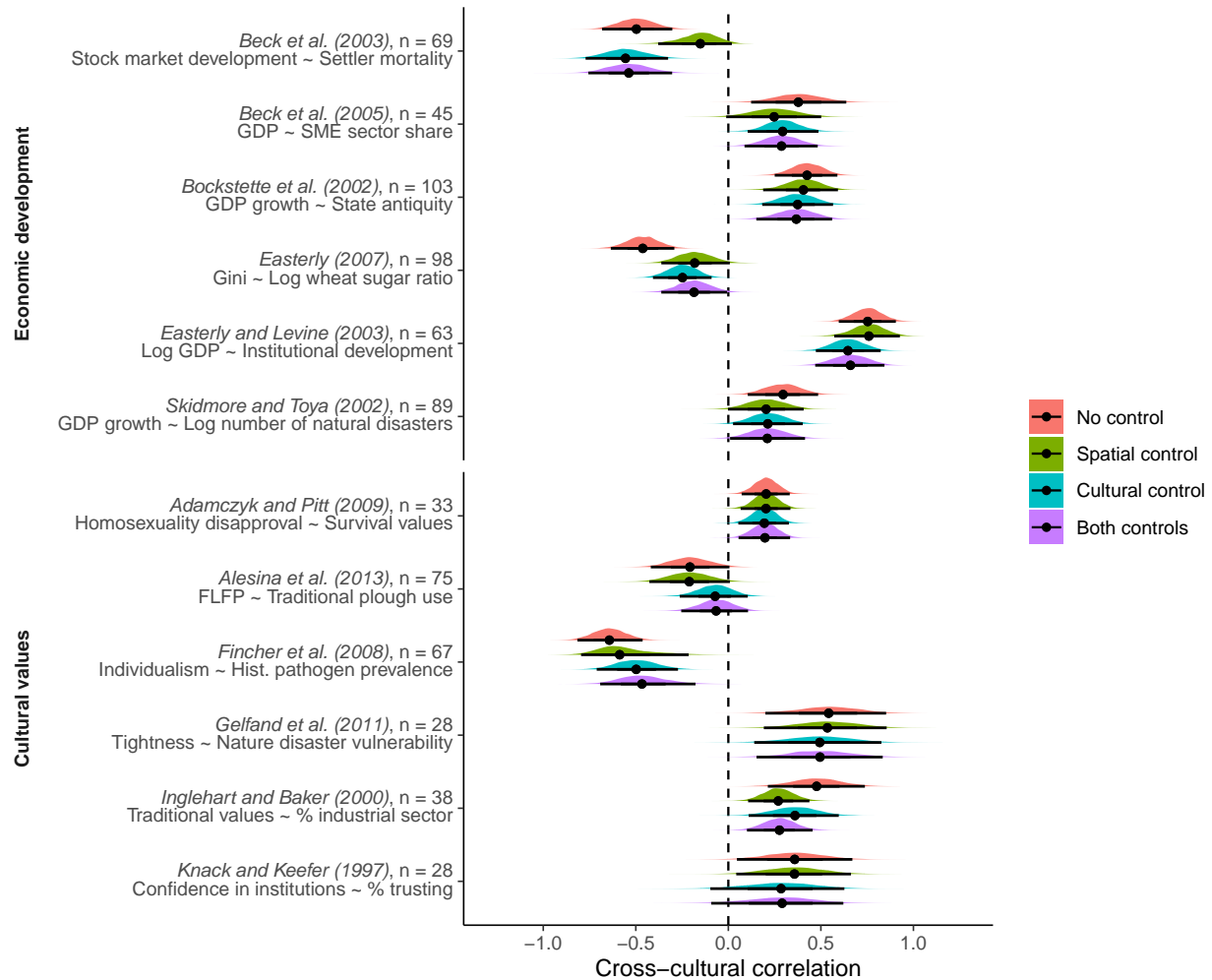posterior uncertainty in effect sizes and estimates of non-independence.

Figure 5. Posterior correlations from our reanalysis of twelve previous cross-national analyses. For each previous cross-national relationship, we plot the posterior slopes from a naive regression (red), a regression controlling for spatial non-independence (green), a regression controlling for cultural phylogenetic non-independence (blue), and a regression controlling for both spatial and cultural phylogenetic non-independence simultaneously (purple). All outcome and predictor variables are standardised. Most analyses are simple bivariate cross-national correlations, but Gelfand et al. (2011) is a partial correlation controlling for log gross national income and Adamczyk and Pitt (2009) is a multilevel model including several covariates. Coloured ridges are full posterior distributions, and points and black lines represent posterior medians and equal-tailed 95% credible intervals. Numbers of observations from the models are as follows, from top to bottom: n = 69, n = 45, n = 103, n = 98, n = 63, n = 89, n = 33, n = 75, n = 67, n = 28, n = 38, and n = 28. GDP = gross domestic product. FLFP = female labour force participation.

Discussion

314   In a literature review and simulation, we found that cross-national studies in

315 economics and psychology rarely account for non-independence between nations, and, when

316 they do, the methods they use are insufficient to reduce false positives in non-independent

317 data. In a reanalysis of twelve cross-national correlations, we further showed that

318 neglecting to account for non-independence has resulted in potentially spurious

319 relationships in the published literature, with half of the correlations failing to replicate

320 when controlling for spatial or cultural non-independence with more rigorous methods.

321 These findings suggest that cross-national analyses in economics and psychology should be

322 interpreted with caution until non-independence is sufficiently accounted for.

323   Our initial analyses add to and clarify existing evidence regarding the degree of

324 non-independence for national-level economic and cultural variables. One previous study

325 suggested that geographic proximity is more important than deep cultural ancestry in

326 explaining the distribution of human development across Eurasian nations, though the

327 authors noted that their small sample of 44 nations and regional focus limited their

328 statistical power[64]. By contrast, our global samples of over 160 nations revealed strong

329 cultural phylogenetic signal, as well as geographic signal, for the Human Development

330 Index, GDP per capita, and the Gini index of inequality. Another previous study found

331 that similarities in the cultural values of nations are predicted by linguistic, but not

332 geographic, distances between those nations[6]. We find this same result for survival

333 vs. self-expression values, cultural tightness, and individualism, but for traditional

334 vs. secular values we find that both linguistic and geographic proximity are important

335 independent predictors of global variation. These findings emphasise the need to account

336 for both spatial and cultural phylogenetic non-independence in cross-national studies of

337 economic development and cultural values.

338   Crucially, our literature review and simulation study revealed that the most

340 commonly used controls for non-independence do not sufficiently deal with the issue. In

341 our simulations, controlling for either latitude or longitude did not reduce false positive

342 rates. This result calls into question the use of controls like distance to the equator to

343 account for non-independence in cross-national regression models, though these controls

344 may still be suitable to account for regional or latitudinal variation in climate, ecology, and

345 natural threats (e.g. pathogens), which we did not simulate. High false positive rates

346 persisted with Conley standard errors, despite recent claims that these standard error

347 corrections are sufficient to deal with spatial non-independence[65]. The simulation also

348 confirmed the assertion that fixed effects for spatial or cultural groupings (e.g. continent or

349 language family fixed effects) are insufficient because non-independence still remains within

350 groupings[43]. This logic further applies to analyses that control for non-independence by

351 separately analysing different regions (e.g.[66]). Controlling for the mean of the predictor

352 variable within a 2000km radius[48] eliminated false positive rates in spatially autocorrelated

353 data, but had reduced statistical power to detect true associations. Across all model types

354 in our simulation, the only methods that reduced false positive rates while retaining high

355 statistical power were the random effects models with covariance matrices. The important

356 advantage of these models is not that they are Bayesian per se (any of the approaches used

357 here could be implemented in a Bayesian framework), but rather that they explicitly model

358 covariance as a function of spatial or cultural distance.

359   There are other approaches to controlling for non-independence that we did not

360 include in our simulation. For example, conditional autoregressive models[34] and generalised

361 additive models[67] are approaches that can be applied in both frequentist and Bayesian

362 frameworks. There are also alternative ways to operationalise cultural distances between

363 nations beyond linguistic distances, including metrics like cultural fixation indices

364 (e.g. cultural F_{ST}[6,68]), covariance based on genetic distances[69], and phylogenetic distances

365 between religious traditions. We see merit in each of these approaches and the use of one

366 over another will depend on the specific question at hand. We decided to focus on

367   linguistic distances in this study since language is a tangible socially-learned trait that has

368   previously been used to successfully track the effects of deep cultural ancestry on modern

369   national outcomes[64,70,71]. Future work should explore whether other approaches are

370   sufficient to reduce false positive rates in spatially and culturally non-independent data. In

371   addition to spatial proximity and shared cultural ancestry, we did not simulate other

372   sources of non-independence that potentially exist in real cross-national datasets, such as

373   modern connections between nations due to flows of people and information (e.g. flight

374   networks, social media networks) and shared histories of colonialism and capitalist modes

375   of production that have shaped today's global landscape[72,73]. Additional controls will be

376   required to ensure that these sources of non-independence do not confound cross-national

377   inferences.

378   Ours is not the first review to show that studies are misapplying statistical methods

379   in ways that inflate false positive rates. For example, other literature reviews have shown

380   that studies in the social sciences tend to use small samples of participants[74], treat ordinal

381   data as metric[75], incorrectly handle missing values[76], and ignore best practices in

382   meta-analyses[77]. Why do cross-national studies also rarely account for non-independence?

383   At the institutional level, one possibility is that such practices are incentivised because

384   they generate statistically significant relationships, which increase the probability that a

385   study is published[74]. Indeed, we found that controls for non-independence were less

386   common among articles published in high impact journals, suggesting that researchers are

387   rewarded for such practices. At the individual level, another possibility is that researchers

388   outside of anthropology and ecology are less aware of the problem, or believe that the

389   problem does not apply to analyses of nations. Even if researchers appreciate the problem,

390   they might not know of suitable controls or perceive the methods to be too complex.

391   These institutional- and individual-level barriers can be combatted. First,

392   cross-national replication studies like ours and others[42–46], combined with the

393   methodological reviews included in Registered Reports[78], might change incentive structures

394 and encourage researchers to analyse the world's nations with more rigorous methods.

395 Second, more explicit descriptions of causal models could promote controls for

396 non-independence by clearly outlining the nature of confounding and the sources of

397 autocorrelation in cross-national data[79]. The causal model outlined in Supplementary

398 Figure 1 is a useful example, but individual studies must outline their own particular

399 causal assumptions, which may include further sources of non-independence and

400 confounding variables to control for (e.g. post-communist status, colony status). These

401 causal models can then be used to design tailored statistical estimation strategies. Indeed,

402 in our review, economists studying economic development dealt with national-level

403 non-independence more than psychologists studying cultural values, likely because

404 economics studies tend to be lengthy statistical exercises that systematically incorporate or

405 exclude numerous variables in an attempt to infer causation. Third, the recent widespread

406 accessibility of open source statistical software, such as the programming language Stan[80]

407 and the R package brms[81], should promote the use of more rigorous methods to control for

408 non-independence. Using brms, for example, Bayesian Gaussian process regression is

409 straightforward to conduct, requiring only longitude and latitude values for nations. We

410 have provided an online tutorial to help researchers apply these methods to their own

411 cross-cultural datasets (https://scottclaessens.github.io/blog/2022/crossnational/).

412   Until such changes are implemented and sufficient controls for non-independence are

413 the norm, existing cross-national correlations should be interpreted with caution. In our

414 reanalyses, we found that half of the cross-national correlations had equal-tailed 95%

415 credible intervals that included zero when controlling for spatial and/or cultural

416 phylogenetic non-independence. While these results are in line with previous reanalyses[42],

417 we note that we are unable to outright reject the claims from these studies, since we only

418 reanalysed the first bivariate regression specifications presented in the papers[65]. More

419 detailed sets of reanalyses would be required to comprehensively challenge the claims from

420 these specific papers. Nevertheless, these reanalyses do show, more broadly, that the

421 problem of statistical non-independence applies to a wider range of national-level variables

422 than those identified by previous work, such as parasite stress and democratic outcomes[43].

423 Moreover, given our finding that most studies in the current cross-national literature do

424 not deal with non-independence at all, our reanalyses raise the worrying possibility that

425 this literature is populated with spurious relationships. Future work should expand our set

426 of reanalyses to determine the extent of this problem in the literature.

427   We do not wish to dissuade researchers from conducting cross-national studies. On

428 the contrary, such work promises to deepen understanding of our world, including the

429 causes and consequences of economic development and cultural values. Moreover,

430 cross-national studies allow social scientists to broaden their scope of study beyond

431 Western populations[7], providing the representative samples necessary to test evolutionary

432 and socio-ecological theories of human behaviour[8,82]. But in order to minimise spurious

433 relationships in global datasets, we urge researchers to control for spatial and cultural

434 phylogenetic non-independence when reporting cross-national correlations. Nations are not

435 independent, and our statistical models must reflect this.

436   **Methods**

437 **Geographic and cultural phylogenetic signal for measures of economic**

438 **development and cultural values**

439   To estimate the degree of spatial and cultural phylogenetic non-independence in

440 economic development and cultural values, we calculated geographic and cultural

441 phylogenetic signal for global measures of development and values. For economic

442 development variables, we retrieved longitudinal data on the Human Development Index[47]

443 (1990 - 2019; n = 189 nations), gross domestic product per capita (1960 - 2021; n = 209

444 nations), annual percentage growth in gross domestic product per capita (1961 - 2021; n =

445 208 nations), and the Gini coefficient of income inequality (1967 - 2021; n = 167 nations).

446 Human development data were retrieved from the United Nations Development

447 Programme (https://hdr.undp.org/en/content/download-data) and data for all other

448 economic development variables were retrieved from the World Bank

449 (https://data.worldbank.org/). For cultural values variables, we retrieved longitudinal data

450 on traditional vs. secular values and survival vs. self-expression values from the World

451 Values Survey[16] (1981 - 2019; n = 116 nations). We downloaded the full Integrated Values

452 Survey, which included all waves from the World Values Survey and the European Values

453 Survey, and computed the two dimensions of cultural values following procedures from

454 previous research[16]. Additionally, we retrieved cross-sectional data on cultural tightness (n

455 = 57 nations) and individualism (n = 97 nations) from previous work[13,83].

456   To calculate geographic and cultural phylogenetic signal, we created two proximity

457 matrices for 269 of the world's nations: a geographic proximity matrix and a linguistic

458 proximity matrix. Geographic distance between two nations was calculated as the logged

459 geodesic distance between nation capital cities (data from the maps R package[84]) using the

460 geosphere R package[85]. The geographic proximity matrix was computed as one minus the

461 log geographic distance matrix scaled between 0 and 1. Linguistic proximity between two

462 nations was calculated as the cultural proximity between all languages spoken within those

463 nations, weighted by speaker percentages. We acquired cultural proximity data by

464 combining the language family trees provided by Glottolog v3.0[86] into one global language

465 tree (undated and unresolved). We calculated cultural proximity s between two languages

466 j and k as the distance (in number of nodes traversed) of their most recent common

467 ancestor i to the root of the tree, through the formula:

$$s_{jk} = \frac{n_r - n_i}{n_r} \tag{1}$$

468   where n_r is the maximum path length (in number of nodes traversed) leading to the

pan-human root r, and n_i is the maximum path length leading to node i. We then

combined these proximities with speaker data from Ethnologue 21[87] and compared every

language spoken within those nations by at least 1 permille of the population, weighted by

speaker percentages, through the formula:

$$w_{lm} = \Sigma\Sigma p_{lj} p_{mk} s_{jk} \tag{2}$$

where p_{lj} is the percentage of the population in nation l speaking language j, p_{mk} is

the percentage of the population in nation m speaking language k, and s_{jk} is the proximity

measure between languages j and k[88]. This calculation resulted in a linguistic proximity

matrix with values between 0 and 1.

We included these matrices in Bayesian multilevel models, allowing nation random

intercepts to covary according to both geographic and linguistic proximity simultaneously.

These models were fitted with the R package brms[81] and converged normally ($\hat{R} < 1.1$).

The assumptions of these models were met: residuals were approximately normally

distributed, though this was not formally tested. Estimates of geographic and cultural

phylogenetic signal were computed as the proportion of national-level variance in these

models explained by geographic and linguistic proximity matrices.

Literature review

We exported two searches from Web of Science (https://www.webofknowledge.com/)

on 27th September 2021, restricting our searches to articles published between 1900 and

2018. The first search was for the terms "economic development" AND ("cross-national"

OR "cross-cultural" OR "cross-country"), which returned 965 articles. The second search

was for the terms "values" AND ("cross-national" OR "cross-cultural" OR

"cross-country"), which returned 6806 articles. As this was not a formal systematic literature review, we did not follow PRISMA[89] guidelines for systematic literature reviews.

Once exported, we ordered the articles by descending number of citations per year since initial publication, using citation counts reported by Web of Science. We then coded each article, in order, for inclusion in our review. Articles were only included if: (1) they were judged to be relevant to economic development or cultural values; (2) they were an original empirical research article; and (3) they contained at least one analysis with national-level outcome or predictor variables. We stopped when we had included 50 articles for the economic development review and 50 articles for the cultural values review.

Within each included article, we exhaustively coded every individual cross-national analysis reported in the main text. We coded mainly correlation or regression analyses, and explicitly excluded meta-analyses, factor analyses, measurement invariance analyses, multidimensional scaling analyses, hierarchical clustering analyses, multiverse analyses, and scale development / validation analyses. We also excluded analyses that compared only two, three, four, five, or six nations. For each included analysis, we recorded the year, impact factor of the journal (retrieved from https://jcr.clarivate.com/jcr/home), outcome variable, all predictor variables, test statistic, p-value, number of nations, number of data points, model type, if the data were available, and whether and how the analysis attempted to control for non-independence.

We coded common attempts to control for non-independence between nations. These included: (1) any higher-level control variables for spatial regional groupings (e.g. continent fixed effects); (2) any geographic distance control variables (e.g. distance between capital cities, distance from equator, latitude); (3) any control variables capturing shared cultural history (e.g. former colony, legal origin fixed effects, linguistic history, cultural influence); and (4) any other control variables, tests, or approaches that were deemed as attempts to control for non-independence (e.g. eigenvector filtering[90], controls for trade-weightings

516 between nations, cross-sectional dependence tests[91], separate analyses for subsets of

517 nations). These were coded by the first author.

518   Once we had compiled our review database, we calculated the proportion of articles

519 attempting to control for non-independence at least once. We also calculated the

520 proportion of articles employing the different types of control listed above at least once:

521 regional fixed effects, distance, shared cultural history, or other. For these proportions, we

522 calculated two-tailed 95% bootstrap confidence intervals with 1,000 bootstrap iterations.

523 Additionally, we predicted the probability of an article attempting to control for

524 non-independence at least once using Bayesian logistic regression, including in separate

525 models log journal impact factor and year of publication as linear and spline predictors,

526 respectively.

527   For individual analyses, we dealt with the nested nature of the data (analyses nested

528 within articles) by fitting Bayesian multilevel logistic regression models with review type

529 (economic development vs. cultural values) as the sole fixed effect and random intercepts

530 for articles. We fitted these models separately for overall attempts to control for

531 non-independence and split by method type. We report the adjusted proportions with

532 equal-tailed 95% credible intervals (see Results and Supplementary Figure 2). Additionally,

533 we predicted the probability of an analysis attempting to control for non-independence

534 using Bayesian multilevel logistic regression with random intercepts for articles. In separate

535 models, we included log journal impact factor and year of publication as linear and spline

536 predictors, respectively. All Bayesian models were fitted with the brms R package[81]. Our

537 priors were informed by prior predictive checks, and all models converged normally ($\hat{R} <$

538 1.1). The assumptions of these models (i.e., binary nested data) were met.

Simulations

540   We simulated data for 236 nations i with varying degrees of spatial or cultural

541   phylogenetic signal for outcome y and predictor x using the following generative model:

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha_y \\ \alpha_x \end{bmatrix}, \mathbf{S}\right) \tag{3}$$

$$\alpha_y \sim \text{Normal}(0, \sqrt{\lambda} \cdot \Sigma)$$

$$\alpha_x \sim \text{Normal}(0, \sqrt{\rho} \cdot \Sigma)$$

$$\mathbf{S} = \begin{pmatrix} \sqrt{1-\lambda} & 0 \\ 0 & \sqrt{1-\rho} \end{pmatrix} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} \sqrt{1-\lambda} & 0 \\ 0 & \sqrt{1-\rho} \end{pmatrix}$$

542   where Σ is a correlation matrix proportional to either geographic or linguistic

543   proximities between nations, λ and ρ are autocorrelation parameters that represent the

544   expected spatial or cultural phylogenetic signal for outcome and predictor variables,

545   respectively, and r is the true cross-national correlation between the variables after

546   accounting for autocorrelation. Importantly, when $r = 0$ in this simulation, we know that

547   there is no direct causal relationship between y and x. Instead, any relationship between

548   the two variables is merely the result of autocorrelation.

549   We set the autocorrelation parameters λ and ρ to either 0.2 (weak), 0.5 (moderate),

550   or 0.8 (strong). We also initially set the true cross-national correlation to 0 in order to

551   determine false positive rates, and then additionally set r to 0.1 (small effect), 0.3 (medium

552   effect), and 0.5 (large effect) in order to determine statistical power to detect true effects.

553   For each parameter combination, we simulated 100 datasets, resulting in 3600 datasets.

554   Each dataset had 236 rows representing different nations, with the following associated

555   data for each nation: latitude, longitude, continent (Africa, Asia, Europe, North America,

Oceania, or South America), language family of the nation's majority spoken language

(Afro-Asiatic, Atlantic-Congo, Austroasiatic, Austronesian, Eskimo-Aleut, Indo-European,

Japonic, Kartvelian, Koreanic, Mande, Mongolic-Khitan, Nilotic, Nuclear Trans New

Guinea, Sino-Tibetan, Tai-Kadai, Tupian, Turkic, or Uralic), the mean of the predictor

variable within a 2000km radius, and coordinates for genetic distances from a previous

study[19] (only available for 177 nations).

With the resulting simulated datasets, we standardised outcome and predictor

variables and fitted eleven different models: (1) naive regression without controls, (2)

regression with latitude control, (3) regression with longitude control, (4) regression with

continent fixed effects, (5) regression with language family fixed effects, (6) regression

controlling for the mean of the predictor variable in a 2000km radius, (7) regression

employing Conley standard errors based on geographic distances, (8) regression employing

Conley standard errors based on genetic distances, (9) Bayesian regression including a

Gaussian process over latitudes and longitudes, (10) Bayesian regression including random

intercepts covarying according to linguistic proximity, and (11) Bayesian regression

including both a Gaussian process over latitudes and longitudes and random intercepts

covarying according to linguistic proximity.

Models employing Conley standard errors either required latitude and longitude

values or coordinates for genetic distances. To determine distance cutoffs, we employed an

approach recommended in previous work[92]: we fitted models with a range of feasible

distance cutoffs and retained the model with the largest standard error for the slope

parameter. These models were fitted using the conleyreg R package[93]. Bayesian models

were fitted using the brms R package[81]. Our choice of priors was based on prior predictive

simulation. All models converged normally ($\hat{R} < 1.1$). Across all model types and

parameter combinations, we calculated the false positive rate as the proportion of models

that estimated slopes with a two-tailed 95% confidence / credible interval excluding zero

when $r = 0$. We calculated statistical power as the proportion of models that estimated

583 slopes with a 95% confidence / credible interval excluding zero when $r > 0$. We calculated

584 two-tailed 95% bootstrap confidence intervals for these false positive rates and statistical

585 power estimates with 1,000 bootstrap iterations.

Reanalyses

587 We searched the individual analyses from our literature review for statistically

588 significant cross-national correlations with available primary or secondary data. We

589 restricted our search to one analysis per paper, and searched until we had a set of twelve

590 analyses, six from economic development papers and six from cultural values papers, for

591 which we were able to replicate the original result (i.e. find a cross-national correlation

592 with the same sign and roughly the same effect size). We also ensured that at least one

593 analysis was a multilevel model, with multiple observations per nation.

594 The twelve analyses that we settled on[13,14,16,55–63] were mostly bivariate cross-national

595 correlations, except for two. One analysis[14] additionally controlled for log gross national

596 income, and another analysis[60] is a multilevel model including random intercepts for

597 nations and several individual-level and national-level covariates (see Model 5 in original

598 paper). Before running any additional models, we pre-registered these twelve analyses on

599 the Open Science Framework on 25th January 2022 (https://osf.io/u8tbf). We endeavoured

600 to keep the sample sizes of our reanalyses as close to the original analyses as possible,

601 though there were some deviations (see Supplementary Table 5). Despite these slight

602 deviations from the original analyses, all models reported in Figure 5 are fitted to the same

603 number of data points, meaning that any changes in effect sizes are solely due to

604 controlling for non-independence.

605 For each individual analysis, we ran four models: (1) a naive regression replicating

606 the original finding, (2) a regression including a Gaussian process allowing nation random

607 intercepts to covary according to a geographic proximity matrix from latitude and

longitude values, (3) a regression including nation random intercepts that covaried
according to a linguistic proximity matrix, and (4) a regression including both a geographic
Gaussian process and nation random intercepts with linguistic covariance. See
Supplementary Methods for full models.

We fitted these models using the brms R package[81]. Our choice of priors was based
on prior predictive simulation. All models converged normally ($\hat{R} < 1.1$), though for some
models we resorted to using approximate Gaussian processes[94] to reach convergence. The
assumptions of these models were met: residuals were approximately normally distributed,
though this was not formally tested.

Reproducibility

All data and code are accessible on GitHub[95]. We used the targets R package[96] to
create a reproducible data analysis pipeline and the papaja R package[97] to reproducibly
generate the manuscript.

Ethics

We did not apply for ethical approval through an ethics board as we analysed only
publicly available and simulated data in this study.

Data Availability

All data to reproduce the statistical analyses in this manuscript can be found on GitHub[95]. Data on human development were retrieved from the United Nations Development Programme (https://hdr.undp.org/en/content/download-data). Data on GDP per capita, annual GDP per capita growth, and the Gini coefficient were retrieved from the World Bank (https://data.worldbank.org/). Data on traditional vs. secular values and survival vs. self-expression values were retrieved from the World Values Survey (https://www.worldvaluessurvey.org/wvs.jsp). Data on cultural tightness were retrieved from the OSF repository for ref[83] (https://osf.io/47pe8/). The review data generated in this study are provided in Supplementary Data 1. All other datasets (e.g., for replications) were retrieved from tables and supplementary tables directly from papers cited in the main text[13,14,16,55–63] - these datasets can be found in our GitHub repository[95].

Code Availability

All code to reproduce the statistical analyses in this manuscript can be found on GitHub[95].

References

1. Caselli, F. Accounting for cross-country income differences. in (eds. Aghion, P. & Durlauf, S. N.) vol. 1 679–741 (Elsevier, 2005).

2. Austin, K. F. & McKinney, L. A. Disease, war, hunger, and deprivation: A cross-national investigation of the determinants of life expectancy in less-developed and sub-Saharan African nations. *Sociological Perspectives* **55**, 421–447 (2012).

3. Rai, D., Zitko, P., Jones, K., Lynch, J. & Araya, R. Country- and individual-level socioeconomic determinants of depression: Multilevel cross-national comparison. *British Journal of Psychiatry* **202**, 195–203 (2013).

4. Diener, E., Diener, M. & Diener, C. Factors predicting the subjective well-being of nations. in *Culture and well-being: The collected works of Ed Diener* (ed. Diener, E.) 43–70 (Springer Netherlands, 2009). doi:10.1007/978-90-481-2352-0_3.

5. Kirkcaldy, B., Furnham, A. & Siefen, G. The relationship between health efficacy, educational attainment, and well-being among 30 nations. *European Psychologist* **9**, 107–119 (2004).

6. White, C. J. M., Muthukrishna, M. & Norenzayan, A. Cultural similarity among coreligionists within and between countries. *Proceedings of the National Academy of Sciences* **118**, e2109650118 (2021).

7. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and Brain Sciences* **33**, 61–83 (2010).

8. Pollet, T. V., Tybur, J. M., Frankenhuis, W. E. & Rickard, I. J. What can cross-cultural correlations teach us about human nature? *Human Nature* **25**, 410–429 (2014).

9. Benhabib, J. & Spiegel, M. M. The role of human capital in economic development: Evidence from aggregate cross-country data. *Journal of Monetary Economics* **34**, 143–173 (1994).

10.   Comin, D., Easterly, W. & Gong, E. Was the wealth of nations determined in 1000 BC? *American Economic Journal: Macroeconomics* **2**, 65–97 (2010).

11.   La Porta, R., Lopez-De-Silanes, F., Shleifer, A. & Vishny, R. W. Legal determinants of external finance. *The Journal of Finance* **52**, 1131–1150 (1997).

12.   Sachs, J. D. & Warner, A. M. The curse of natural resources. *European Economic Review* **45**, 827–838 (2001).

13.   Fincher, C. L., Thornhill, R., Murray, D. R. & Schaller, M. Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society B: Biological Sciences* **275**, 1279–1285 (2008).

14.   Gelfand, M. J. *et al.* Differences between tight and loose cultures: A 33-nation study. *Science* **332**, 1100–1104 (2011).

15.   Hofstede, G. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* (Sage Publications, 2001).

16.   Inglehart, R. & Baker, W. E. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* **65**, 19–51 (2000).

17.   Awad, E. *et al.* The moral machine experiment. *Nature* **563**, 59–64 (2018).

18.   Rhoads, S. A., Gunter, D., Ryan, R. M. & Marsh, A. A. Global variation in subjective well-being predicts seven forms of altruism. *Psychological Science* **32**, 1247–1261 (2021).

19.   Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P. & Henrich, J. The Church, intensive kinship, and global psychological variation. *Science* **366**, eaau5141 (2019).

20.   Thomson, R. *et al.* Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences* **115**, 7521–7526 (2018).

21.   Kissling, W. D. & Carl, G. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography* **17**, 59–71 (2008).

22.   Hewlett, B. S., DeSilvestri, A. & Guglielmino, C. R. Semes and genes in Africa. *Current Anthropology* **43**, 313–321 (2002).

23.   Nunn, N. & Puga, D. Ruggedness: The blessing of bad geography in Africa. *The Review of Economics and Statistics* **94**, 20–36 (2012).

24.   Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**, 234–240 (1970).

25.   Tylor, E. B. On a method of investigating the development of institutions; applied to laws of marriage and descent. *The Journal of the Anthropological Institute of Great Britain and Ireland* **18**, 245–272 (1889).

26.   Naroll, R. Two solutions to Galton's Problem. *Philosophy of Science* **28**, 15–39 (1961).

27.   Naroll, R. Galton's Problem: The logic of cross-cultural analysis. *Social Research* **32**, 428–451 (1965).

28.   Murdock, G. P. & White, D. R. Standard cross-cultural sample. *Ethnology* **8**, 329–369 (1969).

29.   Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).

30.   Green, R. & Pawley, A. Early Oceanic architectural forms and settlement patterns: Linguistic, archaeological and ethnological perspectives. in *Archaeology and language III: Artefacts, languages and texts* (eds. Blench, R. & Spriggs, M.) 31–89 (Routledge, 1999).

31. Jarque, C. M. & Bera, A. K. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique* **55**, 163–172 (1987).

32. Legendre, P. Spatial autocorrelation: Trouble or new paradigm? *Ecology* **74**, 1659–1673 (1993).

33. Roberts, J., Seán AND Winters. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS ONE* **8**, 1–13 (2013).

34. Lichstein, J. W., Simons, T. R., Shriner, S. A. & Franzreb, K. E. Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs* **72**, 445–463 (2002).

35. Loftin, C. & Ward, S. K. A spatial autocorrelation model of the effects of population density on fertility. *American Sociological Review* **48**, 121–128 (1983).

36. Ferguson, J. Anthropology and its evil twin. in (eds. Cooper, F. & Packard, R.) 150–175 (University of California Press Berkeley, 1997).

37. Dow, M. M. & Eff, E. A. Global, regional, and local network autocorrelation in the standard cross-cultural sample. *Cross-Cultural Research* **42**, 148–171 (2008).

38. Eff, E. A. Does Mr. Galton still have a problem? Autocorrelation in the standard cross-cultural sample. *World Cultures* **15**, 153–170 (2004).

39. Symonds, M. R. E. & Blomberg, S. P. A primer on phylogenetic generalised least squares. in *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (ed. Garamszegi, L. Z.) 105–130 (Springer Berlin Heidelberg, 2014). doi:10.1007/978-3-662-43550-2_5.

40. Watts, J., Sheehan, O., Bulbulia, J., Gray, R. D. & Atkinson, Q. D. Christianity spread faster in small, politically structured societies. *Nature Human Behaviour* **2**, 559–564 (2018).

41. Atkinson, Q. D., Coomber, T., Passmore, S., Greenhill, S. J. & Kushnick, G. Cultural and environmental predictors of pre-European deforestation on Pacific islands. *PLOS ONE* **11**, 1–15 (2016).

42. Kelly, M. Understanding persistence. *CEPR Discussion Paper No. DP15246* (2020).

43. Bromham, L., Hua, X., Cardillo, M., Schneemann, H. & Greenhill, S. J. Parasites and politics: Why cross-cultural studies must control for relatedness, proximity and covariation. *Royal Society Open Science* **5**, 181100 (2018).

44. Bromham, L., Skeels, A., Schneemann, H., Dinnage, R. & Hua, X. There is little evidence that spicy food in hot countries is an adaptation to reducing infection risk. *Nature Human Behaviour* **5**, 878–891 (2021).

45. Currie, T. E. & Mace, R. Analyses do not support the parasite-stress theory of human sociality. *Behavioral and Brain Sciences* **35**, 83–85 (2012).

46. Passmore, S. & Watts, J. WEIRD people and the Western Church: Who made whom? *Religion, Brain & Behavior* 1–58 (2022) doi:10.1080/2153599X.2021.1991459.

47. United Nations Development Programme. *Human development report.* http://hdr.undp.org/en/composite/HDI (2021).

48. Schulz, J. F. Kin networks and institutional development. *The Economic Journal* **132**, 2578–2613 (2022).

49. Conley, T. G. GMM estimation with cross sectional dependence. *Journal of Econometrics* **92**, 1–45 (1999).

50. Conley, T. G. Spatial econometrics. in *Microeconometrics* (eds. Durlauf, S. N. & Blume, L. E.) 303–313 (Palgrave Macmillan, 2010).

51. Currie, T. E. & Mace, R. Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Sciences* **106**, 7339–7344 (2009).

52. McElreath, R. *Statistical rethinking: A Bayesian course with examples in R and Stan.* (CRC Press, 2020).

53. Neal, R. M. Regression and classification using Gaussian process priors. in *Bayesian statistics* (eds. Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M.) vol. 6 475–501 (Oxford University Press).

54. Villemereuil, P. de & Nakagawa, S. General quantitative genetic methods for comparative biology. in *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (ed. Garamszegi, L. Z.) 287–303 (Springer Berlin Heidelberg, 2014). doi:10.1007/978-3-662-43550-2_11.

55. Beck, T., Demirgäç-Kunt, A. & Levine, R. Law, endowments, and finance. *Journal of Financial Economics* **70**, 137–181 (2003).

56. Beck, T., Demirgäç-Kunt, A. & Levine, R. SMEs, growth, and poverty: Cross-country evidence. *Journal of Economic Growth* **10**, 199–229 (2005).

57. Bockstette, V., Chanda, A. & Putterman, L. States and markets: The advantage of an early start. *Journal of Economic growth* **7**, 347–369 (2002).

58. Easterly, W. & Levine, R. Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics* **50**, 3–39 (2003).

59. Easterly, W. Inequality does cause underdevelopment: Insights from a new instrument. *Journal of Development Economics* **84**, 755–776 (2007).

60. Skidmore, M. & Toya, H. Do natural disasters promote long-run growth? *Economic Inquiry* **40**, 664–687 (2002).

61. Adamczyk, A. & Pitt, C. Shaping attitudes about homosexuality: The role of religion and cultural context. *Social Science Research* **38**, 338–351 (2009).

62. Alesina, A., Giuliano, P. & Nunn, N. On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics* **128**, 469–530 (2013).

63.   Knack, S. & Keefer, P. Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics* **112**, 1251–1288 (1997).

64.   Sookias, R. B., Passmore, S. & Atkinson, Q. D. Deep cultural ancestry and human development indicators across nation states. *Royal Society Open Science* **5**, 171411 (2018).

65.   Voth, H.-J. Persistence – myth and mystery. in *The handbook of historical economics* (eds. Bisin, A. & Federico, G.) 243–267 (Academic Press, 2021). doi:10.1016/B978-0-12-815874-6.00015-0.

66.   Smith, M. D., Rabbitt, M. P. & Coleman- Jensen, A. Who are the world's food insecure? New evidence from the Food and Agriculture Organization's Food Insecurity Experience Scale. *World Development* **93**, 402–412 (2017).

67.   Hastie, T. J. & Tibshirani, R. J. *Generalized additive models.* (Routledge, 2017).

68.   Muthukrishna, M. *et al.* Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science* **31**, 678–701 (2020).

69.   Spolaore, E. & Wacziarg, R. The diffusion of development. *The Quarterly Journal of Economics* **124**, 469–529 (2009).

70.   Kyritsis, T., Matthews, L. J., Welch, D. & Atkinson, Q. D. Shared cultural ancestry predicts the global diffusion of democracy. *Evolutionary Human Sciences* **4**, e42 (2022).

71.   Matthews, L. J., Passmore, S., Richard, P. M., Gray, R. D. & Atkinson, Q. D. Shared cultural history as a predictor of political and economic changes among nation states. *PLOS ONE* **11**, 1–18 (2016).

72.   Trouillot, M. *Global transformations: Anthropology and the modern world.* (Springer, 2016).

73.   Wolf, E. R. Europe and the people without history. *University of California Press* (2010).

74.   Smaldino, P. E. & McElreath, R. The natural selection of bad science. *Royal Society Open Science* **3**, 160384 (2016).

75.   Liddell, T. M. & Kruschke, J. K. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* **79**, 328–348 (2018).

76.   Nicholson, J. S., Deboeck, P. R. & Howard, W. Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development* **41**, 143–153 (2017).

77.   Tipton, E., Pustejovsky, J. E. & Ahmadi, H. Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods* **10**, 180–194 (2019).

78.   Chambers, C. D. & Tzavella, L. The past, present and future of Registered Reports. *Nature Human Behaviour* **6**, 29–42 (2021).

79.   Deffner, D., Rohrer, J. M. & McElreath, R. A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science* **5**, 25152459221106366 (2022).

80.   Stan Development Team. RStan: The R interface to Stan. (2020).

81.   Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).

82.   Oishi, S. Socioecological psychology. *Annual Review of Psychology* **65**, 581–609 (2014).

83.   Gelfand, M. J. *et al.* The relationship between cultural tightness–looseness and COVID-19 cases and deaths: A global analysis. *The Lancet Planetary Health* **5**, e135–e144 (2021).

84.   Brownrigg, R. *maps: Draw geographical maps.* (2018).

85.   Hijmans, R. J. *geosphere: Spherical trigonometry.* (2019).

86.   Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. *Glottolog 3.0.* (Max Planck Institute for the Science of Human History, 2017). doi:10.5281/zenodo.4061162.

87.   Eberhard, D. M., Simons, G. F. & Fennig, C. D. *Ethnologue: Languages of the world.* (SIL International, 2018).

88.   Eff, E. A. Weight matrices for cultural proximity: Deriving weights from a language phylogeny. *Structure and Dynamics* **3**, (2008).

89.   Page, M. J. *et al.* The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**, (2021).

90.   Tiefelsdorf, M. & Griffith, D. A. Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environment and Planning A: Economy and Space* **39**, 1193–1221 (2007).

91.   De Hoyos, R. E. & Sarafidis, V. Testing for cross-sectional dependence in panel-data models. *The Stata Journal* **6**, 482–496 (2006).

92.   Colella, F., Lalive, R., Sakalli, S. O. & Thoenig, M. Inference with arbitrary clustering. *IZA Discussion Paper No. 12584* (2019) doi:10.2139/ssrn.3449578.

93.   Düben, C. *conleyreg: Estimations using Conley standard errors.* (2021).

94.   Solin, A. & Särkkä, S. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing* **30**, 419–446 (2020).

95. Claessens, S., Kyritsis, T. & Atkinson, Q. D. Cross-national analyses require additional controls to account for the non-independence of nations - ScottClaessens/crossNationalCorrelations: v1.0.0. (2023) doi:10.5281/zenodo.8128702.

96. Landau, W. M. The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* **6**, 2959 (2021).

97. Aust, F. & Barth, M. *papaja: Create APA manuscripts with R Markdown.* (2020).

Acknowledgements

Author Contributions Statement

SC and QDA conceived of and designed the study. SC curated the data, produced all code for analysis and visualisation, and wrote the original draft of the manuscript. TK and QDA developed and compiled the geographic and linguistic distance matrices. QDA provided funding and input on manuscript preparation and revision. All authors reviewed and edited the final draft of the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Supplementary Information

Cross-national analyses require additional controls to account for the non-independence of nations

Scott Claessens[1], Thanos Kyritsis[1], & Quentin D. Atkinson[1,2]

[1] School of Psychology, University of Auckland, New Zealand

[2] School of Anthropology and Museum Ethnography, University of Oxford, United Kingdom

Supplementary Methods

Bayesian models for reanalysis. We provide model formulae for our reanalyses of cross-national correlations, for a general bivariate case with standardised outcome Y and predictor X variables. In the naive regression model without controls for non-independence:

$$Y_i \sim \text{Normal}(\mu_i, \sigma) \tag{S1}$$

$$\mu_i = \alpha + \beta X_i$$

$$\alpha \sim \text{Normal}(0, 0.4)$$

$$\beta \sim \text{Normal}(0, 0.4)$$

$$\sigma \sim \text{Exponential}(5)$$

The priors in this model were arrived at by prior predictive checks, with wider priors making predictions beyond the scale of standardised outcome variables and narrower priors being too informative.

To control for spatial non-independence, we add a Gaussian process to this model and feed it a scaled geographic distance matrix D based on Euclidean distances between latitude and longitude coordinates. This distance matrix is computed internally by the *brms* R package[1]. The Gaussian process uses an exponentiated quadratic covariance kernel, the only covariance kernel currently supported by *brms*. The model formula is:

$$Y_i \sim \text{Normal}(\mu_i, \sigma) \tag{S2}$$

$$\mu_i = \alpha + \kappa_{\text{NATION}[i]} + \beta X_i$$

$$\begin{pmatrix} \kappa_1 \\ \kappa_2 \\ ... \\ \kappa_n \end{pmatrix} \sim \text{MVNormal}\left(\begin{pmatrix} 0 \\ 0 \\ ... \\ 0 \end{pmatrix}, \mathbf{K} \right)$$

$$\mathbf{K}_{ij} = sdgp^2 \exp\left(-D_{ij}^2/(2lscale^2) \right)$$

$$\alpha \sim \text{Normal}(0, 0.4)$$

$$\beta \sim \text{Normal}(0, 0.4)$$

$$\sigma \sim \text{Exponential}(5)$$

$$sdgp \sim \text{Exponential}(5)$$

$$lscale \sim \text{InverseGamma}(?, ?)$$

where n is the number of nations, and D_{ij}^2 reflects the squared Euclidean distances between latitude and longitude coordinates for the i-th and j-th nations. Notice that the inverse gamma prior on *lscale* is left undetermined. This is because the *brms* package intelligently tunes the prior for this parameter based on the covariates of the Gaussian process (see https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html).

To control for cultural phylogenetic non-independence, we manually specify the covariance structure for nation random intercepts using a pre-computed linguistic proximity matrix L (see previous section). The covariance between two nations is assumed to be linearly proportional to the linguistic proximity between those nations. This assumption is justified if we assume that cultural traits evolve neutrally via Brownian motion along a language phylogeny. Although this is a conservative assumption, we follow this approach here rather

than another Gaussian Process for several reasons: (1) to avoid model non-convergence when later including a spatial Gaussian Process in the same model, (2) to work around the lack of a coordinate system for linguistic distances, as opposed to latitude and longitude values for geographic distance, and (3) to showcase different ways that researchers can allow nations to covary to control for non-independence.

The non-centered parameterisation of this model is:

$$Y_i \sim \text{Normal}(\mu_i, \sigma) \tag{S3}$$

$$\mu_i = \alpha + z_{\text{NATION}[i]}\sigma_\alpha L + \beta X_i$$

$$\alpha \sim \text{Normal}(0, 0.4)$$

$$\beta \sim \text{Normal}(0, 0.4)$$

$$z_j \sim \text{Normal}(0, 1)$$

$$\sigma_\alpha \sim \text{Exponential}(5)$$

$$\sigma \sim \text{Exponential}(5)$$

Finally, we can control for spatial and cultural phylogenetic non-independence simultaneously by including both a Gaussian process over latitude and longitude coordinates *and* nation random intercepts that covary according to linguistic proximity. The resulting model is as follows:

$$Y_i \sim \text{Normal}(\mu_i, \sigma) \tag{S4}$$

$$\mu_i = \alpha + \kappa_{\text{NATION}[i]} + z_{\text{NATION}[i]}\sigma_\alpha L + \beta X_i$$

$$\begin{pmatrix} \kappa_1 \\ \kappa_2 \\ \dots \\ \kappa_n \end{pmatrix} \sim \text{MVNormal}\left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \mathbf{K} \right)$$

$$\mathbf{K}_{ij} = sdgp^2 \exp\left(-D_{ij}^2 / (2 lscale^2) \right)$$

$$\alpha \sim \text{Normal}(0, 0.4)$$

$$\beta \sim \text{Normal}(0, 0.4)$$

$$z_j \sim \text{Normal}(0, 1)$$

$$\sigma_\alpha \sim \text{Exponential}(5)$$

$$\sigma \sim \text{Exponential}(5)$$

$$sdgp \sim \text{Exponential}(5)$$

$$lscale \sim \text{InverseGamma}(?, ?)$$

Supplementary Figures
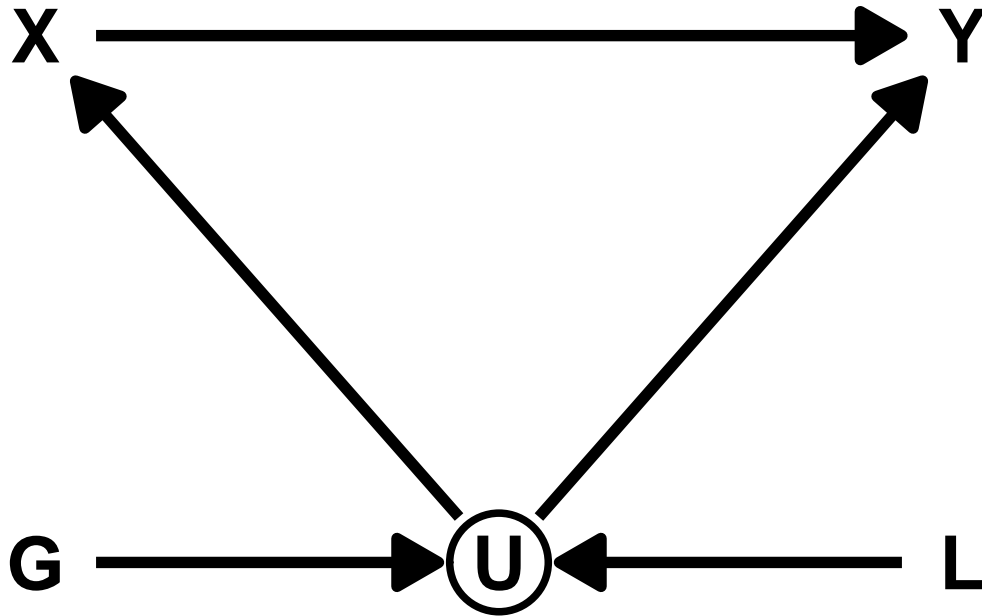
Supplementary Figure 1. A causal directed acyclic graph of spatial and cultural phylogenetic non-independence in cross-national studies. We are interested in estimating the direct effect of national-level exposure X on national-level outcome Y. But these variables are confounded by their common unobserved cause U. U is a stand-in for shared environmental, ecological, and geographic causes (e.g. climate, biodiversity, physical topography) and cultural and institutional causes (e.g. cultural norms, technologies, and institutions). In this causal model, we need to condition on U to estimate the direct path from X to Y, but we cannot since it is unobserved. However, geographic G and linguistic L relationships between societies influence U, since changing a nation's spatial distance to or shared cultural ancestry with other nations will change its environmental and cultural traits. We can thus use G and L to model the covariation between X and Y induced by U. Failing to do this and simply estimating the bivariate correlation between X and Y will produce spurious relationships and residuals that are spatially and culturally non-independent around the world.
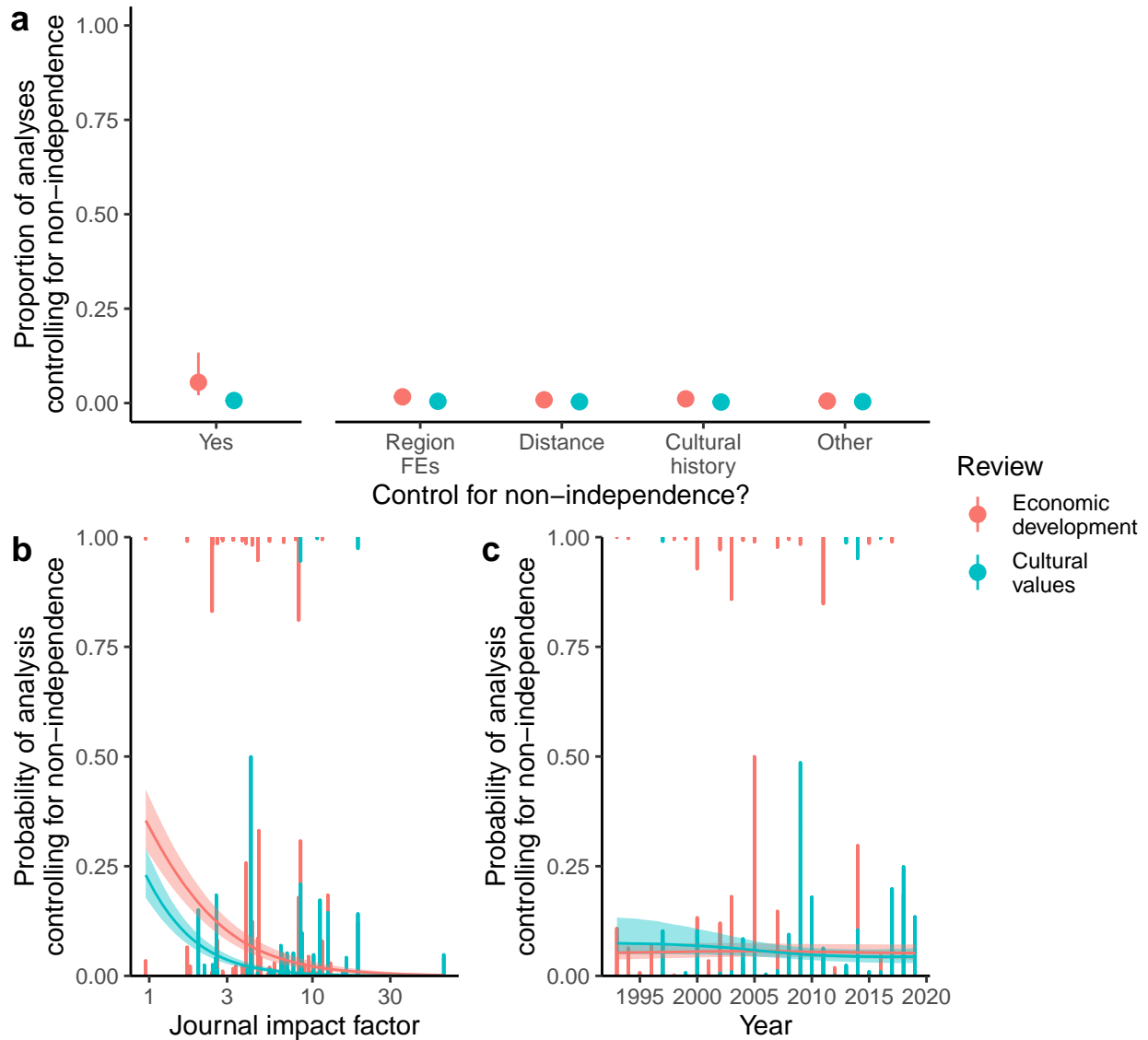
Supplementary Figure 2. *Analysis-level results from literature review of 100 highly-cited cross-national studies of economic development (red) and cultural values (blue).* (a) Proportion of analyses accounting for non-independence, overall and split by common methods of controlling for non-independence. Points represent posterior median proportions and ranges represent equal-tailed 95% credible intervals (n = 4308 observations). (b) The association between journal impact factor and the probability that an analysis accounts for non-independence. (c) Estimated trend over time for the probability that an analysis accounts for non-independence. Lines and shaded areas are posterior median regression lines and equal-tailed 50% credible intervals from Bayesian multilevel models (n = 4308 observations). Histograms represent relative counts for individual analyses that did (top) or did not (bottom) account for non-independence. Region FEs = region fixed effects.
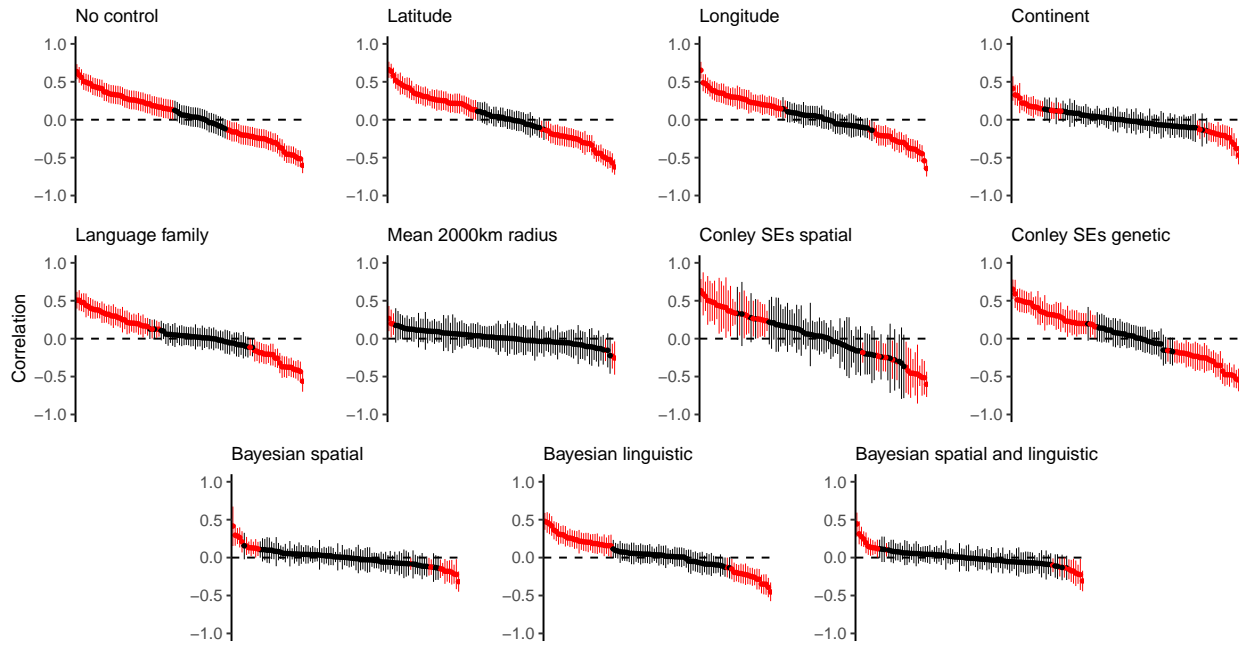
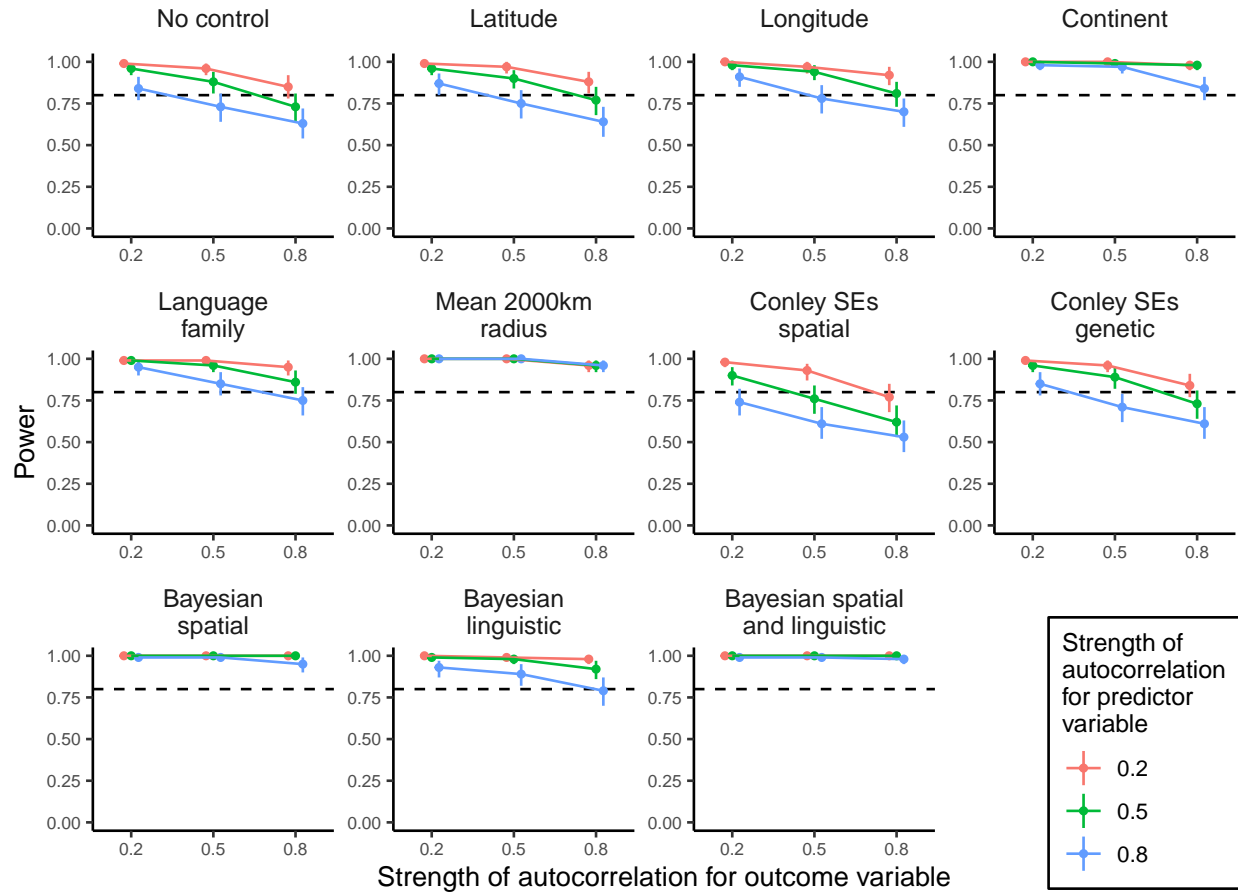*Supplementary Figure 3. Distribution of cross-national correlations from simulation study under strong spatial autocorrelation.* In these simulations, the strength of spatial autocorrelation is set to 0.8 for both outcome and predictor variables, and the true correlation is set to 0. For frequentist regression models, points represent correlation estimates and ranges represent two-tailed 95% confidence intervals (n = 236 observations). For Bayesian regression models, points represent posterior means and ranges represent equal-tailed 95% credible intervals (n = 236 observations). Correlations are ordered by effect size independently in each panel. Red point ranges indicate that the 95% confidence / credible interval excludes zero. Black point ranges indicate that the 95% confidence / credible interval includes zero. SEs = standard errors.
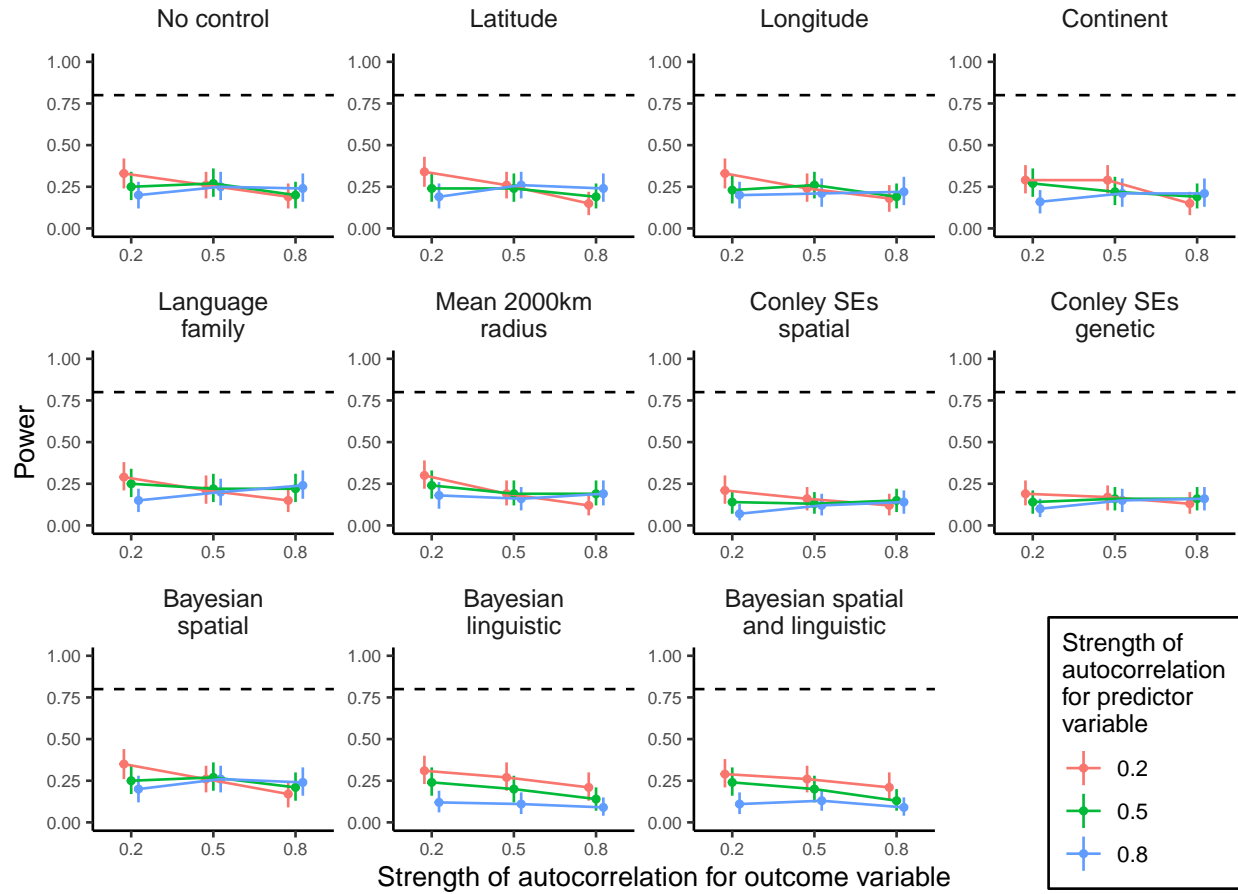
Supplementary Figure 4. Distribution of cross-national correlations from simulation study under strong cultural phylogenetic autocorrelation. In these simulations, the strength of cultural phylogenetic autocorrelation is set to 0.8 for both outcome and predictor variables, and the true correlation is set to 0. For frequentist regression models, points represent correlation estimates and ranges represent two-tailed 95% confidence intervals (n = 236 observations). For Bayesian regression models, points represent posterior means and ranges represent equal-tailed 95% credible intervals (n = 236 observations). Correlations are ordered by effect size independently in each panel. Red point ranges indicate that the 95% confidence / credible interval excludes zero. Black point ranges indicate that the 95% confidence / credible interval includes zero. SEs = standard errors.
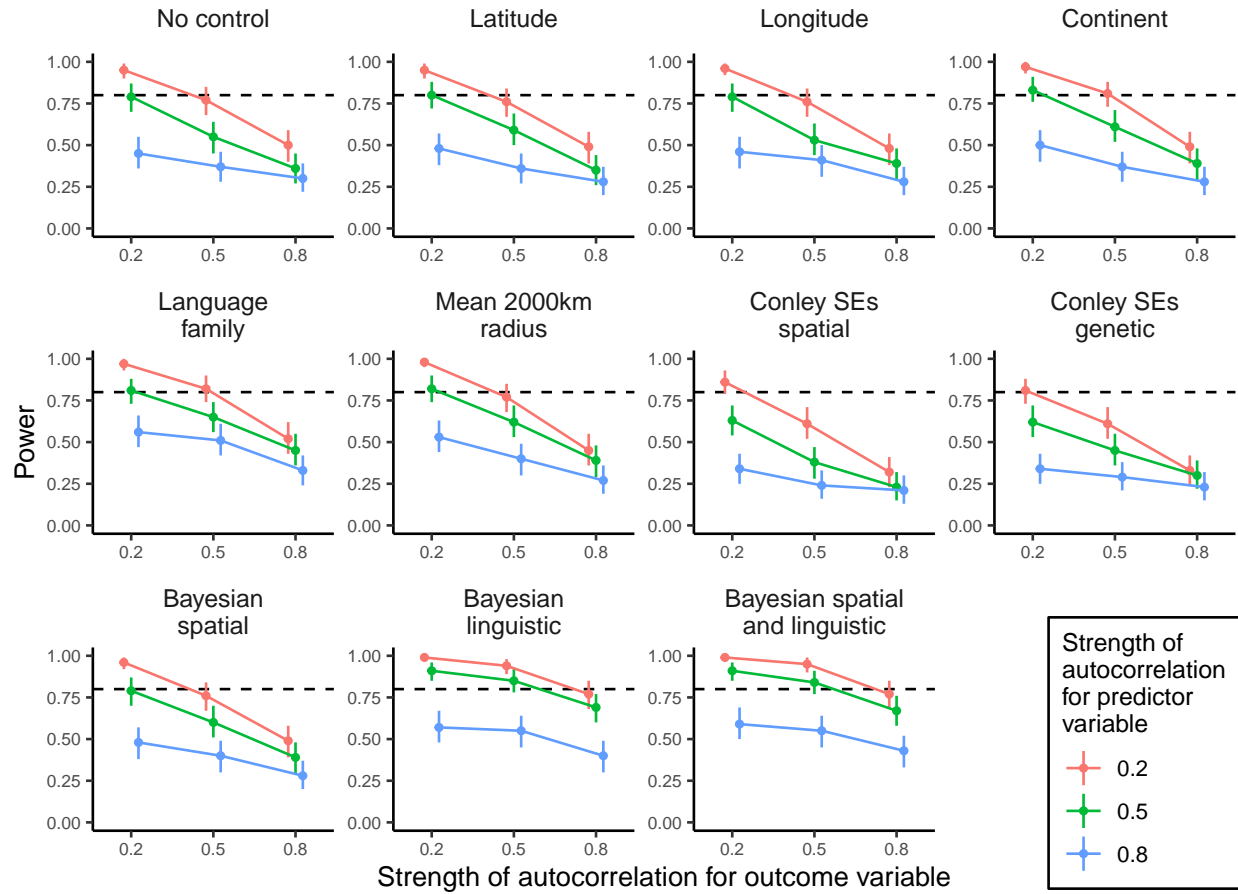
Supplementary Figure 5. Statistical power estimates for different methods of controlling for spatial non-independence in our simulation study, assuming a small true effect size (r = 0.1). For simulated outcome and predictor variables, we systematically varied the strength of spatial autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a small true correlation between variables ($r = 0.1$) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.
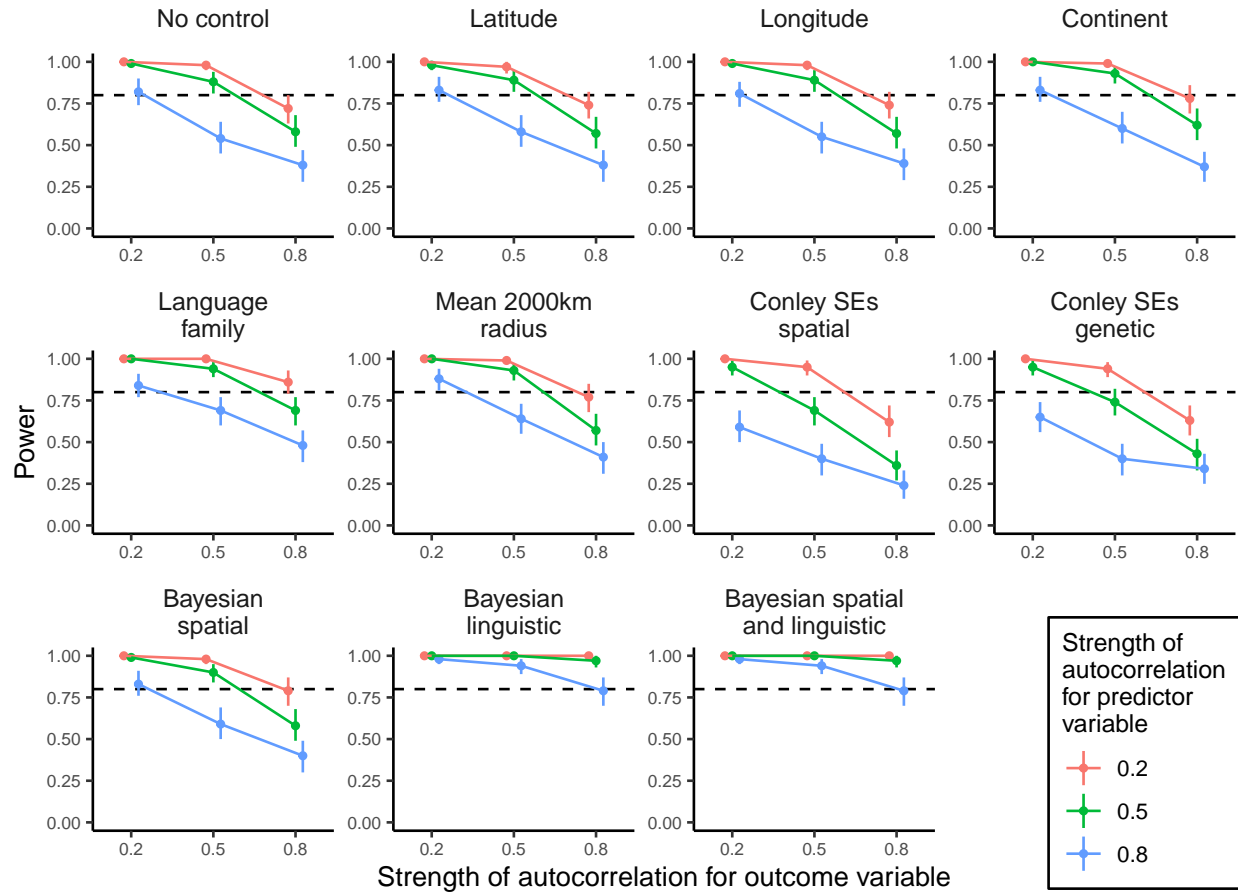
Supplementary Figure 6. Statistical power estimates for different methods of controlling for spatial non-independence in our simulation study, assuming a medium true effect size (r = 0.3). For simulated outcome and predictor variables, we systematically varied the strength of spatial autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a medium true correlation between variables ($r = 0.3$) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.

Supplementary Figure 7. Statistical power estimates for different methods of controlling for spatial non-independence in our simulation study, assuming a large true effect size (r = 0.5). For simulated outcome and predictor variables, we systematically varied the strength of spatial autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a large true correlation between variables ($r = 0.5$) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.
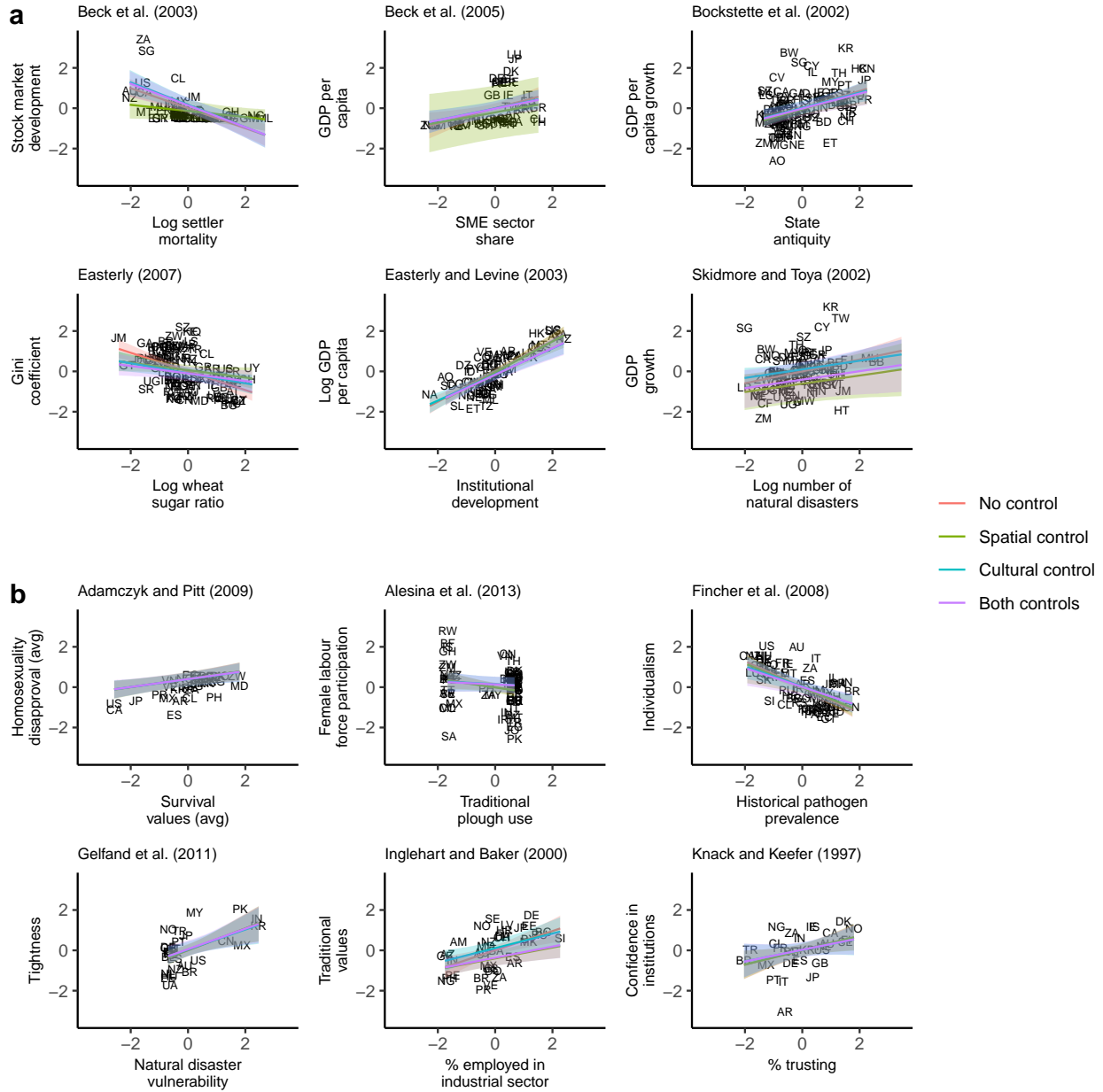
Supplementary Figure 8. Statistical power estimates for different methods of controlling for cultural phylogenetic non-independence in our simulation study, assuming a small true effect size (r = 0.1). For simulated outcome and predictor variables, we systematically varied the strength of cultural phylogenetic autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a small true correlation between variables ($r = 0.1$) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.

Supplementary Figure 9. Statistical power estimates for different methods of controlling for cultural phylogenetic non-independence in our simulation study, assuming a medium true effect size (r = 0.3). For simulated outcome and predictor variables, we systematically varied the strength of cultural phylogenetic autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a medium true correlation between variables ($r = 0.3$) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.

Supplementary Figure 10. Statistical power estimates for different methods of controlling for cultural phylogenetic non-independence in our simulation study, assuming a large true effect size (r = 0.5). For simulated outcome and predictor variables, we systematically varied the strength of cultural phylogenetic autocorrelation, from weak (0.2) to moderate (0.5) to strong (0.8). We simulated 100 datasets per parameter combination assuming a large true correlation between variables (*r* = 0.5) and fitted different models to each dataset. Statistical power was operationalised as the proportion of models that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero. Points represent raw proportions of models with slope 95% confidence / credible intervals excluding zero, ranges represent two-tailed 95% bootstrap confidence intervals (n = 1000 bootstrap samples), and dashed lines indicate 80% power. Colours indicate whether the strength of autocorrelation for the predictor variable is 0.2 (red), 0.5, (green) or 0.8 (blue). SEs = standard errors.
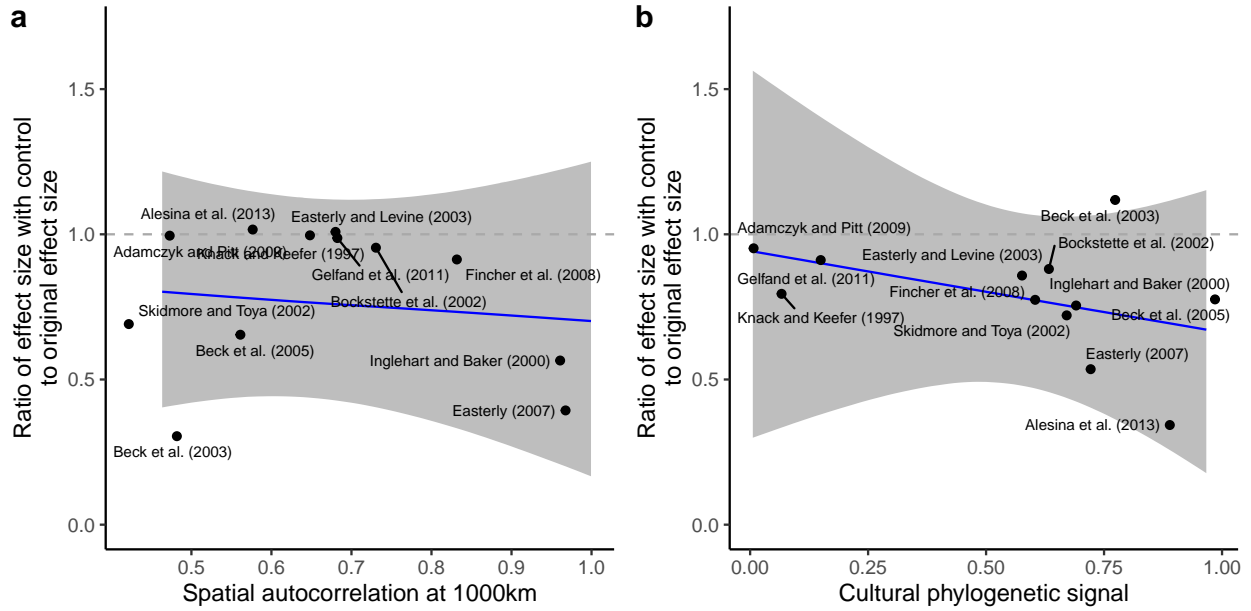
Supplementary Figure 11. *Reanalysis models fitted to raw data, for economic development (a) and cultural values (b) studies.* Data points are labelled using ISO 3166-1 alpha-2 letter country codes. In all reanalyses, outcome and predictor variables are standardised, making regression slopes comparable to Pearson's correlation coefficients. Lines and shaded areas represent posterior median regression lines and equal-tailed 95% credible intervals. Colours indicate predictions from models with no control (red), spatial control only (green), cultural control only (blue), or both spatial and cultural control (purple). For models with covariates (Adamzyck and Pitt 2009; Gelfand et al. 2011), marginal effects are presented holding all covariates at zero or their reference categories. GDP = gross domestic product; SME = small and medium-sized enterprise.
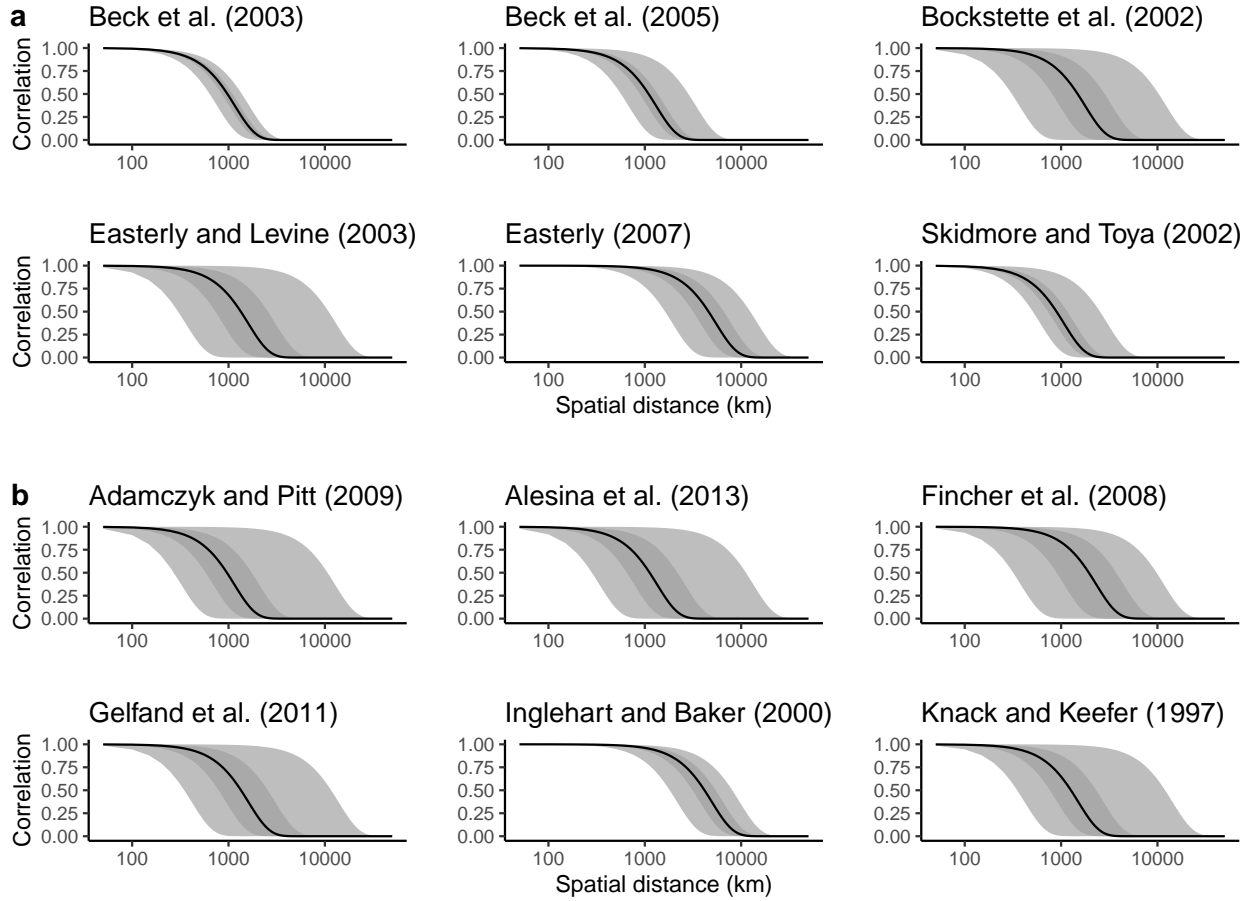
Supplementary Figure 12. Posterior estimates of Gaussian process functions mapping spatial auto-correlation onto geographic distance from our reanalyses of economic development (a) and cultural values (b) studies. The y-axis represents the amount of spatial autocorrelation between data points with increasing distance between those points on the x-axis (logged distance in kilometres). Lines and shaded areas represent median posterior spatial autocorrelation functions and equal-tailed 50% and 95% credible intervals.
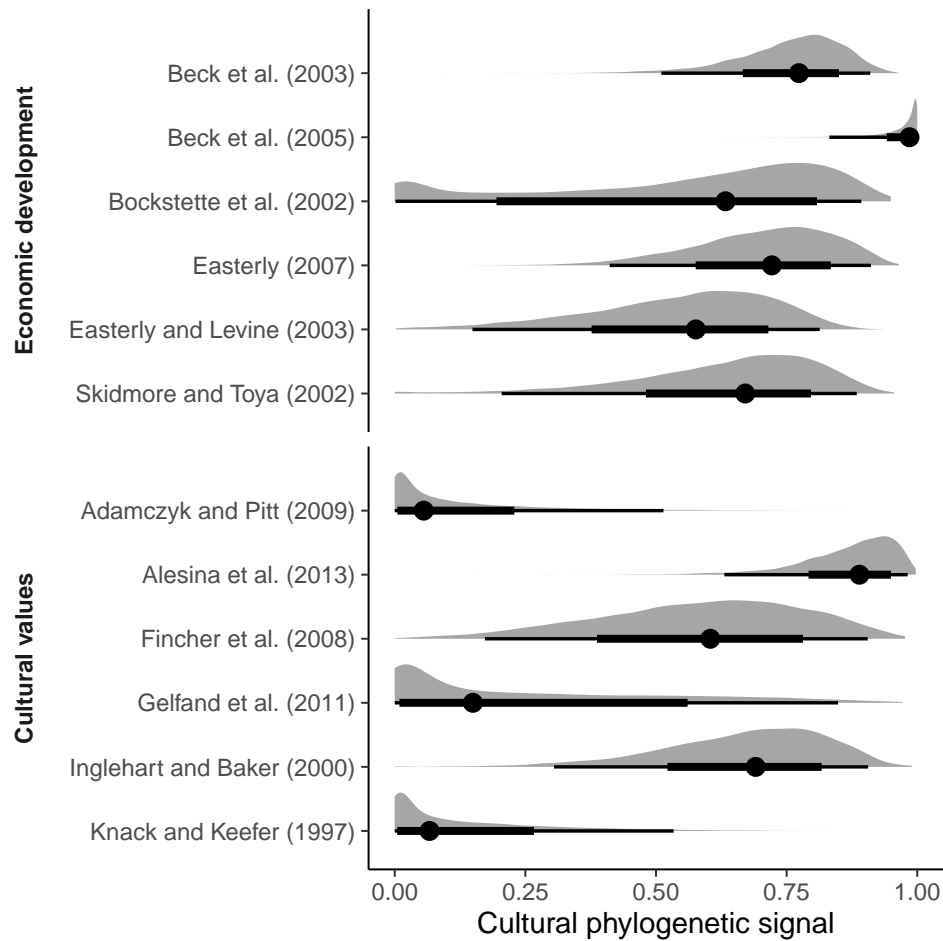
Supplementary Figure 13. Posterior estimates of cultural phylogenetic signal from our reanalyses. Cultural phylogenetic signal is operationalised as the proportion of national-level variance explained by linguistic proximity between nations. Ridges are full posterior distributions, points are posterior medians, and lines represent equal-tailed 50% and 95% credible intervals. Numbers of observations from the models are as follows, from top to bottom: n = 69, n = 45, n = 103, n = 98, n = 63, n = 89, n = 33, n = 75, n = 67, n = 28, n = 38, and n = 28.

Supplementary Figure 14. *The estimated degree of spatial and cultural phylogenetic non-independence predicts reductions in effect size in our reanalysis.* (a) Higher estimated degrees of spatial autocorrelation at 1,000 km distance predict more pronounced reductions in effect sizes when controlling for non-independence. (b) Higher estimated levels of cultural phylogenetic signal predict more pronounced reductions in effect sizes when controlling for non-independence. In both panels, the y-axis represents the ratio of the effect size when controlling for spatial and cultural non-independence to the original effect size (from naive regression model), and the x-axis represents posterior median model estimates. Regression lines are plotted with equal-tailed 95% credible intervals.

Supplementary Tables

Supplementary Table 1

Geographic and cultural phylogenetic signal results for economic development and cultural values variables. Signal estimates reflect the proportion of national-level variance explained by geographic and linguistic covariance matrices. Bayes Factors (BF) reflect support for the hypothesis that the signal estimate differs from zero. HDI = Human Development Index; GDPpc = gross domestic product per capita.

Outcome	Geographic signal	Cultural phylogenetic signal
HDI	0.37, 95% CI [0.18, 0.59], BF > 100	0.62, 95% CI [0.40, 0.80], BF > 100
GDPpc	0.42, 95% CI [0.20, 0.66], BF > 100	0.56, 95% CI [0.33, 0.78], BF > 100
GDPpc growth	0.65, 95% CI [0.09, 0.98], BF = 16.79	0.26, 95% CI [0.00, 0.70], BF = 1.16
Gini index	0.74, 95% CI [0.48, 0.97], BF > 100	0.25, 95% CI [0.02, 0.51], BF = 3.97
Traditional values	0.44, 95% CI [0.17, 0.76], BF > 100	0.54, 95% CI [0.23, 0.79], BF > 100
Survival values	0.20, 95% CI [0.01, 0.45], BF = 1.96	0.78, 95% CI [0.53, 0.96], BF > 100
Tightness	0.09, 95% CI [0.00, 0.32], BF = 0.33	0.84, 95% CI [0.61, 0.97], BF > 100
Individualism	0.26, 95% CI [0.00, 0.66], BF = 1.97	0.69, 95% CI [0.29, 0.96], BF > 100

Supplementary Table 2

False positive rates in simulation with varying levels of spatial autocorrelation for the predictor variable (ρ) and for the outcome variable (λ), assuming that the true correlation is 0. Numbers represent the total number of analyses, out of 100, that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero, falsely inferring a relationship when none is present.

Model	ρ = 0.2			ρ = 0.5			ρ = 0.8		
	λ = 0.2	λ = 0.5	λ = 0.8	λ = 0.2	λ = 0.5	λ = 0.8	λ = 0.2	λ = 0.5	λ = 0.8
No control	18	27	41	27	48	61	40	61	77
Latitude	19	25	39	25	44	60	33	58	71
Longitude	14	20	27	22	34	48	29	49	62
Continent	10	13	16	13	18	24	13	25	35
Language family	12	17	29	20	33	46	27	48	58
Mean 2000km radius	9	12	19	5	9	13	5	5	6
Conley SEs spatial	13	13	14	18	23	25	22	37	43
Conley SEs genetic	17	22	27	20	34	51	28	49	66
Bayesian spatial	14	13	11	13	15	12	16	17	23
Bayesian linguistic	12	20	21	21	26	39	28	42	49
Bayesian spatial & linguistic	11	13	9	14	13	12	17	18	21

Supplementary Table 3

False positive rates in simulation with varying levels of cultural phylogenetic autocorrelation for the predictor variable (ρ) and for the outcome variable (λ), assuming that the true correlation is 0. Numbers represent the total number of analyses, out of 100, that estimated a slope with a two-tailed 95% confidence / credible interval excluding zero, falsely inferring a relationship when none is present.

Model	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 0.8$	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 0.8$	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 0.8$
No control	7	14	15	8	19	30	15	26	36
Latitude	7	10	16	10	19	27	15	22	35
Longitude	5	12	14	9	16	26	13	26	35
Continent	5	6	10	3	11	21	8	20	29
Language family	7	9	12	9	14	22	10	20	32
Mean 2000km radius	5	8	12	2	12	21	4	16	25
Conley SEs spatial	2	6	7	2	9	15	6	14	20
Conley SEs genetic	5	11	13	8	13	21	8	17	26
Bayesian spatial	7	12	12	9	17	25	14	24	32
Bayesian linguistic	5	9	11	6	7	7	4	5	5
Bayesian spatial & linguistic	3	10	11	5	8	8	4	5	5

Supplementary Table 4

Estimated cross-national correlations from our reanalysis of twelve previous cross-national analyses. Numbers are posterior median slopes (equivalent to Pearson's r correlation coefficients) with equal-tailed 95% credible intervals.

Analysis	No control	Spatial control	Cultural control	Both controls
Adamczyk and Pitt (2009)	0.20, 95% CI [0.07, 0.33]	0.20, 95% CI [0.07, 0.34]	0.19, 95% CI [0.05, 0.33]	0.20, 95% CI [0.06, 0.33]
Alesina et al. (2013)	-0.21, 95% CI [-0.42, 0.01]	-0.21, 95% CI [-0.43, 0.01]	-0.07, 95% CI [-0.26, 0.11]	-0.07, 95% CI [-0.25, 0.11]
Beck et al. (2003)	-0.50, 95% CI [-0.68, -0.30]	-0.15, 95% CI [-0.38, 0.02]	-0.56, 95% CI [-0.77, -0.33]	-0.54, 95% CI [-0.76, -0.30]
Beck et al. (2005)	0.38, 95% CI [0.12, 0.64]	0.25, 95% CI [-0.01, 0.50]	0.29, 95% CI [0.10, 0.49]	0.29, 95% CI [0.09, 0.48]
Bockstette et al. (2002)	0.43, 95% CI [0.25, 0.59]	0.41, 95% CI [0.19, 0.59]	0.37, 95% CI [0.18, 0.57]	0.37, 95% CI [0.15, 0.56]
Easterly and Levine (2003)	0.75, 95% CI [0.60, 0.90]	0.76, 95% CI [0.57, 0.93]	0.65, 95% CI [0.47, 0.82]	0.66, 95% CI [0.47, 0.84]
Easterly (2007)	-0.46, 95% CI [-0.63, -0.29]	-0.18, 95% CI [-0.36, 0.01]	-0.25, 95% CI [-0.41, -0.09]	-0.19, 95% CI [-0.36, -0.01]
Fincher et al. (2008)	-0.64, 95% CI [-0.82, -0.46]	-0.59, 95% CI [-0.80, -0.21]	-0.50, 95% CI [-0.71, -0.27]	-0.47, 95% CI [-0.69, -0.18]
Gelfand et al. (2011)	0.54, 95% CI [0.20, 0.85]	0.54, 95% CI [0.19, 0.86]	0.49, 95% CI [0.14, 0.83]	0.50, 95% CI [0.15, 0.83]
Inglehart and Baker (2000)	0.48, 95% CI [0.21, 0.74]	0.27, 95% CI [0.11, 0.44]	0.36, 95% CI [0.11, 0.60]	0.28, 95% CI [0.10, 0.46]
Knack and Keefer (1997)	0.36, 95% CI [0.05, 0.67]	0.36, 95% CI [0.04, 0.66]	0.28, 95% CI [-0.10, 0.63]	0.29, 95% CI [-0.09, 0.62]
Skidmore and Toya (2002)	0.30, 95% CI [0.10, 0.49]	0.20, 95% CI [0.00, 0.41]	0.21, 95% CI [0.03, 0.40]	0.21, 95% CI [0.01, 0.41]

Supplementary Table 5

List of sample size deviations between original analyses and reanalyses.

Analysis	Original N	Reanalysis N	Reason for deviation
Adamczyk and Pitt (2009)	33	33	N/A
Alesina et al. (2013)	>177	75	Original analysis used imputed values for traditional plough use which were unavailable to us. For our reanalysis, we collected data from Tables A4 and A9 and ran our regression on the 75 nations with available data for female labour force participation in 2000 and traditional plough use.
Beck et al. (2003)	70	69	We removed one nation (Zaire) as it was not present in our linguistic distance matrix.
Beck et al. (2005)	45	45	N/A
Bockstette et al. (2002)	94	103	After manually linking state history data from Appendix A with world development indicators, we had complete data for 103 nations.
Easterly and Levine (2003)	72	63	After manually linking log GDP per capita 1995 data from Appendix Table A2 in Acemoglu et al. (2001) with an institutions index self-constructed from world governance indicators, we only had complete data for 63 nations.
Easterly (2007)	118	98	After manually linking log wheat-sugar ratio data from Appendix A with longitudinal Gini data, we only had complete data for 98 nations.
Fincher et al. (2008)	68	67	We combined England and Northern Ireland into a single nation (United Kingdom) for the reanalysis.
Gelfand et al. (2011)	30	28	We combined East and West Germany into a single nation (Germany) and removed a high leverage point with a substantially higher disaster rate than other nations (Venezuela).
Inglehart and Baker (2000)	65	38	After manually linking traditional values data from the World Values Survey with data on percentage total employment in industry from World Development Indicators, we only had complete data for 38 nations.
Knack and Keefer (1997)	28	28	N/A
Skidmore and Toya (2002)	89	89	N/A

Supplementary References

1. Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).