Controls for non-independence should reflect the data-generating process

Scott Claessens* Thanos Kyritsis† Quentin D Atkinson‡

January 16, 2025

In our article[1], we simulated national-level trait data with non-independence due to spatial diffusion or shared language ancestry. These are plausible scenarios from which we can evaluate the ability of different methods to recover the "true" cross-national correlation. We used these simulations to evaluate the performance of widely-used statistical controls for non-independence and showed that these methods do not result in a satisfactory reduction in false positives. This finding holds and is concerning regardless of our ability to identify an alternative class of models that performs better. We then further showed that Bayesian random effects models incorporating explicit assumptions about the data-generating process do indeed perform much better. In particular, the Bayesian model controlling for spatial proximity was most effective at reducing false positives in spatially non-independent data, and the Bayesian model controlling for linguistic proximity was most effective at reducing false positives in "language tree" non-independent data.

In their Matters Arising, Wolfer and Koplenig[2] (henceforth W&K) seem to imply that we conclude from this simulation that models controlling for spatial and linguistic proximity should *always* outperform alternatives and should therefore *always* be the models of choice when analysing autocorrelated cross-national data. To rebut this claim, W&K simulate cross-national data with another form of non-independence: autocorrelation due to shared religious ancestry. From these simulation results, W&K argue that models controlling for spatial and linguistic proximity are no more

*School of Psychology, University of Auckland; School of Psychology, University of Kent; scott.claessens@gmail.com
†School of Psychology, University of Auckland
‡School of Psychology, University of Auckland

effective at reducing false positives compared to models with alternative controls. W&K conclude that our original simulation results were due purely to a methodological artefact that they label "data leakage", where the same covariance matrix is used to generate and analyse the data.

To be clear, we never claimed that models controlling for spatial and linguistic proximity should always be the model of choice when analysing cross-national data. In fact, we explicitly caution against this in our article: "we did not simulate other sources of non-independence that potentially exist in real cross-national datasets… additional controls will be required to ensure that these sources of non-independence do not confound cross-national inferences"[1] (p. 8).

We expand on this point here. Instead of being used as a one-size-fits-all recipe, controls for non-independence should reflect the data-generating process for the particular dataset at hand. As we would expect, if you simulate data with one source of non-independence and run regressions controlling for a different source of non-independence, the performance will be affected. In our own simulations, for example, we already showed that controls for spatial proximity did not necessarily deal with linguistic non-independence, and controls for linguistic proximity did not necessarily deal with spatial non-independence.

The same is true for the simulations from W&K. While we agree that there is a relationship between linguistic and religious traits, the covariation implied by these two forms of cultural ancestry are different, resulting in only a moderate correlation between linguistic and religious proximities in nations around the world ($r = 0.38$) and an even weaker correlation between spatial and religious proximities ($r = 0.20$). Given these differences, it is no surprise that models controlling for spatial and linguistic proximity do not eliminate false positives when dealing with autocorrelation due to religious ancestry. Space and language are poor proxy controls for religion.

It is interesting though that the costs of getting this wrong are not huge. While models controlling for spatial and linguistic proximity fail to completely eliminate false positives in the religious-autocorrelation simulations from W&K, they still arguably perform just as well, if not better, compared to the original models we tested. For example, the Bayesian models controlling for linguistic proximity are the only models that are able to eliminate false positives when there is weak autocorrelation on the predictor variable (the red line in W&K's Figure 1).

We thank W&K for highlighting that we used the same matrices for generating and analysing data in our simulations. While we assumed this was obvious, in hindsight, we could have gone to greater lengths to emphasise this feature of our analysis to readers. However, we do not think "data leakage" is the right label for this. This label is often used in the context of machine learning and other predictive modelling, where information from a training dataset is unintentionally incorporated into a test dataset that the model is later used to predict[3]. There was no such unintentional "leakage" in our article. We used simulated data to evaluate the performance of different causal inference models, including versions of the actual data-generating model, against a known ground truth. This is a standard design feature of simulation-based model validation in the causal inference literature[4,5] and not "leakage".

The broader lesson here is that researchers need to deal with non-independence in their national-level datasets, and when we do, our decisions must be informed by our causal understanding of the processes at work, not the application of a one-size-fits-all recipe. W&K are right to note that the data-generating process is often uncertain to researchers, especially with observational national-level data. But the fact that we may not be able to precisely specify this process does not justify resorting to simpler methods that we know do not capture the process.

There are several ways that researchers can deal with their uncertainty about the true data-generating process. First, researchers can try including multiple matrices in the analysis, as we did in our simulations, with little or no cost in false positives or loss of power. Since the random effects models scale the relative importance of different matrices according to the amount of "signal" in the data, it often does not hurt to include multiple matrices in the analysis, provided that these are motivated by a plausible model of the data-generating process.

Second, researchers can integrate over their uncertainty in the data-generating process. For example, cultural phylogenetic analyses often average over uncertainty in the language ancestry that is thought to have created autocorrelation in the data, iterating the model over many samples from a posterior set of language phylogenies[6,7].

Third, researchers can use an approximation of the data-generating process. In the case of religious ancestry, for example, our simulation results still stand when we use *different* matrices to generate and analyse the data, so long as those matrices approximate the same underlying process (i.e.,

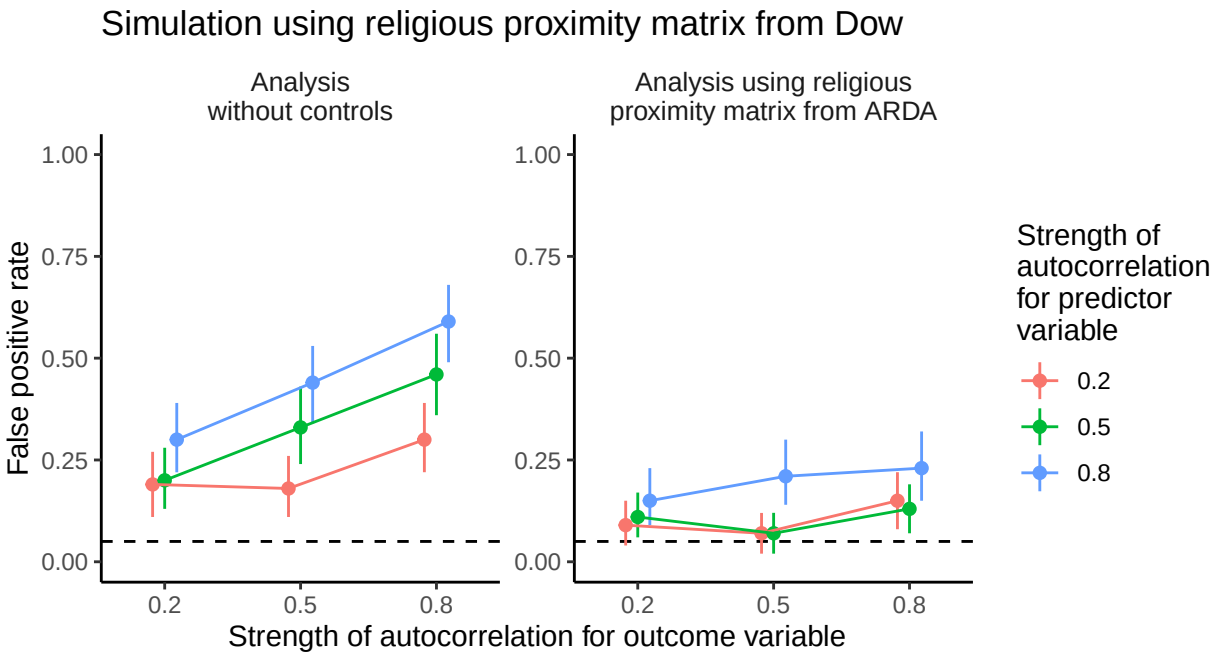

Figure 1: Results from additional simulations using an alternative religious proximity matrix to analyse the data. False positive rates were operationalised as the proportion of models that estimated a slope with a 95% credible interval excluding zero. Points represent raw proportions of false positives out of 100 models, ranges represent 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). ARDA = Association of Religion Data Archives.

not a different process like linguistic ancestry). To show this, we repeat the simulations from W&K but use an alternative religious proximity matrix to analyse the simulated datasets. This alternative religious proximity matrix was created using data from the Association of Religion Data Archives[8,9] (see Supplementary Information). The matrix has higher resolution than W&K's matrix and is more strongly correlated with W&K's matrix (r = 0.81) than language or geography. Our simulations with this alternative matrix show that valid approximations of the true data-generating process can effectively reduce false positive rates, even when the covariance matrices used to generate and analyse the data are not identical (Figure 1).

As a final point, while we agree that these Bayesian random effects models can take a long time to run when iterated across hundreds of simulated datasets, we note that, contrary to the claims from W&K, *individual* models in our simulation took no longer than a minute to sample on a standard laptop.

Data availability

Data on geographic, linguistic, and religious proximity can be found on GitHub: https://github.com/ScottClaessens/crossNationalSimulations

Code availability

The code necessary to reproduce our simulations and generate the manuscript can be found on GitHub: https://github.com/ScottClaessens/crossNationalSimulations

Competing interests

The authors declare no competing interests.

References

1. Claessens, S., Kyritsis, T. & Atkinson, Q. D. Cross-national analyses require additional controls to account for the non-independence of nations. *Nature Communications* **14**, 5776 (2023).

2. Wolfer, S. & Koplenig, A. Leakage explains the apparent superiority of Bayesian random effect models - a preregistered comment on Claessens, Kyritsis and Atkinson (2023). (2024) doi:10.31219/osf.io/ex267.

3. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* **6**, 1–21 (2012).

4. Pearl, J. *Causality: Models, Reasoning and Inference*. (Cambridge University Press, Cambridge, UK, 2000).

5. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. (CRC Press, 2020).

6. Sheehan, O. *et al.* Coevolution of religious and political authority in Austronesian societies. *Nature Human Behaviour* **7**, 38–45 (2023).

7. Watts, J., Sheehan, O., Atkinson, Q. D., Bulbulia, J. & Gray, R. D. Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature* **532**, 228–231 (2016).

8. Brown, D., Mataic, D. R., Bader, C. & Finke, R. The Association of Religion Data Archives: ARDA. (2018).

9. 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).

Supplementary Information

Data-generating model for simulations

Following the approach in our initial article (Claessens et al., 2023), we simulated data for 150 nations i with varying degrees of religious autocorrelation for outcome y and predictor x using the following generative model:

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

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

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

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

In this generative model, Σ_{Dow} is the national-level religious proximity matrix that Wolfer and Koplenig (2024) constructed using data from Dow and Karunaratna (2006). In addition, λ and ρ are autocorrelation parameters that represent the expected religious "signal" for outcome and predictor variables, respectively, and r is the true cross-national correlation between the variables after accounting for religious autocorrelation. We set λ and ρ to 0.2 (weak), 0.5 (moderate), or 0.8 (strong). For simplicity, we always set $r = 0$ in this simulation. For each parameter combination, we simulated 100 datasets, resulting in 900 datasets.

Analysis without controls

We first analysed the simulated data without any control variables. The statistical model is as follows:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta x_i \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(5)
\end{aligned}
$$

We fitted each model using the *brms* package (Bürkner, 2017) with four chains, 1000 warmup samples, and 1000 post-warmup samples. We calculated the false positive rate as the proportion of β slopes with 95% credible intervals excluding zero.

Analysis using alternative religious proximity matrix

For our next analysis, we used a religious proximity matrix of countries from Kyritsis et al. (2022). This matrix is based on national-level data on religious adherence from the Association of Religion Data Archives (ARDA) together with a family tree representing genealogical relationships between 28 religious lineages, informed by historical sources. Since there is evidence for horizontal transmission between some lineages, eight different cladograms were considered, representing alternative paths of inheritance between lineages. For further details on the construction of these religious family trees, see Kyritsis et al. (2022).

For each of the eight cladograms, a pairwise distance matrix was generated based on patristic distances between religious traditions. These distances were then converted to proximities and weighted by adherent percentages from ARDA using the following formula (Eff, 2008):

$$
R_{rk} = \sum_{r} \sum_{k} p_{ik} p_{jr} s_{ij}
$$

where R_{rk} is the religious connection between countries r and k, p_{ik} is the percentage of the population in country k adhering to religion i, p_{jr} is the percentage of the population in country r

adhering to religion j, and s_{ij} is the religious proximity measure between religions i and j. The resulting eight matrices were averaged to produce the final matrix, which we label Σ_{ARDA}.

We then included this Σ_{ARDA} matrix in the statistical model above, allowing nation random effects to covary according to alternative religious proximities. The statistical model is as follows:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + z_{\text{NATION}[i]}\sigma_\alpha\Sigma_{\text{ARDA}} + \beta x_i \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta &\sim \text{Normal}(0, 0.5) \\
z_j &\sim \text{Normal}(0, 1) \\
\sigma_\alpha &\sim \text{Exponential}(5) \\
\sigma &\sim \text{Exponential}(5)
\end{aligned}
$$

We fitted each model using the *brms* package (Bürkner, 2017) with four chains, 1000 warmup samples, and 1000 post-warmup samples. We calculated the false positive rate as the proportion of β slopes with 95% credible intervals excluding zero.

Description of code functionality

The R code in our GitHub repository uses the *targets* package (Landau, 2021) to create the analysis pipeline for this manuscript. The code loads the proximity matrices, calculates the correlations between these matrices, runs the simulations, plots the results of the simulations, and reproducibly generates the manuscript file.

Supplementary References

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

Claessens, S., Kyritsis, T. & Atkinson, Q. D. Cross-national analyses require additional controls to account for the non-independence of nations. *Nature Communications* **14**, 5776 (2023).

Dow, D. & Karunaratna, A. Developing a multidimensional instrument to measure psychic distance stimuli. *Journal of International Business Studies* **37**, 578-602 (2006).

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

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).

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).

Wolfer, S. & Koplenig, A. Leakage explains the apparent superiority of Bayesian random effect models - a preregistered comment on Claessens, Kyritsis and Atkinson (2023). (2024) doi:10.31219/osf.io/ex267.