Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning

Caspar J. Van Lissa1,2

1 Utrecht University faculty of Social and Behavioral Sciences, department of Methodology & Statistics

2 Open Science Community Utrecht

Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning

The Many-Analysts Religion Project illustrates how researcher degrees of freedom cause one research question to be analyzed in many different ways (Team, 2022). Each submission showcases a different perspective on “best practices.” My submission (Team #067, <https://osf.io/n4jcf/>) illustrates two practices I consider important: Firstly, how the Workflow for Open Reproducible Code in Science can be used to create a fully reproducible paper and an unambiguous preregistration (WORCS, Van Lissa et al., 2021). Secondly, how rigorous exploration can complement confirmatory (hypothesis-testing) research, even in a preregistered study.

WORCS is a conceptual workflow based on three principles: 1) writing papers as dynamic documents that combine prose and analysis code and can be reproduced with a single click, 2) using version control to track every change to the project since its inception, and 3) managing dependencies, which means documenting all software required to reproduce the project. These principles are automatically implemented by an Rstudio project template in the worcs R-package. The WORCS-project for my MARP analyses is available at <https://github.com/cjvanlissa/manyanalysts_religion>. Following conventions, a preregistration form was submitted to [the OSF](https://OSF.io). Additionally, the state of the project repository was tagged at time of preregistration (like a time capsule). This so-called “Preregistration As Code” is arguably more comprehensive and unambiguous than a written preregistration form (Peikert, Van Lissa, & Brandmaier, 2021). It contains the exact planned analysis code, complete with a simulated dataset that allowed me to verify that the code worked as expected. I was in the experimental condition, and thus did not receive data until after preregistration - but preregistering before accessing the data is good practice in general. No plan is perfect of course, so after receiving the real data, some adjustments were necessary. Reviewers and readers can see exactly what changes were made by comparing the finished project to the preregistered project using [Git diff](https://github.com/cjvanlissa/manyanalysts_religion/compare/1f2a6bd..e38391d). A changelog indicates why these changes were considered necessary. The finished project is reproducible: Third-party auditors can download the entire repository, and follow the reproduction procedure [described in a vignette](https://cjvanlissa.github.io/worcs/articles/reproduce.html). WORCS provides solutions for specific challenges to reproducibility; for example, if original data cannot be shared, WORCS generates a synthetic dataset that allows auditors to at least verify the correctness of analysis code. WORCS is particularly suited for preregistered studies because it allows researchers to preregister a reproducible draft manuscript with functioning code for planned analyses, evaluated on a simulated dataset. Two advantages over preregistration forms are that 1) it obviates the need to write two distinct documents in separate formats (preregistration versus manuscript); the preregistered manuscript can simply be updated once real data is collected or accessed, and 2) computer code is much less ambiguous than verbal descriptions of planned analyses, thus reducing researcher degrees of freedom.

The second practice I want to address is exploratory research. The open science revolution has prioritized confirmatory research by advocating preregistration and replication. Nevertheless, exploratory research is an integral part of the “empirical cycle”: Data-driven insights inspire new hypotheses and help amend existing theory. Not all exploration is rigorous, however. Many researchers in the social sciences are trained in regression-based methods, and use these for exploration by trying combinations of predictor- and outcome variables until an interesting “significant effect” shows up. At best, this is a labor-intensive exploration method and at worst, a recipe for false-positive results (Wicherts et al., 2016). It has recently been argued that machine learning methods can be used to complement theory-driven research (Van Lissa, 2021). Machine learning methods automate exploration by identifying patterns in data, and ensure robust results by incorporating checks and balances to curtail false-positive findings and maximize generalizability to new data (Hastie, Tibshirani, & Friedman, 2009). Three outcome indices of machine learning models relevant for exploration are: First, the model’s predictive performance in new data. This establishes an upper bound for how well all of the variables included in the study are able to explain the outcome. If the variance explained by the theoretical model is close to that explained by a machine learning model, then the theoretical model might be quite good. If the discrepancy is large, the theoretical model can be improved. Finally, if this upper bound is low, that might give cause to rethink the study design. The second index is the rank-ordered variable importance of each predictor, which refers to that predictor’s relative contribution to the accuracy of the model’s predictions. Highly ranked variables make major contributions to a model’s predictive accuracy, and are thus important to consider when planning future studies or amending theory. The third index consists of the marginal associations of each predictor (or combination of predictors) with the outcome. These marginal associations can reveal non-linearity and even putative interactions.

The choice of a specific machine learning algorithm can be guided by assumptions (e.g., regularized regression assumes associations to be linear), or by an empirical comparison of algorithms’ predictive performance. My MARP submission used a specific machine learning algorithm called random forests (Breiman, 2001), which is well-suited to theory-guided exploration (Brandmaier, Prindle, McArdle, & Lindenberger, 2016). Random forests draw many bootstrap samples from the original data, then estimate a regression tree model on each bootstrapped sample. Each tree splits the sample repeatedly, considering a random subset of predictors at each split point and picking the predictor and value on this predictor that maximizes the homogeneity of the post-split groups. Every effect is represented by splits: A non-linear effect is represented by subsequent splits on the same variable at different values; an interaction is represented by subsequent splits on different variables. To predict new data, predictions are averaged across all trees, thus averaging out prediction error. Random forests can handle many candidate predictors, intrinsically accommodate non-linear associations and interactions between predictors (thus relaxing the assumption of linearity), and can accommodate both individual- and country-level predictors while being robust to measurement variance and random effects across countries. In the latter two cases, the model would incur an interaction to account for different effects between countries. [My results](https://github.com/cjvanlissa/manyanalysts_religion/blob/master/manuscript.md) indicated that, firstly, the predictive explained variance was . This is relatively low; thus we must consider that some important predictors of wellbeing are omitted from the data, or that the wellbeing scale has high irreducible error - which seems less likely due to its high reliability, . Secondly, When examining variable importance (Figure 1), religiosity and perceived cultural norms do not emerge as the most important predictors of wellbeing. Instead, socio-economic status appears to be by far the most important predictor of wellbeing, followed by between-country differences. SES should thus be considered as a relevant covariant, or featured in theories about wellbeing. Finally, the marginal associations reveal that the association of religiosity with wellbeing is likely non-linear (Figure 2), and that the bivariate marginal associations of religiosity and cultural norms with wellbeing show no inkling of an interaction. How does this relate to the interaction term I found in the planned confirmatory analyses, , CI? One potential explanation is that, through the high correlation between religiosity and cultural norms (), the interaction term captures some of the nonlinear effect of religiosity (for a detailed explanation, see Belzak & Bauer, 2019). These exploratory insights help us contextualize our confirmatory findings, provide alternative explanations, and suggest testable hypotheses for future confirmatory research - such as examining the effect of SES, or the potential non-linear effect of religiosity. This is how rigorous exploration can complement confirmatory research.

# References

Belzak, W. C. M., & Bauer, D. J. (2019). Interaction effects may actually be nonlinear effects in disguise: A review of the problem and potential solutions. *Addictive Behaviors*, *94*, 99–108. <https://doi.org/10.1016/j.addbeh.2018.09.018>

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*(4), 566–582. <https://doi.org/10.1037/met0000090>

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second). New York: Springer.

Peikert, A., Van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible research in r: A tutorial on how to do the same thing more than once. *Psych*, *3*(4), 836–867. <https://doi.org/10.3390/psych3040053>

Team, T. M. (2022). A many-analysts approach to the relation between religiosity and well-being.

Van Lissa, C. J. (2021). Mapping Phenomena Relevant to Adolescent Emotion Regulation: A Text-Mining Systematic Review. *Adolescent Research Review*. <https://doi.org/10.1007/s40894-021-00160-7>

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A.-L., Peikert, A., Struiksma, M. E., & Vreede, B. M. I. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, *4*(1), 29–49. <https://doi.org/10.3233/DS-210031>

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1832. <https://doi.org/10.3389/fpsyg.2016.01832>