Global Sensitivity Analysis Unveils the Hidden Universe of Uncertainty in Multiverse Studies

Andrea Saltelli[1], Alessio Lachi[2(✉)], Arnald Puy[3], and Nate Breznau[4]

[1] University Pompeu Fabra, Barcelona, Spain
[2] Saint Camillus University, Rome, Italy
alessio.lachi@unicamillus.org
[3] University of Birmingham, Birmingham, UK
[4] German Institute for Adult Education, Bonn, Germany

Abstract. A wave of recent many-analysts studies struggle to explain the "hidden uncertainty" in their results. We propose global sensitivity analysis (GSA) as a method to better see and understand this uncertainty. By comprehensively exploring the space of all possible model options and pinpointing how much uncertainty in the results is due to single or to high-order model specifications, GSA will improve current many-analysts study methods. In particular, it allows for an explained-variance feasibility calculation to determine if it is worth running a costly many-analysts study in the first place. We demonstrate the effectiveness of GSA by replicating the many-analysts study by Breznau, Rinke and Wuttke et al., which hoped to shed light on how immigration impacts public preferences for social policy but left 95% of the variance in the results unexplained.

Keywords: Global sensitivity analysis · Modelling of the modelling process · Many-analysts studies · Garden of forking paths

1 Introduction

The process of statistically modelling and testing a hypothesis has been compared to a garden of forking paths [7], because researchers are faced with a host of methodological and analytical choices that collapses with each path taken at every fork. Many-analyst studies attempt to explore the cumulative effect that these choices have on the results by gathering different investigators or teams who then independently test a single hypothesis using the same dataset [2,11]. These experiments show radical variation in both numerical results and the conclusions of the analysts, raising serious concerns about the state of scientific evidence [4]. Furthermore, many works are unable to identify what drives the uncertainty in the results. A recent study by [3], for instance, left 95% of the variance in the results unexplained. This means that one of the key goals of many-analysts studies (pinpointing which research choices drive the results in specific directions and

E. di Bella et al. (Eds.): SIS 2025, ISSSAS, pp. 299–304, 2025.
https://doi.org/10.1007/978-3-031-96033-8_49

magnitudes) is not reached [15]. Part of the problem may be caused by overlooking the missing cells in the potential decision matrix; that is, that there might not be enough teams to cover all scientifically feasible model choices. To this end, multiverse analysis [5,16] can be applied post-hoc to try and fill in the missing cells [1]. This has two limitations: the first is that multiverse analysis looks for the impact of unique modelling decisions in isolation, and not the combined effect of pairs, triplets, or higher-orders of choices. The second is that, as the potential forking paths in the matrix grow exponentially with each model choice, it becomes unfeasible to run all possible models. We argue that concentrating on single modelling choices in isolation as is commonly done in multiverse analysis, may be a large part of why so much variance is hidden in these many-analysts studies. Many-analysts studies are especially applied in the social and behavioral sciences to analyze topics like racial prejudice or brain functioning during different activities [2,14]. Unlike physics, human behaviors and social interactions do not clearly follow fundamental laws. Looking for single modelling choices in isolation may fail to account for this complexity. The output uncertainty is likely due to high-order interaction effects: the combined effect of two or more specific methodological decisions in the modelling process. Therefore it may be that we face "a universe of uncertainty hiding in plain sight", as one author recently suggested [6]. Here we show how Global Sensitivity Analysis (GSA, [10,12]) can overcome these limitations and provide a valuable tool for those planning a many-analysts study. GSA efficiently explores the modelling choices by sampling from all possible combinations of decisions using quasi-random numbers, which fill out the space of the choices more efficiently and evenly than random numbers. Because of sampling, GSA can explore a modelling space that would be computationally unfeasible for complex data structures with many potential modelling choices. Furthermore, GSA decomposes the output variance into orders of effects, quantifying the impact of each specific choice in the research pipeline and the interaction between different choices on the outcome variance. The application of GSA allows analysts to develop prior expectations about outcome variance and post hoc full decomposition of variance in many-analyst and multiverse studies. Finally, it can help save human and financial resources and allow researchers to walk all tracks of the garden of the forking paths at once with minimal computational power and without invoking dozens of other researchers.

2 Methods

We illustrate the potential of GSA for many-analyst studies by replicating via computer simulation the work of Breznau et al. [3], who mobilized 162 researchers in 73 teams to test the hypothesis that immigration reduces support for social policy. Researchers were given the dependent and independent variables, but afforded methodological freedom to chose their measurement strategies, statistical controls and the estimation techniques in the testing of the hypothesis. Overall, the teams reported 1,261 different results with 103 research decisions,

and the effects of immigration on social policies ranged from large negative to large positive. [3] concluded that only 5% of this uncertainty could be explained by the methodological choices made by the research teams. Based on [3], we compile the dependent variables, predictors and controls used by the research teams to test the hypothesis. We mobilize one dependent variable with 14 possible levels, nine control variables with two levels each (included or not), three control variables with four, three and two levels respectively, and four possible models (logistic or linear with or without clustered standard errors). We exclude predictors and controls when their inclusion in the simulations lead to singular matrices or uneven arrays. Overall, this approach creates a space formed by 14 uncertain parameters and 2.7M possible models, approximately 2,000 times larger than that effectively explored in [3]. The functional form of the models selected by the research teams in [3] can be summarized as $y = f(\boldsymbol{x}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where y is the dependent variable, $\mathbf{X}$ is the matrix of independent variables, $\boldsymbol{\beta}$ is the vector of model coefficients and $\boldsymbol{\varepsilon}$ is the statistical error. We use quasi-random numbers to construct a $2^{13} \times 14$ sample matrix $\mathbf{M}$, where $x_i^{(v)}$ denotes the element in the v-th row of the i-th column. In $\mathbf{M}$, X_i is linked to a specific uncertain parameter and treated as a trigger, a condition that randomly defines whether the parameter and which of its levels is included in the model. We then run $f(\boldsymbol{x})$ row-wise throughout the matrix and calculate the Average Marginal Effect (AME) for each model execution to estimate, on average, how much does policy support change given a change in immigration. Finally we calculate how much of the variance in the AME, our dependent variable y, is due to first, second, and higher-order interactions using Sobol' indices. The workflow of our analysis, jointly with the code and data, can be found in GitHub (https://github.com/arnaldpuy/universe_of_uncertainty).

3 Results

We replicate the trend in the distribution of AME produced by the 73 teams in Breznau et al. [3] in a GSA-based many-analyst setting. However, our results highlight a much larger variation in the outcome. While the AME range was bounded in $(-0.57, 1.30)$ in Breznau et al., an exploration of the uncertain decision space through a GSA indicates that AME values actually range between $(-2, 4)$. This means that, on average, a one-unit change in the independent (immigration) variable can lead from a 2-unit decrease to a 4-unit increase on the respective dependent variable measuring policy support (Fig. 1 a-b). The larger AME span covered by our analysis is because GSA efficiently explores the corners of the uncertainty space [13], and by design never walks the same path (to use the language of Gelman and Loken to describe unique model choice combinations) twice [9]. In other words, our GSA visits more paths than the analysts, who explore certain combinations of modelling choices more than others, or who simply are too few in number to recover a realist sample of paths.

Of the 14 uncertain parameters reflecting the model decision space created by the research teams in [3], six did not condition the final AME value. These

predictors were the age, the level of education, the gender (male or female), the public social welfare expenditure as a percentage of Gross Domestic Product (socx_oecd), whether to bootstrap the data and whether to run logistic/linear models with or without clustered errors (SE). The decision as to the inclusion of the other eight parameters did have an influence (Fig. 1c). Interestingly, the decision about which country or dependent variable (y) to select had a clear first order effect, contributing c. 23% and 7% of the AME uncertainty respectively. The remaining 70% was due to interactions between specific model decisions. Particularly significant in that regard is the high-order effect derived from selecting the dependent variable (y) and the policy variable (X1) (Fig. 1c). To better understand these interactions, we calculated second-order effects and found that there were four pairwise interactions which altogether contributed 18% of the AME uncertainty (Fig. 1 d). This means that 52% of the variance in the AME values was caused by the combination of three or even more specific methodological decisions. Such results explain why [3] were unable to pinpoint "key variables" responsible for shaping the relationship between immigration and social policy preference: the "hidden universe of uncertainty" only comes to light when the unique, interactive nature of certain model specifications are taken into account.

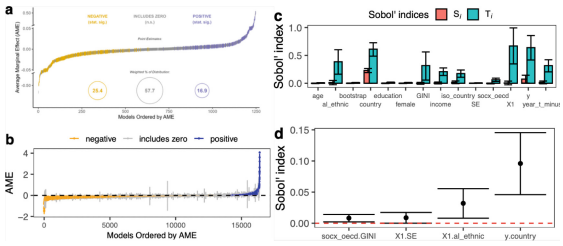

Fig. 1. A GSA-based many-analyst approach. a) Distribution of Average Marginal Effects (AME) in Breznau et al. [3]. b) Distribution of AME after a GSA. c) Sobol' first (S_i) and total-order (T_i) effects. S_i shows the proportion of variance contributed to the output by X_i, whereas T_i shows the proportion of variance contributed by X_i jointly with its interactions with all the other parameters. When $S_i < T_i$, X_i is involved in interactions. The x-axis shows the parameters accounted for in this analysis. d) Second-order (S_{ij}) effects. All error bars show the 95% confidence intervals after bootstrapping the Sobol' indices. The red, horizontal dashed line is the approximation error threshold and is at 0.05. We only show pairwise interactions whose contribution to the output variance is higher than the approximation error.

4 Discussion

Seen from the perspective of practitioners of GSA, the "hidden" universe of uncertainty is much more knowable than many analyst studies, or even multiverse analysis, suggest. [3] could only explain 5% of the real outcome variance,

and in a multiverse simulation to benchmark how much variance they "should" explain they only got 12%. This is because they searched almost entirely for first order variance only. Social and behavioral sciences particularly deal with some of the most complex phenomena we know of, human brains interacting with the environment, social groups, economic and political forces and geography; all in combination and constantly changing. Testing for the effect of independent model choices in these settings leads to little knowledge gain because the effect is driven by combinations of choices, often of a high-order. Referring back to the original [3] hypothesis: immigration, regardless of how it is measured, cannot explain variation in policy preferences because the variation is largely driven by the joint effect of at least three specific model choices. Furthermore, the direction of change can go in any direction depending on the combination. It is only at the second and higher orders of interactions that we acquire useful knowledge. Had [3] conducted a GSA prior to running their many-analyst study (which cost the participants roughly 3,138 h of work), they might have decided not to run it. They had no chance of finding key decisions that made certain outcomes more likely due to the strong weight of high-order effects. Alternatively, had the run a prior GSA, they might have re-framed their goals to find which unique combinations mattered most. Since testing for generalized singular effects in meta-science without running a GSA may be futile, a funding agency considering a many-analyst study should ask for a GSA as part of its design. Moreover, by conducting a GSA, they could have identified the factors that most strongly influenced the outcomes, leading to more robust patterns of association and, consequently, stronger conclusions. As the choice combinations that researchers might take are not known in advance, then exploring the space potentially spanned by these with a GSA will set the stage for a more efficient and possibly fruitful investigation. Similarly, a GSA is ideal for post hoc analysis of a model multiverse. Rather than prioritizing single model choices, it covers all possible model combinations. Doing this manually is not possible in many cases due to computing time. If we have learned anything from many-analyst studies, it is that the number of forking paths in the garden is far beyond what we previously imagined. The GSA approach solves this by comprehensively sampling from the universe of possible combinations of model specifications using a Monte Carlo quasi-random approach. Rather than 17 years to analyze a choice set of 30 model specifications, a GSA would instead allow someone with an average powered laptop to do this in just a few days. Today scientists point out that in conducting empirical research, the results of testing the same hypothesis likely vary depending on the study and model population sampled, study design and analysis choices [8]. We suggest the use of GSA beforehand to chart the potential space of variability, as to maximize efficiency, investment returns, policy impact and reliability in modern meta-science research. The [3] dataset is not only a Multiverse, but also a crowd-funded one, indicating that higher-order interactions may play a particularly critical role in crowd-funded multiverses. These structurally distinct settings demand careful consideration-underscoring

the need for GSA not just as a methodological tool, but as a prerequisite for drawing credible and generalizable conclusions.

References

1. Katrin, A., Josef, B.: Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts, One Data Set" Project". In: Socius, vol. 7, p. 23780231211024421, January 2021, issn: 2378- 0231
2. Rotem, B.-N., et al.: Variability in the analysis of a single neuroimaging dataset by many teams. Nature **582**(7810), 84–88 (2020)
3. Nate, B., et al.: Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Proc. Natl. Acad. Sci. **119**(44), e2203150119 (2022)
4. Camerer, C.F.: The apparent prevalence of outcome variation from hidden "dark methods" is a challenge for social science. Proc. Natl. Acad. Sci. **119**(52), e2216020119 (2022)
5. Cantone, G.G., Tomaselli, V.: Characterisation and calibration of multiversal methods. In: Advances in Data Analysis and Classification, pp. 1–33 (2024)
6. Per, E.: A universe of uncertainty hiding in plain sight. Proc. Natl. Acad. Sci. **120**(2), e2218530120 (2023)
7. Andrew, G., Eric, L.: The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time". In: Department of Statistics, Columbia University, vol. 348, no. 1–17, p. 3 (2013)
8. Felix, H., et al.: Heterogeneity in effect size estimates. Proc. Natl. Acad. Sci. **121**(32), e2403490121 (2024)
9. Puy, A., et al.: Models with higher effective dimensions tend to produce more uncertain estimates. Sci. Adv. **8**, eabn9450 (2022)
10. Puy, A., et al.: sensobol: an R package to compute variance-based sensitivity indices. J. Stat. Softw. **102**(5) (2022)
11. Matthew, J.S., et al.: Measuring the predictability of life outcomes with a scientific mass collaboration. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 117, pp. 8398–8403, 15 April 2020, . issn: 10916490
12. Andrea, S., et al.: Global sensitivity analysis: the primer. Wiley (2008)
13. Andrea, S., et al.: Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. Environ. Model. Softw. **114**, 29–39 (2019). issn: 1364-8152
14. Raphael, S., Eric, L.U.: Crowdsourced research: many hands make tight work. Nature **526**(7572), 189–191 (2015)
15. Uri, S., Simmons, J.P., Nelson, L.D.: Specification curve analysis. Nat. Hum. Behav. **4**(11), 1208–1214 (2020)
16. Sara, S., et al.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. **11**(5), 702–712 (2016)