



A Multiversal Model of Vibration of Effects of the Equitable and Sustainable Well-Being (BES) on Fertility

Giulio Giacomo Cantone^{1,2} · Venera Tomaselli^{1,2}

Accepted: 10 June 2023
© The Author(s) 2023

Abstract

This study checks the hypothesis that sustainable well-being is a determinant factor of fertility through the application of a multiversal method based on the assumptions of Vibration of Effects (VoE) model of multiversal sampling on the Italian Equitable and Sustainable Well-being (BES) dataset. For 48 social indicators of well-being, adopting the Province as statistical unit, the estimate of linear effect has been systematically sampled 1024 times. This multiversal sampling allows to estimate the component of the hypothetical variability of the estimate of the effect size that depends on the theoretical choices of the analysis. This component is commonly underestimated, hidden, or ignored in regular multivariate analysis or classical bootstrapping. Results do not allow to claim that any of the 48 social indicators is a robust linear predictor of fertility. Indicators associated with high statistical significance are also at risk of misleading the causal relationship. Implications for policy-making and the future direction of the research on BES are discussed in the conclusive remarks.

Keywords Multiversal models · Janus effect · Robustness · Panel data · BES · Fertility

1 Introduction

The present study is an application of a “Multiverse”-based model on the Italian Equitable and Sustainable Well-being indicators (a.k.a “Benessere Equo e Sostenibile”, BES). The proposed exploratory analysis regards the hypothetical claim that variation in future fertility can be explained by indicators of well-being.

Giulio Giacomo Cantone and Venera Tomaselli have contributed equally to this work.

✉ Venera Tomaselli
venera.tomaselli@unict.it

Giulio Giacomo Cantone
giulio.cantone@phd.unict.it

¹ Science Policy Research Unit, University of Sussex, Jubilee Building G08, Falmer, Brighton BN1 9SL, England, UK

² Department of Economy and Business, University of Catania, Corso Italia, 55, 95129 Catania, Italy

The concept of the Vibration of Effect, originally introduced by Patel et al. (2015), has been applied to biomedical databases (Tierney et al., 2021; Palpacuer et al., 2019). The underlying idea is to test a hypothesis of dependency of an outcome variable on various explanatory factors, all gathered from a single, large source. Thus, the main purpose of VoE is to test the evidential strength of a causal claim of a variable within the dataset in relation to the outcome variable. This is achieved by constructing a synthetic sample of estimates that constitute, at least theoretically, a multidimensional density distribution. Differently from methods based on *bootstrap*, this distribution has no element of randomness. It is instead fully determined by a process of differentiation on the specifications of, similarly to methodologies for sensitivity analysis proposed by Leamer (1983, 1985), Raftery (1995); Sala-I-Martin (1997), and other computationally intensive methods of inference.

The ultimate goal of performing a VoE analysis on a dataset is to identify whether or not regressors are evidently relevant variables in an explanatory model of the outcome. While other techniques for reducing the number of variables in a multivariate model, such as penalised regressions (LASSO, Ridge, etc), exist, the VoE methodology allows for the evaluation of informative statistics that directly interpret the uncertainty regarding the effectiveness of individual regressors, similar to quasi-experimental methods. Compared to *penalised regressions*, which tend to select a singular model, functional for decision-making (James et al., 2017; Oelker & Tutz, 2017; Young, 2019), VoE enables testing the hypothesis that there is no empirical basis for a dataset to explain an outcome. Furthermore, in this application, the VoE methodology has been expanded with some later introductions derived from the methods of Multiverse Analysis (Gelman & Carlin, 2014; Steegen et al., 2016; Muñoz & Young, 2018; Simonsohn et al., 2020). These introductions were necessary to better adapt the method to the panel structure, which is particularly recurrent in socio-demographic inferential data analysis, too.

Section 2 offers an account of this theoretical framework, and of the implementation in the adopted methodology, providing the application of the underlying principles and in the process of inference. The adoption of a multiversal model to sample distribution of estimates for a parameter allows not only to quantify the uncertainty regarding the estimate of that parameter, but also to assess how much of the hypothetical variation in that estimate is related to the sample distribution of the fit data (e.g. the *standard error*), and how much is related to the theoretical choices of the analyst in specifying the statistical model (e.g. the so-called *modelling error*), instead. In a regular analysis, the latter component of variation of the estimate is often not explicit and even neglected, so the uncertainty behind a scientific claim is systemically underestimated. The adoption of a multiversal model as an extended framework for VoE is aimed at correcting the statistical estimation of the hypothetical variance of the estimate, in order to better assess the overall robustness of statistically significant results. Peculiar attention is reserved for the quantification of the risk of claiming an effect when there is a consistent number of alternative statistically significant specifications of the model that would lead to claiming an effect of a different sign. This case, particularly relevant for therapeutical choices as for social policies, is formalised into a multiversal estimator of this risk (the “Type Sign Error”, or TSE), and this estimator is employed to corroborate the analysis of the suitability of the social indicators of BES as determinants or predictors of future fertility.

In Sect. 3 the hypothesis that sustainable well-being predicts future fertility is tested on a panel dataset of 48 BES indicators observed at Province level over a time span of 7 years. For each indicator, a multiverse of 1024 specifications of the regression is fit, based on an enriched version of the so-called Vibration of Effect (VoE) model of multiversal sampling. The results of the global analysis are not robust enough to assert that there is

evidence for linear effects of BES indicators on Italian fertility. Another relevant result is that the model-based component of the variance of the multiversal estimates is higher than the sampling-based component. It means that, at least for the proposed task of investigating the effectiveness of well-being on fertility, the statistical significance of a linear effect depends more on the pertinence of the theoretical background of the model than on the observed variability in BES data. The multiverse of estimates for some individual indicators is represented through volcano plots; these are discussed with the aim of highlighting common issues in modelling causal effects on fertility.

The employment of a multiversal method serves to illustrate the problem of legitimisation of ‘science-based’ (‘evidence-based’, etc.) decisions on policies, too. Since the scientific knowledge behind the adoption of a public policy can be characterised *a posteriori* as ideologically driven, excessive sensitivity of the estimates to the assumptions of the specification of the model specification should be more explicit, in order to allow a better-informed scientific evaluation of the evidence for the proposed policy. This problem is commented in Sect. 4, where some methodological developments based on ‘Multiverse’-based designs are discussed, too.

2 Theoretical Background

2.1 Vibration of Effect and other Multiversal Methods

Multiversal methods include all the methods based on the practice of fitting on the same data¹ multiple specifications of the same conceptual model of a scientific theory, and then reaching a scientific result that accounts for the variability observed across the sample of these specifications (Gelman & Loken, 2014; Patel et al., 2015; Steegen et al., 2016; Young & Holsteen, 2017; Simonsohn et al., 2020). All multiversal methods have in common collecting and comparing fit statistics from a systematically differentiated multiplicity of specifications of the same conceptual model.

The classic example of systemic differentiation of a single conceptual model into multiple specifications is the following: for any logistic regression that adopts a *logit* link function, exists also a logistic regression that adopts a *probit* link function (Gelman & Loken, 2014). The estimates (and the associated p-values) between these two specifications are expected to be close to each other, yet not identical. Other elements of the regression, such as the operational definition of the variables, the control structure and the functional form of the model (i.e. the “formula”), the estimators, etc. are other sources of differentiation of the specifications of the same conceptual regression model (Simonsohn et al., 2020; Del Giudice & Gangestad, 2021). The difference between estimates is quantifiable hence it has a variance. This variance is considered the model-based component of the hypothetical variance of estimates of the parameter.

Among the other fit statistics that can be collected *per* specification, there are the *p*-values, R^2 , etc. After differentiation of many specifications into a set \hat{J} , this set can be treated as a (non random²) ‘multiversal’ sample. Statistics can be derived from a ‘multiversal’ sample with a loose analogy to regular sampling. For example, assuming that the model

¹ If not from the same sample dataset, at least from samples that strictly share the same process of data generation.

² It does not allow to draw twice the same specification, so it is functionally equivalent to sampling without replacement.

variance in the estimate is finite, it makes sense to compute a multiversal average, multiversal standard deviation, etc. (Young & Holsteen, 2017). The commonly analysed statistics in multiversal methods are estimates of the coefficient of the ‘first regressor’ $\hat{\beta}_x$ and the *p*-values of their tests of statistical significance in simple or multiple regression.

Vibration of Effects analysis (VoE) has been developed as a confirmatory method to check the statistical robustness of etiological models in observational biomedical databases (Patel et al., 2015; Palpacuer et al., 2019). Tierney et al. (2021) extended the goals of the method to elicit hypotheses of causal structure in observational studies. In the latter case, VoE acts as a formalisation of the customary research practice to check the coefficients of effect size on a singular outcome variable (a ‘*Y*’) of many potential linear predictors from a pre-collected database (the ‘main regressors’, *X*). As an exploratory method, it looks for empirical insights before formalising any explicit hypothesis on the elementary biomarkers.³ The procedures to collect a multiversal sample for VoE Analysis are the following:

1. A coherent database of indicators (or, biomarkers, etc.) is identified.
2. A model of regression is fully specified.
3. An outcome variable *Y* is identified. An exemplary *y* is a variable of mortality in the statistical units (e.g. patients, etc.). In Medicine, mortality is both a relevant issue and a reliable one: it is a naturally binary state with no ambiguity in measurement, e.g. compared to psychometric indicators. These are excellent characteristics of an ideal *Y*.
4. A set $\mathbf{Z} : \{Z_1, Z_2, \dots\}$ of control variables is identified within the variables of the dataset.
5. The other variables in the database that are not $Z \in \mathbf{Z}$ are identified as independent *X*.
6. Data are fit on each regression $Y \sim X$, adopting all the elements of the power set $2^{\mathbf{Z}}$ as control structures in the functional form (pre-specified in the regression model). Hence, each vector \mathbf{x} of the database is fit $2^{|\mathbf{Z}|}$ times⁴. This operation is the “vibration”, while the “effect” refers to the parameter β_x .
7. The vectors of all the estimates $\hat{\beta}_x$, and of their associated standard errors ($se(\hat{\beta}_x)$) and *p*-values, grouped *per X* ($2^{|\mathbf{Z}|}$ rows for each *X*), make the \hat{J} multiversal sample for the VoE analysis.

VoE assumes that if these seven steps are conducted ‘reasonably well’, then the sample statistics of the multiverse are representative of or at least informative about the latent uncertainty of the inferential model.⁵ In the application (Sect. 3), the VoE template is altered in 2 ways: (i) $\hat{\beta}_x$ is estimated through four general models and not only one, and (ii) *Y* is specified in 8 different fertility rates.

³ On the perils of conducting this practice without a formal multiversal model, see Rubin (2017)

⁴ For example, for $\mathbf{Z} = \{Z_1, Z_2\}$, a linear model would get specified $2^{|\mathbf{Z}|} = 4$ times:

- (a) $y = \beta_x x + \epsilon$
- (b) $y = \beta_x x + \beta_{z_1} z_1 + \epsilon$
- (c) $y = \beta_x x + \beta_{z_2} z_2 + \epsilon$
- (d) $y = \beta_x x + \beta_{z_1} z_1 + \beta_{z_2} z_2 + \epsilon$

whereas the assumptions on ϵ are pre-specified and unique ($\epsilon \in r.v. V$, and *V* is unique).

⁵ Other multiversal methods differ in the assumptions about what can be inferred from this variability and, as a consequence, they differ in the set of procedures allowed to differentiate the sampled specifications. For example, Young and Holsteen (2017); Muñoz and Young (2018) imply that researchers are allowed to assume that the best estimator for a true value of a parameter must be an average of the estimates of that parameter in the sampled ‘multiverse’ of specifications, while Slez (2019) suggests to include at least a weighting scheme to such estimator.

Indeed, in Young and Holsteen (2017) the sample variance of the standard estimates $s^2(\hat{\beta}_x)$ (estimator for “model variance”) is explicitly assumed as informative about the sensitivity of the inferred effectiveness of the relation between the independent X and the outcome Y : a lower model variance is indicative of a robust linear effect, mirroring the standard error of the estimate.

Behind this assumption is the idea that the interplay of total survey errors (Groves & Lyberg, 2010) and misspecification error can induce a rate of false positives higher than the significance level α of the test, helping mischievous (or just unaware) authors to propose as statistically validated theories that would be not, under a more rigorous set of assumption. This specific mispractice in scientific research is known as p -hacking (Head et al., 2015; Imbens, 2021).

In a VoE analysis is checked the so-called Type Sign Error (Gelman & Carlin, 2014), or “Janus effect” (Patel et al., 2015), too. This error concerns the case when the sign (i.e. positive vs. negative) of $\hat{\beta}_x$ is not unique across the sub-samples of the statistically significant estimates. A high Error Type Sign (TSE) indicates that the estimate is sensitive towards conditional adjustments of the model.

2.2 Sustainable Well-Being and Fertility

In the words of Alaimo et al. (2022), the Italian Equitable and Sustainable Well-being (Benessere Equo e Sostenibile, BES) is “probably the worldwide most advanced attempt to pursue the beyond GDP perspective effectively”. The purpose of this panel dataset is to develop a multidimensional approach to complement the indicators related to economic measures together with measures of inequality and sustainability with the aim to depict the complex system of indicators produced in a comprehensive way through the analysis of a very wide annual updated set of measures, organised into 12 conceptual “domains”, each corresponding to related dimensions along the path of well-being.⁶

Mostly collected by the Italian National Institute of Statistics, the coverage of indicators keeps growing over time, including data from social surveys and administrative records. These records offer some thematic and analytical insights both at the territorial level and with respect to the different socio-demographic characteristics of the population. They are provided at different levels of territorial administration: some indicators are also recorded at Provincial level (Chelli et al., 2016) although these may suffer from missingness of data.

This work is constantly updated to reflect all improvements in the production of statistical measures within the National Statistical System and the progressive extension of the “mapping” of the indicators proposed by United Nations Inter-agency and Expert Group on Sustainable Development Goals (UN-IAEG-SDGs).

In the present study, the indicators of Italian well-being are hypothesised to be determinant causes of fertility, that is the average number of children *per* woman in fertile age. This hypothesis is not novel in the literature on BES: De Rose et al. (2020) found significant results for both a positive impact on fertility of some composite indicators of sustainable well-being (e.g. “quality of work”, “quality of services”) and a negative

⁶ Since 2016, some selected indicators of BES are included in developing the Italian government’s annual Economic and Financial Document (DEF) to analyse the more recent trends of well-being and assess the impact of the policies. Furthermore, since 2018 ISTAT publishes the “SDGs Report. Statistical information for 2030 Agenda in Italy” about Sustainable Development Goals to explain the intrinsic complexity of sustainable development, employing some BES indicators to define the 17 goals of sustainable development for 2030 Agenda programme of United Nations.

impact of others dimensions ("level of employment", "level of education"). The authors admit that these results are truly substantial only for Northern territories and that in general estimates may differ across territorial clusters of the country. This territorial differentiation seems in line with similar findings on determinants of health (Ciccarelli et al., 2022).

Including literature on the impact of well-being on fertility in other countries, there are three reasons for adopting a multiversal approach to check the soundness of the theory that variation in fertility rate can be determined by sustainable well-being:

1. Studies on human fertility, and in particular on the estimation of the impact of social determinants in national fertility, are of particular relevance in Social Sciences. There is a conspicuous international *programme of research* devoted to understanding why fertility rates are low in Western countries. Historically, this *programme* has been related to variation in economic performance and economic inequalities (Billari & Kohler, 2004; Sobotka et al., 2011; Balbo et al., 2013; Barbieri et al., 2015), yet these economic models should always face concurrent, and possibly more complex explanations for the variation of fertility across time and nation. For example, cultural shifts or changes in family structure are well-researched, too (Tropf & Mandemakers, 2017; Rodrigo-Comino et al., 2021; Giuntella et al., 2022). For BES or similar projects willing to "go beyond the GDP perspective" the development of rich-but-robust models of fertility may be one of the most prestigious challenges. Italy, with its persistent, yet paradoxical (Kertzer et al., 2009), low fertility is a good candidate to approach the riddle of fertility in the interplay of economical, sociological, and cultural determinants.
2. By the admission of almost all authors who approached it, a paradigmatic theory like Demographic Transition (Galor, 2012) is maybe not sufficient to fully explain (and govern) the observed shifts in fertility and in the age pyramids of nations (Graham, 2021; Sigle, 2021). Given the abundance of studies, this theoretical *impasse* should not be attributed to a lack of effort but to the inherent complexity of human behaviour about reproductive choices. Another, related yet different, issue regards the operational definition of fertility itself. The Total Fertility Rate (TFR) accounts for the number of newborns divided by the number of women within a range of years for which women are considered fertile. This indicator lends itself to confounding those effects that are localised across specific age ranges. Let us assume, just for example, that an increase in female employment implies a decrease in fertility under 30, but also an increase above 40. This dynamic of 'postponement' (Billari & Borgoni, 2005; Ni Bhrolchain & Beaujouan, 2012) would be lost by any estimator of a linear effect adopting TFR as unique Y . There are many age-specific fertility rates to test, so to test them all requires an analytical framework that accounts for the risk of p -hacking or of TSE. This gap can be filled by multiversal methods.
3. Demographic models of fertility, as background for social policies, have the tendency to be objects of political controversy. For example, in Hungary, policies implemented for raising fertility have been accused to reflect a biased political agenda (Szekulesz, 2022). More in general, hypotheses on determinants of fertility can often be characterised *a posteriori* with a political bias. This is more an epiphenomenon of the complexity and the relevance of fertility: many variables may predict variation in fertility with only a spurious connection to it, so different political actors have preferences for some theories and some policies that are more or less related to religious, moral, or ideological beliefs. Under these conditions, there are personal incentives for p -hacking a significant effect in

order to show “to be right about it”. In Saraceno et al. (2021), Multiversal models have been successfully employed to channel a diversity of personal ‘priors’ into a scientific methodology, and parties converged towards a shared and univocal interpretation of results.

3 Application

3.1 Aims of the Application

Compared to other multiversal methods, VoE is convenient for checking the risks associated with modelling of complex dynamics in databases that collect an extensive number of variables but a relatively low number of observations and statistical units. These features make VoE a desirable VoE ‘template model’ for investigating large socio-economic observational datasets.

Building an enriched model of VoE on the BES database has two main goals. The first is to perform a global exploratory evaluation regarding the hypothesis that some indicators of sustainable well-being can explain or predict the variation in Italian fertility. The second goal is to propose the multiversal method as a general methodology for detecting variables that are:

1. robust predictors of future fertility,
2. definitely not related to future fertility,
3. ambiguous, because they could lead to false, ideologically driven claims. For example, a variable with a high rate of significant estimates across \hat{J} that is also affected by high TSE (Janus effect) is an ambiguous predictor.

The prior expectation is that this exploratory analysis should not identify a high number of significant variables, because most BES variables have theoretically no direct connection with human fertility. No prior expectation can be elicited on the frequency of Janus effect. An ideal and desirable result is to separate a small cluster of significant variables (with a unique sign for each), and a majority of never-significant variables.

3.2 Data Sources

Data on fertility has been imported by the Health For All database (HFA) edited by the Italian National Institute of Statistics (ISTAT).⁷ A fertility rate is defined as follows: given an interval of age (in years, e.g. from 15 to 19 years old), a fertility rate is the total number of newborns from mothers aged within the interval, divided by the number of women aged within that interval. In HFA are reported 8 intervals of mother’s age: 15 : 49 (Total Fertility Rate, TFR); 15 : 19, 20 : 24, 25 : 29, 30 : 34, 35 : 39, 40 : 44, 45 : 49. TFR can be assumed at the ‘main’ fertility rate, the other operative definitions are its alternative specifications.

From HFA are imported two other variables: *frgn_rate*, the percentage of registered inhabitants without a Citizenship of European Union; and *marrg_rate*, the rate of marriages *pro capita*, a measure indicative of the propensity of people to form new families. In

⁷ HFA can be downloaded following the procedures indicated on this webpage: <https://www.istat.it/en/archivio/14562>.

Italy, new families and a higher presence of households of immigrants are positively correlated with fertility in the population.

The statistical units of the BES are 100 Italian Provinces⁸. The panel regards recorded values in BES database from the year 2013 to the year 2020. In the years before 2013, Italy sustained a considerable period of economic and social crisis and initiated significant reforms with an impact on the job market, demographic dynamics, and lifestyles of its inhabitants. After 2020, job markets, demographic trends, and lifestyles have been severely impacted by the pandemic threat of COVID-19. This is also the same timespan considered by De Rose et al. (2020).

The criterion for a BES variable to be included in the analysis is the following: there are recorded values for at least two consecutive years between 2013 and 2020, for at least one Province. This criterion is functional to adopting advanced procedures of estimation (see Sect. 3.3). The indicator for the extension of “Urban Green” (extension of non-historical parks and gardens in urban areas) has been excluded too because the data are unreliable, showing excessive variance. After this selection, of the original 78 variables of the BES at the Province level, 51 BES variables are left, for a total of 53 variables adding up *frgn_rate* and *marrg_rate*. Of these, in Sect. 3.4, 5 variables have been selected as Z controls, so the analysis regards only 48 independent X. A list of these 48 independent variables is in Table 2.

3.3 Type of Regressions and Estimators

3.3.1 Lagged Fixed Effects (LFE)

A panel dataset is characterised by a vector of identifier values $\mathbf{i} : \{i_1, i_2, \dots\}$ of the statistical unit and by a vector of time units $\mathbf{t} : \{t_1, t_2, \dots\}$. A common estimator for ascertaining if the value y_i of variable Y depends on the value x_i of an independent variable X is the parameter for the β_x standardised linear coefficient of the lagged regression:

$$y_i = \alpha_i + \beta_x x_{i,t-1} + \beta_z \mathbf{z}_{i,t-1} + \epsilon_{i,t} \quad (1)$$

where: α_i is a parametric effect that varies across statistical units, but it is invariant (‘fixed’) across time; $\mathbf{z} : \mathbf{z}_1, \mathbf{z}_2, \dots$ is the set of covariates with a function as control variables, with β_z being the vector of their coefficients; and $\epsilon_{i,t}$ is the error term (Reed, 2015; Bellemare et al., 2017). α_i and $x_{i,t-1}$ are the salient features of the model in Eq. 1, which in this context takes the name of 1-Lagged Fixed Effect model (1LFE).

If a statistical test finds evidence to reject the *null hypothesis* for $\beta_x = 0$ in Eq. 1, the statistical significance of the estimate $\hat{\beta}_x$ needs an explanation. Indeed, the reason to adopt the lagged regression is connected to the goal to connect the predictive model with a causal interpretation of the estimates: $x_{i,t-1}$ happens before y_i , so it must be generally excluded the hypothesis that the estimate $\hat{\beta}_x$ can be influenced by the effect of y on x , i.e. the effect of the “reverse causality” (Leszczensky & Wolbring, 2022). Through lagging x , assuming the

⁸ The dataset can be downloaded from this webpage: <https://www.istat.it/en/archivio/275374>. A Province is a local administration commonly used as a territorial unit in statistical surveys on the whole Italian territory. These 100 are all the Provinces of Italy (including the Autonomous Provinces of Trento and Bolzano - Südtirol) but excluding the Provinces in Region Sardinia. The administrative subdivision of Provinces in Sardinia, indeed, has been reformed in 2016 and these provinces are considered unreliable as statistical units.

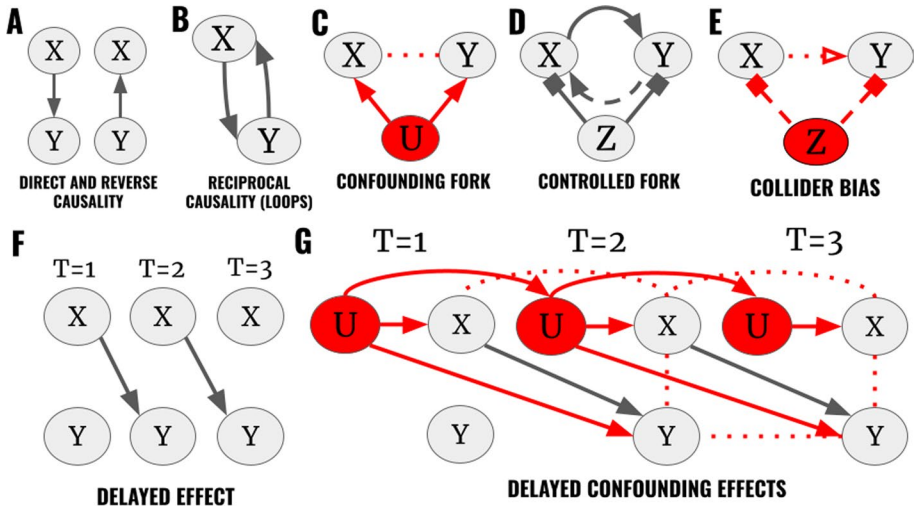


Fig. 1 Models of causal effects. A black straight arrow symbolises a correctly identified causal effect. A dashed line represents a hypothetical causal connection that is not specified. A dotted line with no arrows represents correlation. The red colour represents the induction of a bias. A red dotted line is a spurious correlation. In the model **D**, the reciprocal dashed arrows represent that the model does not identify univocally the direction of the effect. This identification happens instead in model **F**, through the inclusion of lagged time variable. In model **E**, controlling for **Z** induces a collider bias (Pearl, 2015). Collider bias induces an artificial correlation between **X** and **Y**, which is wrongly identified as a causal effect of **X** on **Y**, hence the empty arrow. In model **G** delayed confounding effects from autocorrelated **U** affect both **X** and **Y** over time. These effects induce spurious correlation between **X** and **Y** and spurious serial correlations in **X** and **Y**. (Color figure online)

absence of serial correlation over time (autocorrelation) in **X** and **Y**⁹, the model excludes hypotheses of reverse and reciprocal causality (Fig. 1, models **A** and **B**).

Since between conception and birth, there is a time of pregnancy, results from 1RLE should be compared with a 2-Lagged Fixed Effect model (2LFE):

$$y_t = \alpha_i + \beta_x x_{t-2} + \beta_z z_{t-2} + \epsilon_{i,t} \tag{2}$$

3.3.2 Lagged First Differences (LFD)

The case of covariation of **x** and **y** can also be explained through triadic causal structures called forks: a third value **u** determines both **x** and **y** (Pearl, 2015). In this case, even if **x** has a non-zero mediating effect on **y**, $\hat{\beta}_x$ would be biased (the correlation is ‘spurious’) unless it holds that if $u \in U$ then $u \in Z$, where **u** is a vector in the panel that is a representative sample of **U**.

This bias is commonly referred as ‘confounder bias’: the **U** in the fork, when not controlled, is the ‘confounder’ (Fig. 1, model **C**). Assuming **Z** does not contain **u**, then **U** is ‘hidden’: the necessary condition for unbiasedness of lagged models is if the effects of **U**

⁹ The postponement effect implies long-term autocorrelation in the time series of fertility. However, the considered sample (see Sect. 3.2) is too small to manifest it.

on X and of U on Y are simultaneous, which is the canonical case for forks. However, if U is autocorrelated and the effect of U on Y is delayed, compared to the effect of U on X , then lagged models are biased by the confounder U . This latter case is atypical, yet is structurally equivalent to the canonical case of spurious serial correlation over time in X (see Fig. 1, model G).

While the assumption that values in the vector \mathbf{y} are not autocorrelated in the short term is sound, it is not plausible to assume that it holds for each \mathbf{x} in the BES. To de-bias the estimate it can be adopted the estimator for $\hat{\beta}_x$ in the 1-Lagged First Difference model (1LFD):

$$\begin{aligned}\Delta(y) &= y_t - y_{t-1}; \\ \Delta(x) &= x_{t-1} - x_{t-2}; \\ \Delta(\mathbf{z}) &= \mathbf{z}_{t-1} - \mathbf{z}_{t-2}; \\ \Delta(\epsilon) &= \epsilon_t - \epsilon_{t-1}; \\ \Delta(y) &= \beta_x \Delta(x) + \beta_z \Delta(\mathbf{z}) + \Delta(\epsilon)\end{aligned}\quad (3)$$

that resilient to serial correlations in X and to delayed confounding effects of hidden U : even if u biases x , this bias affects equally x_{t-1} and x_{t-2} . Since these terms enter the regression with different signs, the bias on β_x from u is stochastically neutered, i.e. it holds

$$\mathbf{E}(\text{bias}(x_{t-1}) - \text{bias}(x_{t-2})) = 0 \quad (4)$$

For the same principle, since α_i are assumed time-invariant, they are neutered in the LFD, i.e. it holds $\Delta(\alpha_i) := 0$ for all statistical unit i (Allison, 2009; Levanon et al., 2009).

Again, by analogy to Eq. 2, the lag in Eq. 3 can be raised up to two years (2LFD):

$$\begin{aligned}\Delta(y) &= y_t - y_{t-1}; \\ \Delta(x) &= x_{t-2} - x_{t-3}; \\ \Delta(\mathbf{z}) &= \mathbf{z}_{t-2} - \mathbf{z}_{t-3}; \\ \Delta(\epsilon) &= \epsilon_t - \epsilon_{t-1}; \\ \Delta(y) &= \beta_x \Delta(x) + \beta_z \Delta(\mathbf{z}) + \Delta(\epsilon)\end{aligned}\quad (5)$$

In this regard, as demonstrated by Vaisey and Miles (2017), opting for an LFD model prevents confounding if the lags in a fork are correctly identified by the lags in the Deltas¹⁰.

3.4 Control Structure

Vibration of Effects analysis (VoE) requires identifying the set $\mathbf{Z} : Z_1, Z_2, \dots$ in order to expand it into its power set $2^{\mathbf{Z}}$. The elements of $2^{\mathbf{Z}}$, combined by addition, are the specifications of the control structure of the VoE model.

¹⁰ Still, all the causal structures in Fig. 1 represent optimal cases for adopting a regression type instead of another, but they are toy models that rarely resemble complete models of real demographic dynamics. For the data structures involved in databases like BES, complete models of a determination of a dependent variable resemble complex networks, as those shown in Onori and Jona Lasinio (2022); Alaïmo et al. (2022). Generally, as a form of ecological fallacy, panel models cannot capture significant heterogeneity at a level inferior to the primary statistical unit (Gnaldi et al., 2018). Luckily, works as Liu et al. (2012); Ding and Miratrix (2015) concluded that in complex causal structures, multiple confounders have an asymptotic tendency to mutually annihilate and not compound into more severe biases.

The causal relation between Y and Z is not the object of investigation of the VoE. Indeed, even if the Z s are technically treated as control variables, their role is not to adjust the estimate $\hat{\beta}_x$ towards a value that is *a priori* assumed in the hypothesis of the specification to be closer to the true value. Instead, across their power set, they alternate their presence and absence in the estimation function, allowing $\hat{\beta}_x$ to express variation across the specifications. Since the variation is to be attributed to the elements 2^Z , the choice of \mathbf{Z} is not trivial nor a-theoretical. In the scientific literature on multiversal methods, Z s are usually chosen because they are consistently associated with the outcome variable Y or because there is a strong principle for Z being a cause Y .

In the current application, five control variables are elicited. The first two are already mentioned in Sect. 3.2: `frgn_rate` and `marrg_rate`. The rate of foreigners is consistently positively correlated to an increase in fertility in Italy, to the point that the literature suggests that foreigners sustain a healthier age pyramid in Italy (Strozza et al., 2016; Benassi & Carella, 2022). The relationship between marriage and newborns is more ambiguous because is routed in cultural schemes that may change over time. However, the presence of marriage rates as controls is pervasive in the literature about fertility in Italy, hence the inclusion in \mathbf{Z} .

The third variable is the rate of graduated inhabitants of the Province (`grad_rate`), which is included in BES. It has been selected as a proxy value to measure the general level of education in the population. Education is considered a strong predictor of fertility, especially in the long run, both because it has a direct effect (e.g. through ‘postponement’) both because a variation in the rate of highly educated people spuriously reflects cultural changes (Impicciatore & Dalla Zuanna, 2017; Hahn et al., 2018; Impicciatore & Tomatis, 2020).

The fourth and fifth variables are included in BES, too. These are `income_pc` average (the income *pro capita*), and `employment` (the rate of people with an occupation). Income and employment are associated with lower fertility in the popular “Theory of Demographic Transition”: it states that as incomes and standards of living increase through modernisation, the fertility rate decreases; as people become more economically secure, they will concentrate investments in higher education of a small number of children. Additionally, income raises because a larger share of people, including a fair share of women, is employed. As a result marriages and pregnancies are delayed, and families have fewer children overall (Galor, 2012). Recent developments question whether this relationship should be modelled as a linear effect: the hypothesis itself is that there is also evidence that in advanced societies, richer families have more children because they can assure a sufficiently high standard of living for all of them (Turner & Robbins, 2022). So, as a variant of the $\mathbf{z} = \text{income_pc}$, it included also the quadratic term `income_pc2`; the same addition is included for `employment` (Cazzola et al., 2016). In addition, these two variables are collinear: in provinces with higher employment, a higher family income is expected. In a regular control structure, this fact would not be ideal, yet given that VoE is an analysis of robustness, stressing the model introducing collinearity in the predictor helps to track how it would inflate variation in the estimates.

In Table 1 are reported the standardised estimates for coefficient, and the associated standard error and p -value for the five \mathbf{z} in the multivariate regression of y on \mathbf{z} .

Table 1 helps to determine the control structure of the enriched VoE model: assuming $\alpha = .05$, `frgn_rate` reaches significance in three models, yet it is positively associated only in one case over three. `marrg_rate` presents a much more robust scheme: it is consistently associated when the lag is minimal, in accordance with the theory that variation in the rate of marriages determines natality in the next year because people marry with already the plan of a future pregnancy. `grad_rate` exhibit a double Janus effect: it is significant for 1LFE and

Table 1 Multivariate regression of the control structure on TFT

	1LFE	2LFE	1LFD	2LFD
<i>frgn_rate</i>	-0.272	-0.388	0.053	0.199
<i>st.error</i>	0.128	0.182	0.051	0.057
<i>p-value</i>	0.037	0.036	0.298	0.001
<i>marrg_rate</i>	0.153	0.026	0.175	0.008
<i>st.error</i>	0.068	0.076	0.045	0.049
<i>p-value</i>	0.028	0.732	< 0.001	0.877
<i>grad_rate</i>	-0.074	-0.018	-0.029	0.097
<i>st.error</i>	0.035	0.050	0.046	0.049
<i>p-value</i>	0.035	0.721	0.525	0.048
<i>income_pc</i>	-1.117	-1.880	0.020	-0.014
<i>st.error</i>	0.235	0.257	0.059	0.062
<i>p-value</i>	< 0.001	< 0.001	0.739	0.817
<i>income_pc</i> ²	-0.129	-0.289	0.038	-0.075
<i>st.error</i>	0.135	0.275	0.047	0.049
<i>p-value</i>	0.344	0.295	0.411	0.124
<i>employment</i>	-0.210	-0.083	0.008	-0.019
<i>st.error</i>	0.179	0.199	0.052	0.055
<i>p-value</i>	0.244	0.679	0.882	0.733
<i>employment</i> ²	-0.247	-0.181	0.144	0.031
<i>st.error</i>	0.092	0.142	0.046	0.049
<i>p-value</i>	0.009	0.204	0.002	0.526

Only the value in bold are significant

2LFD, but with opposite signs. Income shows a strong significance in Fixed Effects models but not in First Difference models, and the hypothesis that it could manifest a quadratic form (convex effect) is unsubstantiated. Employment, on the other hand, shows significant *p*-values of the quadratic term for 1-Lagged model. The identified control set **Z** will contain these five variables, with the only addition that *employment* will always show up with the addition of its quadratic term.

3.5 Evaluation Measures

3.5.1 Robustness Ratio

A summary statistic that is helpful to evaluate independent variables is the Young-Holsteen Robustness Ratio, or *RR* (Young & Holsteen, 2017):

$$RR(X) = |\hat{\beta}_x| \cdot \frac{1}{\sqrt{(\widehat{se}(\hat{\beta}_x))^2 + (s_j(\beta_x))^2}} \quad (6)$$

where $\bar{\beta}_x$ represents the average estimate of β_x , $\bar{se}(\hat{\beta}_x)$ is the average parametric standard error of β_x and $s_j(\beta_x)$ is the sample standard deviation of the vector of estimates of β_x , which is also called ‘model error’ (Young & Holsteen, 2017).

Young and Holsteen (2017) give a parametric interpretation of $RR(X)$: given an exploratory analysis including many X independent variables, $X | RR(X) < 2$ should not be claimed as causal determinants of Y . As aforementioned in Sect. 3.1 the ideal case is to observe a few variables with a high RR , and many others with a low RR , since this would help to identify patterns of causality excluding unrelated variables.

Since the composition is still linear, if $s_j(\beta_x) > \bar{se}(\hat{\beta}_x)$, then the estimates are very sensitive to model variance, which is usually hidden in regular analysis. This is a severe red flag that only a multiverse analysis can test the statistical significance of the relation between the variables.

3.5.2 Type Sign Error Estimator

To estimate TSE, the following procedure is proposed: let

$$n_{+\star}(x) = n(p_{\hat{\beta}_x} < \alpha | \hat{\beta} > \beta_{H_0}) \tag{7}$$

hence $n_{+\star}(x)$ is the number of specifications associated and to estimates that are simultaneously (i) statistically significant and (ii) higher than the null value β_{H_0} (e.g. $\beta_{H_0} = 0$), and let:

$$f_{+\star}(x) = \frac{n_{+\star}(x)}{n_j(x)} \tag{8}$$

where $f_{+\star}(x)$ is the relative frequency over the whole number of specifications of X ($n_j(x)$).

Then for asymptotically large $n_j(x)$, it is assumed:

$$Pr.(\beta_x > \beta_{H_0}) \sim f_{+\star}(x) \tag{9}$$

Equation 9 can be treated as an estimator of the probability to draw by chance a significant result higher than the null value among all the specifications sharing the same modelling process (i.e. the VoE model) that sampled the \hat{J} multiverse.

Likewise, the probability to draw by chance a significant result lower than the null is the same, but switching the operators in Eqs. 7 and 8:

$$\begin{aligned} n_{-\star}(x) &= n(p_{\hat{\beta}_x} < \alpha | \hat{\beta} < \beta_{H_0}); \\ f_{-\star}(x) &= \frac{n_{-\star}(x)}{n_j(x)}; \\ Pr.(\beta_x < \beta_{H_0}) &\sim f_{-\star}(x) \end{aligned} \tag{10}$$

To quantify the risk of TSE, consider the two relevant hypotheses:

1. $\beta_x > \beta_{H=0}$. When this hypothesis is true, the TSE has a probability $\sim f_{-\star}(x)$.
2. $\beta_x < \beta_{H=0}$. When this hypothesis is true, the TSE has a probability $\sim f_{+\star}(x)$.

The probabilities of TSE should be weighted for the probabilities of the hypothesis being true. In Eqs. 9 and 10 there is a method to estimate the probabilities for these hypotheses to be true. It follows:

$$\begin{aligned} Pr_{\cdot TSE}(X) & \\ & \sim f_{+\star}(x) \cdot f_{-\star}(x) + f_{-\star}(x) \cdot f_{+\star}(x) \\ & = 2(f_{+\star}(x) \cdot f_{-\star}(x)) \end{aligned} \quad (11)$$

From Eq. 11 follows that $\max(Pr_{\cdot TSE}(X)) = .5$. This is coherent with the assumption that if $f_{+\star}(x) > .5$, then $f_{-\star}(x) < .5$ by definition (see Eq. 8 and 10). Indeed, it holds that if $f_{+\star}(x) = .5$ and $f_{-\star}(x)$, then

$$2(f_{+\star}(x) \cdot f_{-\star}(x)) = .5 = \max(Pr_{\cdot TSE}(X)) \quad (12)$$

3.6 Multiversal Model

For each of the 48 X , it is generated a multiversal sample of 1, 024 specifications, summing up to a total of 49, 152 total estimates. This model, while keeping the general assumptions of VoE, includes three additions to how the specifications ‘vibrates’:

- The estimation technique of $\hat{\beta}_x$ vibrates across two types of panel regression: Fixed Effects and First Differences.
- Lag vibrates across 1 year of lag and 2 years of lag.
- Y vibrates across 8 specifications of the rate of fertility across different age ranges

As a consequence:

$$2\mathcal{E} \times 2\mathcal{L} \times 8\mathcal{Y} \times 2^5\mathcal{Z} = 1, 024\mathcal{J} \quad (13)$$

where \mathcal{E} are the estimation methods, \mathcal{L} are the lags, \mathcal{Y} are the specifications of the dependent variable, \mathcal{Z} are the control variables and \mathcal{J} is the number of specifications.

3.7 Results

The list of 48 independent variables with their evaluation statistics is presented in Table 2. Of these, 23 have a positive average of estimates in the multiverse and 25 have it negative. Assuming $\alpha = .05$, a median rate of significance $f_{\pm\star} = .2$ implies that, globally, the evidence to claim linear effects of the independent X on Italian fertility is very poor. For all X , the sample standard deviation in the multiverse of their estimates is always higher than their average parametric standard error.

It is possible to qualify the BES database as an insufficient empiric source to claim that, in general, sustainable well-being has a direct impact on Italian fertility: no one of the 48 variables reaches the goal of $RR > 2$ (Young & Holsteen, 2017). The highest RR is reached

Table 2 Multiverse statistics on the 48 main regressors from BES database

Domain	Label	$\tilde{\beta}_x$	$\tilde{se}(\beta_x)$	$s_j(\beta_x)$	RR	$f_{\pm*}$	TSE
Econ	Average wage	-0.05	0.18	0.64	0.07	0.38	0.07
Econ	Family debts	0.04	0.04	0.05	0.65	0.22	0.00
Econ	Low income pensions	0.12	0.16	0.39	0.28	0.46	0.09
Econ	Pro capita net wealth	0.04	0.21	0.34	0.09	0.13	0.00
Econ	Retirement incomes	-0.23	0.10	0.59	0.40	0.61	0.16
Educat	Children scholarisat	0.03	0.06	0.12	0.23	0.24	0.02
Educat	Children services	-0.03	0.10	0.22	0.12	0.23	0.02
Educat	H. School scholarisat	-0.05	0.08	0.14	0.33	0.10	0.00
Educat	Lifelong learning	-0.02	0.05	0.08	0.17	0.22	0.02
Educat	NEET Youth	0.05	0.07	0.11	0.38	0.23	0.02
Educat	Univ. scholarisat	-0.03	0.07	0.10	0.22	0.14	0.01
Environ	Elect. from renewables	0.02	0.09	0.14	0.10	0.15	0.01
Environ	Pollution (PM10)	0.00	0.05	0.06	0.06	0.06	0.00
Environ	Pollution (PM2.5)	-0.00	0.06	0.06	0.06	0.10	0.00
Environ	Urb. waste coll. (km)	-0.10	0.08	0.27	0.35	0.41	0.05
Health	Avoidable mortality	0.04	0.05	0.09	0.33	0.21	0.01
Health	Cancer mortality	0.03	0.04	0.08	0.28	0.33	0.04
Health	Dementia mortality	-0.02	0.05	0.14	0.14	0.39	0.07
Health	Infant mortality	0.01	0.04	0.04	0.25	0.06	0.00
Health	Life expectancy	-0.06	0.06	0.13	0.39	0.26	0.01
Health	Road mortality (Youth)	-0.01	0.04	0.03	0.29	0.02	0.00
Innov	Digital skills	0.01	0.09	0.10	0.11	0.06	0.00
Innov	Patent propensity	-0.00	0.06	0.06	0.06	0.10	0.00
Instit	Prison density	-0.01	0.04	0.07	0.13	0.18	0.02
Instit	Tax capacity (Municip.)	0.02	0.04	0.09	0.18	0.38	0.07
Instit	Tax capacity (Province)	-0.01	0.04	0.03	0.19	0.01	0.00
Instit	Women local admin	-0.02	0.05	0.17	0.09	0.49	0.12
Instit	Young local admin	0.03	0.05	0.08	0.33	0.19	0.01
Landsc	Farmhouses	-0.03	0.14	0.23	0.10	0.09	0.00
Landsc	Historic parks	0.27	0.17	1.08	0.25	0.37	0.06
Safety	Intentional homicides	0.02	0.04	0.03	0.38	0.12	0.00
Safety	Not violent crimes	0.20	0.08	0.31	0.61	0.51	0.05
Safety	Road mortality (rural)	-0.04	0.04	0.05	0.54	0.19	0.00
Safety	Violent crimes	0.05	0.07	0.17	0.29	0.47	0.10
Servic	Hosp. patients emigr	-0.04	0.16	0.45	0.08	0.26	0.03
Servic	Hospital beds	0.06	0.08	0.14	0.35	0.21	0.01
Servic	Irreg. elect. power	0.00	0.05	0.05	0.03	0.02	0.00
Servic	Public transportation	0.00	0.15	0.15	0.00	0.00	0.00
Servic	Special. hosp. beds	-0.02	0.10	0.16	0.10	0.13	0.01
Servic	Specialised medics	-0.20	0.14	0.36	0.51	0.46	0.01
Servic	Urb. waste coll. (pop.)	-0.08	0.06	0.15	0.51	0.36	0.02
Social	No profit organizations	-0.24	0.35	0.63	0.34	0.18	0.01
Work	Employment (Youth)	-0.02	0.07	0.10	0.17	0.11	0.00
Work	Occupational injuries	0.02	0.07	0.12	0.11	0.19	0.01

Table 2 (continued)

Domain	Label	$\tilde{\beta}_x$	$\tilde{se}(\beta_x)$	$s_j(\beta_x)$	RR	$f_{\pm*}$	TSE
Work	Pay days	-0.00	0.10	0.32	0.01	0.36	0.06
Work	Unemployment	0.07	0.15	0.32	0.20	0.15	0.01
Work	Unemploy. (Youth)	0.07	0.08	0.16	0.37	0.16	0.00

by the variable “Family debts” ($RR = .65$).¹¹ As a consequence, the model presented in De Rose et al. (2020) deserves further investigation.¹²

Nonetheless, this outcome was not unexpected, given the heterogeneity of social indicators in BES. The risk of making a claim of a causal effect in the wrong direction is relatively contained: the estimates for $Pr_{.TSE}$ rarely are above .05. However, the Kendall correlation between $f_{\pm*}$ and $Pr_{.TSE}$ is .77; given the structure of the estimator of $Pr_{.TSE}$ (see Sect. 3.5), this result is informative about the fact that most of the variance in the risk of observing a significant wrong sign depends by the probability to observe a significant result, which is another red flag.

3.7.1 Individual Analysis

These results lead to the individual investigation of those variables showing a severe “Janus effect” (high $Pr_{.TSE}$ with high $f_{\pm*}$).

There are only 6 variables with a rate of significant specification $> .45$. These are: “Retirement incomes” (.61, $\tilde{\beta}_x = -0.23$), “Low income pensions” (.46, $\tilde{\beta}_x = 0.12$), “Not violent crimes” (.51, $\tilde{\beta}_x = 0.20$), “Violent crimes” (.47, $\tilde{\beta}_x = 0.05$), “Women local administrators” (.49, $\tilde{\beta}_x = -0.02$), and “Specialised medics” (.46, $\tilde{\beta}_x = -0.20$).

Two hypothetical patterns are worthy of mention: (i) a rise in the allocation of pensions after retirement predicts (small) decreases in future fertility and (ii) higher crime is associated with future fertility. Even after adopting lagged models (see Sect. 3.3), both hypotheses do not configure necessarily a linear causal effect. Consider that “Retirement incomes”, the index of *pro capita* incomes after retirement (e.g. pensions), has the highest $Pr_{.TSE}$ of the whole BES at .16, and the correlated quota of “Low-income pensions” has a $Pr_{.TSE}$ on the higher end of the distribution (.9), too. This condition can be visualised in Fig. 2.

The difference in the magnitude of the estimated effect sizes between Lagged Fixed Effects (LFE) and Lagged First Difference (LFD) is striking. This pattern repeats for the models of the impact of the diffusion of not violent crimes (Fig. 3) and of the rate of women as local administrators (Fig. 4) on generational fertility. Other variables exhibit a similarly patterned ‘Janus effect’, even if less accentuated.

Observing the volcano plots, the principal cause of Janus effect seems the adoption of LFE models. This models, which is adopted in De Rose et al. (2020), inflates the estimate compared to LFD. Another common pattern in these three plots is that the Janus effect in LFE models is also connected to the age of the mothers (darker colours on one side, lighter on the other). For example, a causal interpretation of the effect of the rise of women administrators of Italian fertility based exclusively on LFE models could lead to think that

¹¹ This variable is the odd ratio of non-performing loans to families.

¹² De Rose et al. (2020) adopted Regions as statistical units and composite indices as latent variables, so the number of observations is reduced and part of the variance of the BES indicators is suppressed.

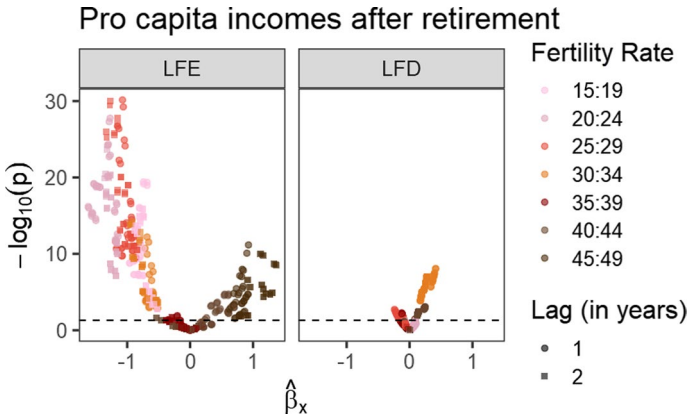


Fig. 2 Volcano plot of estimates of the effect sizes of *pro capita* incomes after retirement on Italian fertility. In different colours are reported 7 age-specific fertility rates, hence 896 specifications are represented in the model. Darker colours are associated with fertility in older women. The dashed line is the significance level $\alpha = .05$ (notice the y-axis is log-scaled)

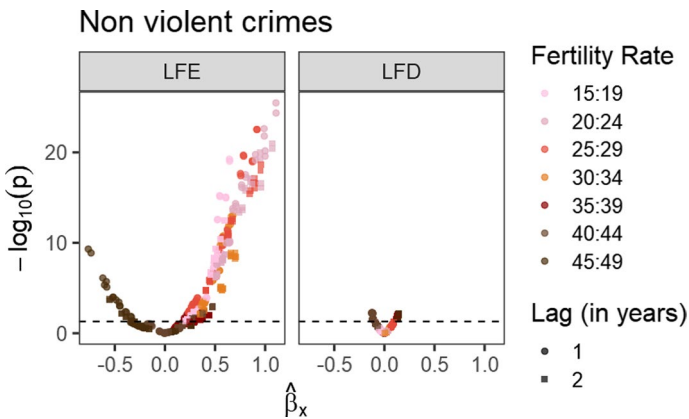


Fig. 3 Volcano plot of estimates of the effect sizes of diffusion of non-violent crimes on Italian fertility

the administration of women directly promotes fertility in older women but at the cost of disadvantaging it in young women ('postponement' effect on fertility). However, the LFD estimates tell totally another story: the presence of women as local administrators has a slightly, hard-to-catch, positive impact on the fertility of young and middle-aged women.

Which estimation method is best to listen to? Across all X , the first difference is more conservative in the magnitude of the effect. Compared to Fixed Effects, it suffers less lag-induced confounding effects (since it 'neuters' these) and more the misspecification of the lag structure itself. Overall, the observed results are coherent with the following explanation: LFE models cannot properly de-confound the estimate by latent structures. So, re-connecting this last assertion to the example of the effect of women administrators on fertility, one could say that the patterns observed in Fig. 4 are coherent with a phenomenology where women administrators have a very tiny positive direct effect on fertility (e.g. the promote pro-fertility policies) but the prerequisites for a rise in the rate of women

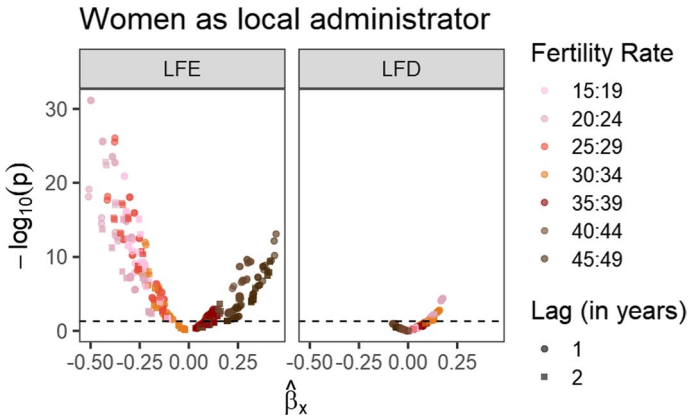


Fig. 4 Volcano plot of estimates of the effect sizes of the rate of women as local administrators on Italian fertility

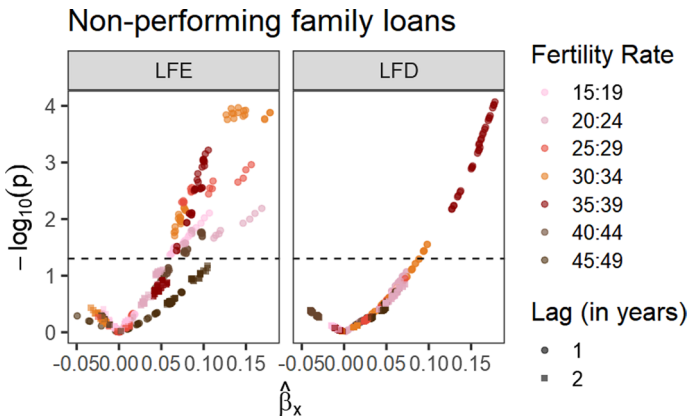


Fig. 5 Volcano plot of estimates of the effect sizes of non-performing family loans on Italian fertility

administrators (e.g. ‘women emancipation’) also induce a latent, lagged, ‘postponement’ effect in fertility - which are responsible for the technical negative average effect on fertility.

It is also possible to illustrate the analysis of a variable not affected by the Janus effect: the impact of non-performing family loans (“Family debts” in Table 2) on fertility (Fig. 5).

This is a peculiar case because the average of the effect size is very small and negligible ($\hat{\beta}_x = 0.04$), yet is very consistent across specifications and even across methods of estimation. A very small but robust effect size can be interpreted as the reflection of a persistent, but not very extensive, social dynamic. For example, poor families contract relevant debts during pregnancy. In this sense, debts would not be the cause of fertility, but a small contingency that manifests 1 year before birth. Indeed, a careful eye can catch that no specification lagged to 2 years is above the threshold of significance (the dashed line in Fig. 5). This is another perk of the enhanced VoE employed in the present analysis, as an exploratory method: it helps at isolating short-termed effects (when only 1-lagged models

are significant) from effects with an unclear lag structure (both 1-lagged and 2-lagged are significant¹³).

4 Conclusive Discussion and Future Directions

Results of Sect. 3 are coherent with the hypothesis that the effect of 'well-being' is too small to justify a policy based on well-being to raise fertility in Italy. The major argument to support this thesis is the moderate magnitude of the average multiversal estimates. No $|\bar{\beta}_x| > .3$ in Table 2 and the standard deviation of the multiversal estimates $s_j(\beta_x)$ is consistently higher than the average standard error of the estimate of the coefficient of linear effect, so much of the total variance of the multiversal estimates depends on the hidden model-dependent variance and not on sampling variance.

However, individual analyses on some of the most promising variables help to understand that "there is no evidence for linear effects of BES indicators on Italian fertility" does not imply that the effect of well-being on fertility is uniformly zero for all women. Two different arguments are provided to clarify why reasonable policies may be ineffective in supporting fertility in Italy.

The first argument regards the implication of 'linearity' of the effect. In Figs. 2, 3, and 4, it can be noticed that the darker cluster of the scatterplot of estimates, referring to older women, goes in one direction, while the lighter cluster of younger women, goes into another. Without this distinction, this could be referred to as a Janus effect, but through the differentiation across ages, this effect is totally explainable, in the sense that the same regressor may induce a negative effect in the reproductive decision in one cluster of women and positive effects in another. Here the role of the methodology is to visualise clearly this duality. This statistical behaviour would be coherent with the hypothesis of postponement of firstborn in young women. Given this specific configuration, it would also be possible to weight the multiverse estimates based on the actual contribution to the fertility of different age clusters of women. However, it should be noted that this solution would not be free from technical, operational, and even moral issues. From a technical standpoint, a correction criterion based on cluster-wise estimators (Abadie et al., 2023) would be necessary when computing the multiverse error. In practice, there is still no theoretical formulation of Young and Holsteen (2017) RR statistic that harmonizes standard error with cluster error. From an operational perspective, this solution would completely evade the concept of sustainability implicit in the BES. The idea of promoting social policies to support women's fertility should be aimed at a specific purpose, which is to rebalance the decision-making power of women who would like to have children but are not in a position to bear the costs. However, a criterion that favours certain clusters over others seems to be out of line with the foundational spirit of the BES and more in line with the aforementioned initiatives of the Hungarian government. In this sense, multiverse analysis can also be a starting point for an ethical debate on reproductive rights. A clear case for this argument is that, adopting the LFE model in Fig. 4, it can be argued that women should face more obstacles to reach the position of administration, in order to raise fertility in young women. This is unlikely an empowering solution for young women! In other words, the first argument against the claim that well-being is a determinant of fertility is that, even when local evidence is found

¹³ Given a larger timespan, one can include 3-lagged models, 4-lagged models, etc.

for it, this effect will be nested within a complexity of other variables, so the policy could easily backfire (or ‘vibrate back’) in the long run.

The second argument for avoiding to claim that policies of well-being can help to regularise fertility in Italy regards the extent to which collected evidence is truly sufficient for accurately refusing the null hypothesis of no effect. The study considered a panel of 100 provinces, across 6 years, and two different modelling approaches to account for the panel structure. The differences in the magnitude of the values between LFE and LFD estimates are striking. Aside from the interpretation of these differences (LFD may be seen as more conservative), there are more differences between designs of the inference than within them. This condition alone implies that the collected evidence ‘is not sufficient’ because it alone does not explain (‘is not the main reason’) for a specification to be statistically significant: the significance of a result depends more on the estimator than on the estimated.

In this regard, a limitation of the study is that, contrarily to De Rose et al. (2020), the modelling of estimation never involves multi-level assumptions. The Italian provinces are treated as an unstructured collection of statistical units, without the typical clusterisation in geographic macro-areas. Multi-level multiversal modelling has been attempted by Harder (2020), and it represents an ambitious future direction for multiversal modelling for demography.

From this consideration follows a second limitation, regarding the possibility of ecological fallacies across the multiple inferences (Gnaldi et al., 2018). For example, the model accounts for the presence of immigrants in the province, but then the outcome is the pooled fertility rate between natives and immigrants. In this regard, it can be said that the general interpretation of the results is more related to evaluating the uncertainty regarding a very general hypothesis rather than the possibility of simultaneously testing multiple hypotheses. If the research question specifically aims to quantify the uncertainty regarding the estimation of the effects of policies aimed at promoting the sustainability of well-being on fertility for Italian or foreign-born individuals, then it would be absolutely necessary to diversify the output variable.

On the extendibility of the methodology for purposes other than the evaluation of the robustness of a scientific claim, there is an ongoing debate about the usability of RR for model selection (Slez, 2019; Young, 2019). In this application, the VoE model produced 1024 specifications of the same conceptual model. How should one pick one? Aside from traditional methods based on Information Theory (Burnham & Anderson, 2002), Prof. Breznau wrote extensively on this problem from the perspective of the social scientist (Breznau, 2021). He favours the crowdsourced (or, meta-analytical) solution to multiversal sampling: the same dataset and the same research question are provided to multiple research teams, in line with the spirit of the “Many Labs” protocol (Stroebe, 2019). Researchers send back a sample of specifications and the whole multiverse is the union of the proposals. This crowdsourced method is technically superior: the same specification can be drawn in the sample more than one time (up to times equal to the number of teams), so the assumptions about the random structure of the multiverse are relaxed. BES has excellent features to be the empirical basis for reproducing this crowdsourced method; it could answer some of the most complex questions regarding the social life of people living in Italy. However, this path is not without adversities. Breznau’s experiment demonstrated that a relevant research question, namely: “Does immigration induce variation in the desirability of welfare?” still exhibits an irreducible Janus effect even in the crowd-sourced sample (Breznau et al., 2022). In this sense, the proposal of Slez (2019) to adopt a system of weightings sounds sensible.

To conclude, it has been aforementioned that the present study adapts a template originated in Biomedical Sciences for the new frontiers of Social Research. A relevant future direction of the research is envisioned reconnecting the Multiverse with the developments of the research on BES indicators. Part of the research on BES is focused explicitly on methods for aggregation of the elementary indicators into composite indicators of well-being for the 12 latent domains, with the Adjusted Mazziotta–Pareto method (Mazziotta & Pareto, 2016), the one referred in De Rose et al. (2020), being prevalent (Ciommi et al., 2017; Alaimo & Maggino, 2020), but not the only one (Davino et al., 2018; D’Urso & Vitale, 2021; Tomaselli et al., 2021; Giacalone et al., 2022; Monte & Schoier, 2022). The robustness of the inferences based on these composite indicators can be checked by adopting different alternatives and checking if these analytical choices, for the operational definition of the latent variables, induce significant differences in variance or effect size, and even if some methods of composition are consistently at risk of TSE.

Acknowledgements We thank Dr. Mal Lecoursonnais and Dr. Robert Hacken, who optimised the algorithm for the detection of J_{δ} . This study is part of the research line on vulnerability and risk management of the project GRIDAVI Risk Management, Decision Uncertainties and Social Vulnerabilities (Gestione del Rischio, Incertezze DecisionAli e Vulnerabilità sociali) by the University of Catania Research Incentive Plan 2020/2022 PIACERI.

Author Contributions All authors contributed to the study’s conception and design. Data collection was performed by VT and data analysis by GGC. The manuscript was written and approved by both authors.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abadie, A., Athey, S., Imbens, G. W., et al. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1), 1–35. <https://doi.org/10.1093/qje/qjac038>
- Alaimo, L. S., Arcagni, A., Fattore, M., et al. (2022). Measuring equitable and sustainable well-being in Italian regions: The non-aggregative approach. *Social Indicators Research*, 161(2), 711–733. <https://doi.org/10.1007/s11205-020-02388-7>
- Alaimo, L. S., & Maggino, F. (2020). Sustainable development goals indicators at territorial level: conceptual and methodological issues-The Italian perspective. *Social Indicators Research*, 147(2), 383–419. <https://doi.org/10.1007/s11205-019-02162-4>
- Allison, P. (2009). *Fixed effects regression models*. SAGE Publications. <https://doi.org/10.4135/9781412993869>
- Balbo, N., Billari, F. C., & Mills, M. (2013). Fertility in advanced societies: A review of research. *European Journal of Population / Revue européenne de Démographie*, 29(1), 1–38. <https://doi.org/10.1007/s10680-012-9277-y>

- Barbieri, P., Bozzon, R., Scherer, S., et al. (2015). The rise of a Latin model? Family and fertility consequences of employment instability in Italy and Spain. *European Societies*, 17(4), 423–446. <https://doi.org/10.1080/14616696.2015.1064147>
- Bellemare, M. F., Masaki, T., & Pepinsky, T. B. (2017). Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics*, 79(3), 949–963. <https://doi.org/10.1086/690946>
- Benassi, F., & Carella, M. (2022). Modelling geographical variations in fertility and population density of Italian and foreign populations at the local scale: a spatial Durbin approach for Italy (2002–2018). *Quality and Quantity*. <https://doi.org/10.1007/s11135-022-01446-1>
- Billari, F. C., & Borgoni, R. (2005). Assessing the use of sample selection models in the estimation of fertility postponement effects. *Statistical Methods and Applications*, 14(3), 389–402. <https://doi.org/10.1007/s10260-005-0122-x>
- Billari, F., & Kohler, H. P. (2004). Patterns of low and lowest-low fertility in Europe. *Population Studies*, 58(2), 161–176. <https://doi.org/10.1080/0032472042000213695>
- Breznau, N., Rinke, E.M., Wuttke, A., et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. In Proceedings of the National Academy of Sciences 119(44):e2203150,119. <https://doi.org/10.1073/pnas.2203150119>
- Breznau, N. (2021). I saw you in the crowd: Credibility, reproducibility, and meta-utility. *PS: Political Science and Politics*, 54(2), 309–313. <https://doi.org/10.1017/S1049096520000980>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.
- Cazzola, A., Pasquini, L., & Angeli, A. (2016). The relationship between unemployment and fertility in Italy: A time-series analysis. *Demographic Research*, 34(1), 1–38. <https://doi.org/10.4054/DemRes.2016.34.1>
- Chelli, F. M., Ciommi, M., Emili, A., et al. (2016). Assessing the equitable and sustainable well-being of the Italian Provinces. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24(Suppl. 1), 39–62. <https://doi.org/10.1142/S0218488516400031>
- Ciccarelli, A., Fabrizi, E., Romano, E., et al. (2022). Health, well-being and work history patterns: Insight on territorial differences. *Social Indicators Research*, 161(2), 629–647. <https://doi.org/10.1007/s11205-020-02393-w>
- Ciommi, M., Gentili, A., Ermini, B., et al. (2017). Have your cake and eat it too: The well-being of the Italians (1861–2011). *Social Indicators Research*, 134(2), 473–509. <https://doi.org/10.1007/s11205-016-1450-y>
- Davino, C., Dolce, P., Taralli, S., et al. (2018). A quantile composite-indicator approach for the measurement of equitable and sustainable well-being: A case study of the Italian provinces. *Social Indicators Research*, 136(3), 999–1029. <https://doi.org/10.1007/s11205-016-1453-8>
- De Rose, A.D., Racioppi, F., & Sebastiani, M.R. (2020). Fertility and well-being in the Italian regions. *Annali del Dipartimento di metodi e modelli per l'economia, il territorio e la finanza* pp 47–61. <https://doi.org/10.13133/2611-6634/510>
- Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1), 2515245920954925. <https://doi.org/10.1177/2515245920954925>
- Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, 3(1), 41–57. <https://doi.org/10.1515/jci-2013-0021>
- D'Urso, P., & Vitale, V. (2021). Modeling local BES indicators by Copula-based Bayesian networks. *Social Indicators Research*, 153(3), 823–847. <https://doi.org/10.1007/s11205-020-02530-5>
- Galor, O. (2012). The demographic transition: Causes and consequences. *Cliometrica*, 6(1), 1–28. <https://doi.org/10.1007/s11698-011-0062-7>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives in Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–466.
- Giacalone, M., Mattera, R., & Nissi, E. (2022). Well-being analysis of Italian provinces with spatial principal components. *Socio-Economic Planning Sciences*, 84(101), 377. <https://doi.org/10.1016/j.seps.2022.101377>
- Giuntella, O., Rotunno, L., & Stella, L. (2022). Globalization, fertility, and marital behavior in a lowest-low fertility setting. *Demography*, 59(6), 2135–2159. <https://doi.org/10.1215/00703370-10275366>
- Gnaldi, M., Tomaselli, V., & Forcina, A. (2018). Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review*, 86(1), 119–135. <https://doi.org/10.1111/insr.12244>

- Graham, E. (2021). Theory and explanation in demography: The case of low fertility in Europe. *Population Studies*, 75(sup1), 133–155. <https://doi.org/10.1080/00324728.2021.1971742>
- Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879. <https://doi.org/10.1093/poq/nfq065>
- Hahn, Y., Islam, A., Nuzhat, K., et al. (2018). Education, marriage, and fertility: Long-term evidence from a female stipend program in Bangladesh. *Economic Development and Cultural Change*, 66(2), 383–415. <https://doi.org/10.1086/694930>
- Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5), 1158–1177. <https://doi.org/10.1177/1745691620917678>
- Head, M. L., Holman, L., Lanfear, R., et al. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3), 1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, 35(3), 157–174. <https://doi.org/10.1257/jep.35.3.157>
- Impicciatore, R., & Dalla Zuanna, G. (2017). The impact of education on fertility in Italy. Changes across cohorts and south-north differences. *Quality and Quantity*, 51(5), 2293–2317. <https://doi.org/10.1007/s11135-016-0388-0>
- Impicciatore, R., & Tomatis, F. (2020). The nexus between education and fertility in six European countries. *Genus*, 76(1), 35. <https://doi.org/10.1186/s41118-020-00104-4>
- James, G., Witten, D., Hastie, T., et al. (2017). *An introduction to statistical learning: With applications in R* (2013th ed.). Springer Verlag.
- Kertzer, D. I., White, M. J., Bernardi, L., et al. (2009). Italy's path to very low fertility: The adequacy of economic and second demographic transition theories. *European Journal of Population / Revue européenne de Démographie*, 25(1), 89–115. <https://doi.org/10.1007/s10680-008-9159-5>
- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), 31–43.
- Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3), 308–313.
- Leszczensky, L., & Wolbring, T. (2022). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. *Sociological Methods and Research*, 51(2), 837–865. <https://doi.org/10.1177/0049124119882473>
- Levanon, A., England, P., & Allison, P. (2009). Occupational feminization and pay: Assessing causal dynamics using 1950–2000 U.S. census data. *Social Forces*, 88(2), 865–891. <https://doi.org/10.1353/sof.0.0264>
- Liu, W., Brookhart, M. A., Schneeweiss, S., et al. (2012). Implications of M bias in epidemiologic studies: A simulation study. *American Journal of Epidemiology*, 176(10), 938–948. <https://doi.org/10.1093/aje/kws165>
- Mazziotta, M., & Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicators Research*, 127(3), 983–1003. <https://doi.org/10.1007/s11205-015-0998-2>
- Monte, A., & Schoier, G. (2022). A multivariate statistical analysis of equitable and sustainable well-being over time. *Social Indicators Research*, 161(2), 735–750. <https://doi.org/10.1007/s11205-020-02392-x>
- Muñoz, J., & Young, C. (2018). We Ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1), 1–33. <https://doi.org/10.1177/0081175018777988>
- Ni Bhrolchain, M., & Beaujouan, E. (2012). Fertility postponement is largely due to rising educational enrolment. *Population Studies*, 66(3), 311–327. <https://doi.org/10.1080/00324728.2012.697569>
- Oelker, M. R., & Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, 11(1), 97–120. <https://doi.org/10.1007/s11634-015-0205-y>
- Onori, F., & Jona Lasinio, G. (2022). Modeling “Equitable and sustainable Well-being” (BES) using Bayesian networks: A case study of the Italian regions. *Social Indicators Research*, 161(2), 1003–1037. <https://doi.org/10.1007/s11205-020-02406-8>
- Palpacuer, C., Hammam, K., Duprez, R., et al. (2019). Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Medicine*, 17(1), 174. <https://doi.org/10.1186/s12916-019-1409-3>
- Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058. <https://doi.org/10.1016/j.jclinepi.2015.05.029>
- Pearl, J. (2015). Causes of effects and effects of causes. *Sociological Methods and Research*, 44(1), 149–164. <https://doi.org/10.1177/0049124114562614>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163. <https://doi.org/10.2307/271063>

- Reed, W. R. (2015). On the practice of lagging variables to avoid simultaneity. *Oxford Bulletin of Economics and Statistics*, 77(6), 897–905. <https://doi.org/10.1111/obes.12088>
- Rodrigo-Comino, J., Egidi, G., Sateriano, A., et al. (2021). Suburban fertility and metropolitan cycles: Insights from European cities. *Sustainability*, 13(4), 2181. <https://doi.org/10.3390/su13042181>
- Rubin, M. (2017). When does HARKing Hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, 21(4), 308–320. <https://doi.org/10.1037/gpr0000128>
- Sala-I-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2), 178–183.
- Saraceno, J., Hansen, E. R., & Treul, S. A. (2021). Reevaluating the substantive representation of Lesbian, Gay, and Bisexual Americans: A multiverse analysis. *The Journal of Politics*, 83(4), 1837–1843. <https://doi.org/10.1086/712139>
- Sigle, W. (2021). Demography’s theory and approach: (How) has the view from the margins changed? *Population Studies*, 75(sup1), 235–251. <https://doi.org/10.1080/00324728.2021.1984550>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Slez, A. (2019). The difference between instability and uncertainty: Comment on young and Holsteen (2017). *Sociological Methods and Research*, 48(2), 400–430. <https://doi.org/10.1177/0049124117729704>
- Sobotka, T., Skirbekk, V., & Philipov, D. (2011). Economic recession and fertility in the developed world. *Population and Development Review*, 37(2), 267–306. <https://doi.org/10.1111/j.1728-4457.2011.00411.x>
- Steege, S., Tuerlinckx, F., Gelman, A., et al. (2016). Increasing transparency through a multiverse analysis. *Perspectives in Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Stroebe, W. (2019). What can we learn from many labs replications? *Basic and Applied Social Psychology*, 41(2), 91–103. <https://doi.org/10.1080/01973533.2019.1577736>
- Strozza, S., Benassi, F., Ferrara, R., et al. (2016). Recent demographic trends in the major Italian urban agglomerations: The role of foreigners. *Spatial Demography*, 4(1), 39–70. <https://doi.org/10.1007/s40980-015-0012-2>
- Szekulesz, D. (2022). All planned babies must be born: Women’s experience of infertility and assisted reproductive technologies in Hungary. *Intersections East European Journal of Society and Politics*, 8(3), 30–47. <https://doi.org/10.17356/ieejsp.v8i3.874>
- Tierney, B. T., Anderson, E., Tan, Y., et al. (2021). Leveraging vibration of effects analysis for robust discovery in observational biomedical data science. *PLoS Biology*, 19(9), e3001398. <https://doi.org/10.1371/journal.pbio.3001398>
- Tomaselli, V., Fordellone, M., & Vichi, M. (2021). Building well-being composite indicator for micro-territorial areas through PLS-SEM and K-means approach. *Social Indicators Research*, 153(2), 407–429. <https://doi.org/10.1007/s11205-020-02454-0>
- Tropf, F. C., & Mandemakers, J. J. (2017). Is the association between education and fertility postponement causal? The role of family background factors. *Demography*, 54(1), 71–91. <https://doi.org/10.1007/s13524-016-0531-5>
- Turner, N., & Robbins, K. (2022). Association between county-level natality and income in the US, 2000–2020. *JAMA Pediatrics*. <https://doi.org/10.1001/jamapediatrics.2022.4814>
- Vaisey, S., & Miles, A. (2017). what you can-and can’t-do with three-wave panel data. *Sociological Methods and Research*, 46(1), 44–67. <https://doi.org/10.1177/0049124114547769>
- Young, C. (2019). The difference between causal analysis and predictive models: Response to “Comment on Young and Holsteen (2017)”. *Sociological Methods and Research*, 48(2), 431–447. <https://doi.org/10.1177/0049124118782542>
- Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods and Research*, 46(1), 3–40. <https://doi.org/10.1177/0049124115610347>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.