US FERTILITY THROUGH THE LENS OF GRAPHICAL CAUSAL MODELS[1]

Giambattista Salinari, Gianni Carboni, Gustavo De Santis, Federico Benassi

Abstract. This article explores the application of the Structural Causal Models (SCM) approach in the field of demography, discussing the PC algorithm to identify the causal chain, and the backdoor criterion, to identify the variables that need to be controlled for. Using a subset of the Panel Study of Income Dynamics (PSID) dataset, we applied the SCM approach to investigate the causal effects of women's age at first child on completed family size and household income, with the aim of simulating potential interventions designed at promoting an earlier onset of fertility. We found contrasting effects: inducing women to have their first child one year earlier could result in a 5% increase in their completed fertility, but it would also lead to a 4% reduction in their household income.

1. Introduction

The 1920s represented a turning point in the history of causal inference. During this decade, Ronald Fisher (1926) developed the idea of randomized experiments, today considered the most reliable method for causal inference, and Jerzy Neyman (1923), in his PhD dissertation, proposed a mathematical notation allowing for the rigorous treatment of causality. Around the same period, Sewell Wright (1921, 1934) proved that it was possible to represent causal dependencies among a set of variables using a graph where variables are represented as nodes and causal links as arrows (edges). Wright also showed that, in a linear system, it is possible to estimate the causal path coefficients following a set of simple rules.

About 50-60 years later, these seminal papers led to the development of two different approaches to causal inference. The first, based on the work of Fisher and Neyman, was developed in the field of statistics by Donald Rubin (2015) under the name of Potential Outcome Framework (POF). The second approach, developed by Judea Pearl (2009) in the field of computer science, expanded on Wright's original

[1] Giambattista Salinari was the progenitor of the core idea and provided methodological expertise for the study. Gianni Carboni was responsible for data curation and preparation, ensuring robustness and integrity in the analytical process. Gustavo De Santis and Federico Benassi contributed significantly through providing overarching critical feedback and refining the manuscript with insightful conceptual enhancements.

idea of representing causal dependencies using graphs, and introduced the concept of Structural Causal Models (SCM).

The two approaches are perfectly consistent, as a theorem developed by Pearl proves. However, they pursue partially different goals. Broadly speaking, causal inference faces two fundamental problems. First, it is necessary to establish whether a given causal question can be answered unambiguously using available data. For example, it can be proved that in a randomized experiment, the question about the existence of a causal effect of the treatment (e.g., aspirin) on an outcome (e.g., headache duration) can be answered univocally. This problem is known as *identification*. At that point, a second problem arises, known as *estimation*: measuring the causal effect of interest from a finite sample.

The POF and the SCM approaches differ, in our view, in the relative importance they attribute to these two problems: SCM focuses primarily on identification, whereas POF is mainly concerned with estimation. In short, the two approaches can be considered as complementary.

Another difference is in the fields of applications. POF has been widely applied in the social sciences, where SCM has been relatively neglected, despite its potentialities. For instance, it proves helpful in highlighting causal structures, and in the last 20-30 years, many algorithms have been developed in the SCM field to automatically identify causal connections between variables in a dataset, the so-called causal structure. Verma and Pearl, back in 1990, were the first to propose an algorithm with this purpose, called inductive causation (IC). Despite its limitations (the method is slow and inefficient), the algorithm has the merit of showing that it is possible to derive a causal structure from purely observational (non-experimental) data. Within limits, of course, as Verma and Pearl (1990) themselves pointed out. Depending on the characteristics of the underlying causal structure (to be discovered), the *direction* of some causal connections cannot be determined from observational data alone, even with an infinite set of observations. In these cases, the direction of causality can be determined only through randomized experiments, or based on a-priori, background knowledge, or by introducing assumptions. For example, if the IC algorithm cannot establish the direction of the causal connection between gender and income, we can conduct an experiment where we artificially increase the income of randomly selected individuals with the aim of observing whether this intervention leads to a change in their gender. Alternatively, we can rely on our background knowledge to conclude that a causal effect of income on gender is unlikely: the true causal connection goes the other way.

After IC, several other algorithms were proposed in the literature. The PC algorithm for instance (after its inventors Peter and Clark) is an improved – faster and more efficient – version of IC (Spirtes et al. 2000). Today, both the IC and PC algorithms are considered part of a broader class of learning algorithms known as constraint-based algorithms, which involve local testing of causal connections between variables.

An additional advantage of the SCM approach is the possibility of automating the search for the solution of the identification problem. Imagine that the causal structure of a dataset has been identified (e.g. via PC) and that we are interested in particular in the causal effect of X on Y (both variables). One of options offered by SCM is the now well-known *backdoor criterion* (Pearl 2009), which allows researchers to identify the set of causal variables, excluding "bad controls". For example, if X and Y share a common *cause* Z, the backdoor criterion will indicate the need to control for Z to correctly estimate the causal effect of X on Y. However, if Z is a shared *effect* of X and Y, the backdoor criterion will indicate that no control is necessary (keeping in mind that controlling for Z in this case could bias the results, introducing a spurious association between X and Y).

The backdoor criterion is sufficient, but not necessary and sufficient. When its conditions are satisfied, the identification of a causal effect is possible. However, there are contexts where, although the conditions for the backdoor criterion are not satisfied, the causal effect can still be identified using more advanced criteria, such as those provided by the *do-calculus*. The do-calculus has been shown to be complete, meaning that if a causal effect is identifiable, it can be demonstrated to be so using the do-calculus.

If a given causal effect can be identified, it can be estimated with several different methods: regression analysis, inverse probability weighting, matching etc.

To conclude, with the SCM approach, the computation of a causal effect can be broken down into three distinct phases:

1. Discovery of the causal structure;
2. Identification of the causal effect of interest;
3. Estimation.

In this paper we apply the SCM approach to answer two questions related to fertility in the US, and in particular to the causal effect of a woman's age at first birth: a) How does it affect completed family size? b) How does it affect lifetime income?

2. Causal discovery

For our analysis, we used the Panel Study of Income Dynamics (PSID) (Survey Research Center, Institute for Social Research, University oùf Michigan, 2022), a comprehensive US longitudinal dataset launched in the 1960s.

The PSID is a long-running, national panel survey of American families that collects data on economic, social, and health aspects. It is a valuable resource for researchers studying income dynamics, family structure, and other demographic variables. It has achieved a remarkably high wave-to-wave re-interview response rate of more than 90%. This rate indicates the percentage of participants who continue to participate in each successive wave of the study. The PSID started in 1968: as of 2019, 41 waves of data collection were carried out, highlighting a wide range of factors affecting the well-being of American families.

We focused on a subset of the PSID, known as the Childbirth and Adoption History dataset, with fertility information on more than 50,000 individuals of both genders. This data was subsequently integrated with an array of ancillary sub-datasets, encompassing a wide spectrum of demographic and socioeconomic variables, both at the household and individual levels. This integration facilitated a comprehensive reconstruction of individual life trajectory. For our study we retained only women whose fertility history was known for the ages between 20 and 45 years. This allows us to analyze sufficiently long fertility histories without limiting too much our sample size. Of course, something gets lost in the way, such as the possibility to investigate phenomena such as teen pregnancies.

This selection process left us with 2,531 women, for whom we retrieved also additional information from other sections of the PSID, creating the following variables (all referred to Ego, i.e. each woman in our sample):

- *Children*: Number of Ego’s biological children (ever had).
- *Year*: Ego’s year of birth.
- *Ethnicity*: Ego’s ethnic group.
- *ChildhoodIncome*: Ego’s past economic situation (when they were young).
- *EduExp*: Total education expenditure in Ego’s household.
- *Siblings*: Number of Ego’s siblings.
- *AgeCh1*: Ego’s age at first birth.
- *RelStatus*: The proportion of years (age 20 to 45) Ego lived with a partner.
- *EmplStatus*: The proportion of years (age 20 to 45) Ego had a payed job.
- *IncomePre*: Ego’s household income in the year preceding the first birth.
- *IncomePost*: Ego’s household income after the birth of the first child.

Our purpose is to detect a possible causal relationship between these variables. We employed the PC algorithm, which works under three main assumptions:

1. Reverse causality is not allowed (no loops):
2. There are no hidden confounders;
3. The joint probability of the variables in our dataset is stable (faithful).

The first condition states that it is not possible for a variable to be both cause and effect of any other at the same time. The key emphasis here is on simultaneity. Feedback processes are allowed, but it is crucial that causal effects, if any, take place in different periods. For example, income measured prior to the start of reproduction may influence the age at first child, and this can affect later income (for instance, parents with higher incomes may want to delay reproduction not to harm their careers). Feedback of this type can be represented without loops; in fact, it suffices to distinguish between two types of income, before and after the birth of the first child:

$$\text{IncomePre } \rightarrow \text{AgeCh1} \rightarrow \text{IncomePost}$$

According to the second assumption there must be no omitted variable that causally affects at least two variables in our dataset, where “omitted” means that this variable is not included in our dataset.

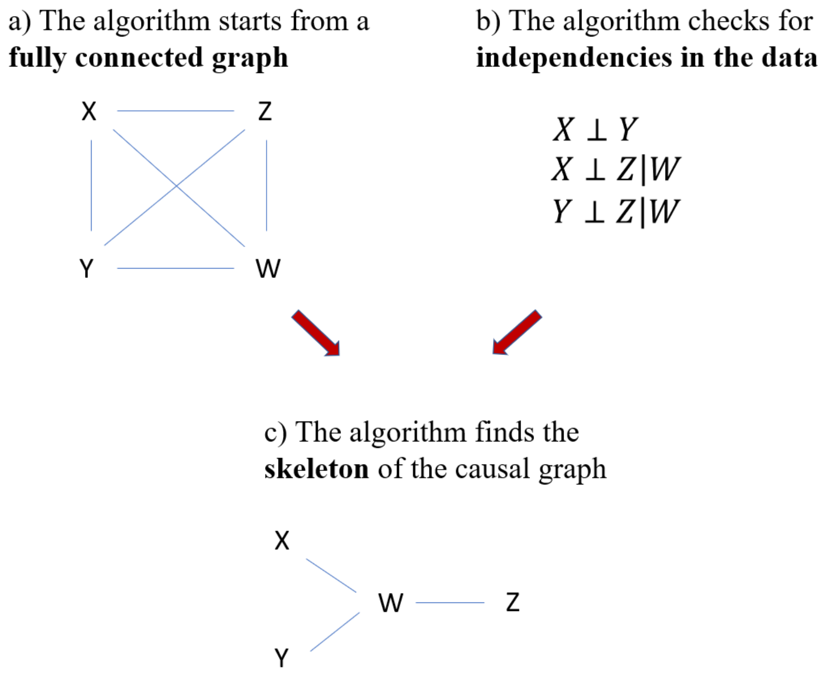
Note that this assumption is only provisional. Based on our current knowledge, all the relevant variables are included in the dataset to which we apply the PC algorithm to infer its causal structure (local Markov condition). Of course, if some relevant variable is missing, the resulting causal structure may be biased. However, this is not overly concerning because further analyses may later reveal that some extra variables should be considered. In other words, the causal structure is learnt through an incremental process, wherein the Markov condition is assumed to hold until proof of the contrary.

The third assumption states that some rare situations such as those occurring in a Simpson’s paradox do not affect our data (on this, see Spirtes et al. 2000).

Based on these assumptions, the PC algorithm systematically checks for independence in the relationships within our dataset. The logic behind it is that a direct causal connection between two variables exists only if they cannot be rendered independent by conditioning on all possible subsets of the remaining variables. In simpler terms, if two variables, X and Y, can be made independent of each other by conditioning on a subset **Z** of the remaining variables, a direct causal link between X and Y is excluded.

The PC algorithm initially assumes that all variables are causally linked (Figure 1a). It then proceeds to identify the independence relationships between the variables in the dataset, as depicted in Figure 1b. This is done with the help of standard statistical tools such as the chi-squared test of independence or the likelihood-ratio test. In the case of linear systems, linear regression and vanishing partial-correlation coefficients can also be used. In the example of Figure 1, we assume that three independence relationships emerge: 1) X and Y are marginally independent ($X \perp Y$); 2) X and Z are conditionally independent given W ($X \perp Z|W$); and 3) Y and Z are conditionally independent given W ($Y \perp Z|W$).

Figure 1 – *The PC algorithm finds the skeleton of the causal structure.*



a) The algorithm starts from a **fully connected graph**

b) The algorithm checks for **independencies in the data**

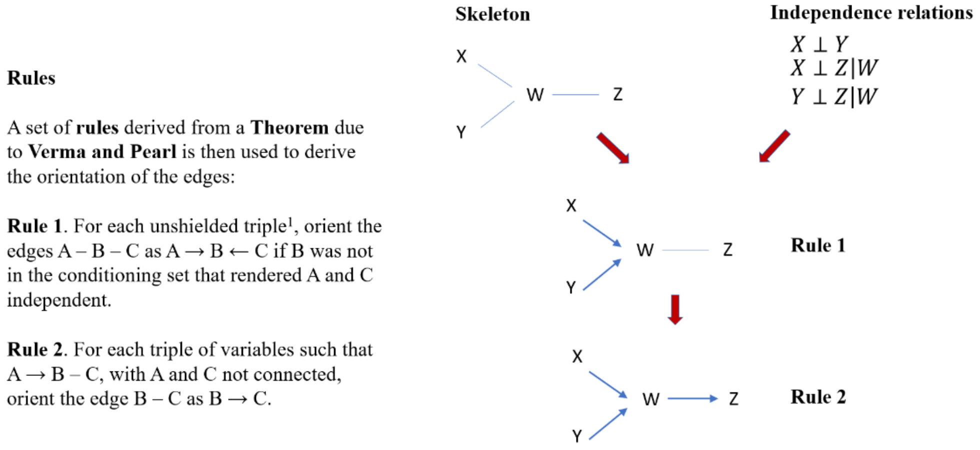
c) The algorithm finds the **skeleton** of the causal graph

Since two variables cannot be directly connected if they are found to be marginally or conditionally independent, in Figure 1c we simply remove the edges between X and Y, X and Z, and Y and Z from our original graph. This results in the "skeleton" of the causal structure, where the causal links are represented without their orientation (the arrowheads).

In its final step, the PC algorithm determines the orientation of the edges in the skeleton applying a set of rules derived from a theorem developed by Pearl and Verma (1990), as depicted in Figure 2.

Eventually, PC concludes that the only causal structure consistent with the observed independence relationships in our dataset is that shown at the bottom of Figure 2, where both X and Y cause W, and W causes Z. In this example, the PC algorithm successfully orients all the edges of the skeleton, but this is not always the case. There may be situations where the correct orientation of these causal connections cannot be derived from observational data alone.

Figure 2 –*The PC algorithm finds the orientation of the causal links*[2]



Rules

A set of **rules** derived from a **Theorem** due to **Verma and Pearl** is then used to derive the orientation of the edges:

Rule 1. For each unshielded triple[1], orient the edges A – B – C as A → B ← C if B was not in the conditioning set that rendered A and C independent.

Rule 2. For each triple of variables such that A → B – C, with A and C not connected, orient the edge B – C as B → C.

As said, the PC algorithm can be broken into two distinct phases: a) finding the skeleton, and b) orienting the edges. While the first phase generally produces stable and reliable results, the second phase may not (Spirtes et al. 2000), because the number of possible causal structures increases super-exponentially with the number of nodes. For example, with a set of 10 nodes (variables), there are approximately 4.2×10^{18} different possible causal structures, which increases the likelihood of picking up the wrong structure. This problem can be mitigated by narrowing down the search space, i.e. by incorporating background knowledge. In our case, for instance, the year of birth and the ethnic group of a woman are two “root” nodes in our graph, that is, variables that cannot be influenced by any other. Therefore, if there is a causal connection with, let us say, education, the direction must be from these

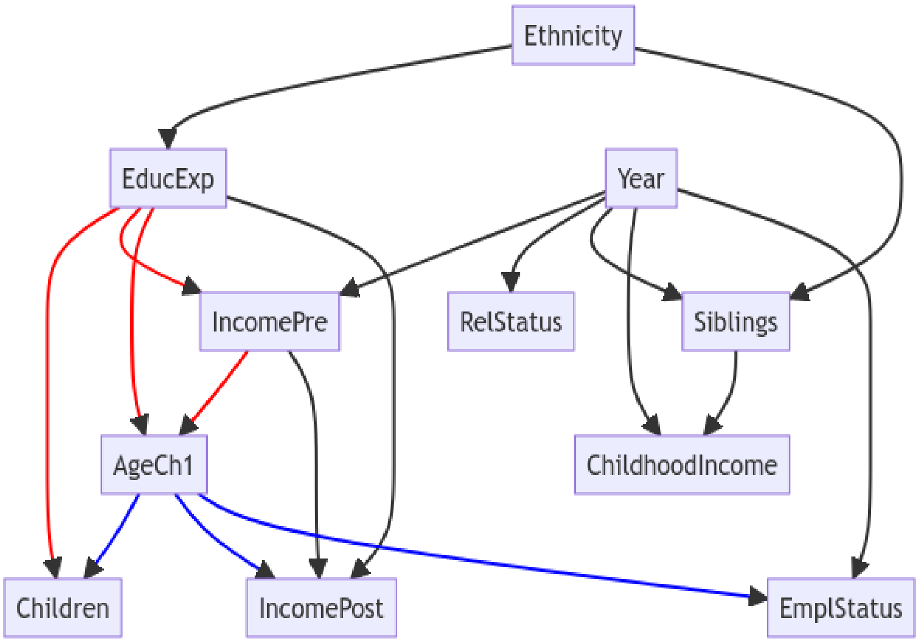
[2] Consider the triple A, B, C; this is referred to be unshielded if and only if A is adjacent to B (A–B), B is adjacent to C (B–C), and A is not adjacent to C (A C) Then A–B–C is an unshielded triple.

two variables towards education, and not vice versa. This type of information can be added to the algorithm in the form of a blacklist of prohibited causal connections.

This restricts the search space for possible causal structures, enhances the reliability of the results, and makes the Structural Causal Model (SCM) approach particularly promising in the field of social science. As actors in the social structure, we naturally possess a significant amount of background knowledge about the phenomena under study. In some cases, such as fertility for instance, the variables are naturally ordered by age, which guides the orientation of edges in the causal graph.

In our dataset, we were able to specify a set of 41 prohibited causal links, which guided the PC algorithm[3] in the discovery of the causal structure depicted in Figure 3.

Figure 3 – *The causal structure produced by PC using PSID data.*



[3] To recover this causal graph, we employed the bnlearn library of R. This library includes many different functions for recovering a causal structure. In the present case, we used the pc.stable() function (originally part of the pcalg library of R).

3. Causal effects identification and estimation

Below-replacement fertility is an issue in almost all advanced economies, and pro-natalist policies of some kind are frequently advocated. One among these involves lowering age at first birth. Would that be effective? To answer this question, we need to estimate the causal effect of the age at first birth on the total number of children.

To identify the causal effect of AgeCh1 (the treatment) on Children (the outcome), we begin by eliminating from the causal structure of Figure 3 all the edges emitted by the treatment (in blue). These edges are removed because they are the beginning of a series of paths through which the genuine causal effect flows. The attention then shifts to the remaining five paths that connect the two variables, as these are the paths that can potentially produce a spurious (non-causal) association between the treatment and the outcome. Two of these paths are of particular interest (in red in Figure 3):

$$\text{AgeCh1} \leftarrow \text{Edu.} \rightarrow \text{Children}$$
$$\text{AgeCh1} \leftarrow \text{IncomePre} \leftarrow \text{Edu.} \rightarrow \text{Children}$$

These two paths are classified as active (or open or unblocked) by the backdoor criterion, indicating that a bias is flowing through them, as the variable Edu. (EducExp) influences both age at first birth and the overall number of children. In other words, Edu. is a common cause, of the treatment and of the outcome, which qualifies it as a confounder. We must therefore control for Edu., netting out the spurious associations it may trigger. Surprisingly enough, in our case, this is the only control that is needed. This happens because in the three remaining paths connecting the treatment (AgeCh1) and the outcome (Children):

$$\text{AgeCh1} \leftarrow \text{IncomePre} \rightarrow \text{IncomePost} \leftarrow \text{Edu.} \rightarrow \text{Children}$$
$$\text{AgeCh1} \leftarrow \text{IncomePre} \leftarrow \text{Year} \rightarrow \text{ChildInc.} \leftarrow \text{Siblings} \leftarrow \text{Ethn.} \rightarrow \text{Edu.} \rightarrow \text{Children}$$
$$\text{AgeCh1} \leftarrow \text{IncomePre} \leftarrow \text{Year} \rightarrow \text{Siblings} \leftarrow \text{Ethn.} \rightarrow \text{Edu.} \rightarrow \text{Children}$$

we always find a collision node (or collider), that is a node influenced by two other variables (the collision nodes are respectively: IncomePost, ChildhoodIncome, and Siblings). Pearl showed that the presence of a collision node deactivates that path, through which, therefore, no spurious association flows.[4]

We conclude that controlling for EducExp is sufficient to block all the confounding paths connecting AgeCh1 and Children.

[4] The set of variables to be controlled for (the adjustment set) can be automatically detected. In the present case we used the dagitty library of R. dagitty is also available online at the following link: http://www.dagitty.net/.

We can now proceed to estimate the casual effect of interest through a Poisson model (because the outcome, Children, is a count variable):

$$log\big(E(Y|x,z)\big) = \beta_0 + \beta_1 x + \beta_2 z, \qquad (1)$$

where Y represents the number of children, X is the age at first child, and Z is the variable EducExp (Table 1). Our estimates indicate that a delay of one year in the age at first birth entails a 5% reduction in the overall number of children.

Table 1 – *Estimate of the causal effect of AgeCh1 on EducExp*

Coeff.	Estimate	SE	z	Pr(>|z|)
Intercept	1.806e+00	6.350e-02	28.442	< 2e-16
AgeCh1	**-4.100e-02**	2.911e-03	-14.084	< 2e-16
EducExp	1.765e-05	3.579e-06	4.933	8.09e-07

However, this analysis is somewhat incomplete. Let us also assess the possible causal effect of AgeCh1 (the treatment) on IncomePost (the outcome). The (backdoor) procedure is the same as before, and it tells us that it is necessary to control for both EducExp and IncomePreChild1 (because these two variables represent shared causes, or confounders, of both treatment and outcome).

This time the causal effect of interest can be estimated by mean of a linear model given the continuous nature of the response variable:

$$E(Y|x,z_1,z_2) = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2, \qquad (2)$$

where Y represents the log-income after the birth of the first child, X is the age at first birth, Z_1 the cost of education, and Z_2 income before the beginning of the reproductive phase. The estimates of this model are shown in Table 2. Age at first birth has a positive effect on post reproductive income, and lowering by one year the age at first birth causes a 4% reduction in the income measured after the birth of the first child.

Table 2– *Estimate of the causal effect of AgeCh1 on IncomePost*

Coeff.	Estimate	SE	z	Pr(>|z|)
Intercept	1.806e+00	6.350e-02	28.442	< 2e-16
AgeCh1	**-4.100e-02**	2.911e-03	-14.084	< 2e-16
EducExp	1.765e-05	3.579e-06	4.933	8.09e-07
IncomePre	1.787e-05	1.036e-06	17.25	<2e-16

In short, policies aimed at promoting an earlier onset of fertility, if successful, are likely to have two effects. On the one hand, they do increase fertility, as desired (+5% for every year of fertility anticipation); on the other, however, they tend to depress income, by 4%. Individuals may be unhappy with the latter consequence and refuse to comply with pro-natalist (anticipation) policies, unless additional interventions are foreseen to offset income reduction.

4. Conclusion

The main purpose of this paper is to unveil the potentials of the SCM approach in social studies and policy interventions. We presented an example where we simulated an intervention which induces an earlier onset of reproduction. While possible, this intervention may have unintended and undesired consequences, such as a decrease in post-reproductive income.

We presented a very simple case in this paper, but similar analyses can be conducted on a larger population, including males and employing a broader set of variables, among which personality traits, childhood economic conditions, job uncertainty, gender equity, personal aspirations, and several others, all of them, ideally, collected longitudinally, so as to preserve the dynamic nature of these relationships. In this case, the SCM approach offers a powerful identification tool, the "sequential backdoor criterion", which enables the identification of the effects of time-varying causes on time-varying outcomes. The graphical framework developed by the SCM approach can thus prove especially useful in a life course perspective, currently emphasized in most demographic and social science research.

It should be noted that the primary objective of this paper was to compute Average Causal Effects (ACE). In other words, our focus was on calculating the mean effect across all individual causal effects. Nevertheless, it is crucial to acknowledge that causal effects may vary within specific segments of the population. For example, the anticipation of fertility might have a more significant impact on completed family size for couples who are “family-oriented” rather than “work-oriented”. Characteristics such as being “family-oriented” or “work-oriented”, which modify the causal effect of the treatment, are referred to as “effect modifiers”. Identifying these effect modifiers is essential for policy design, as they facilitate the targeting of interventions to segments of the population that are more responsive to the treatment. Structural Causal Models (SCM) offer various tools to ascertain the existence of such effect modifiers (see Pearl 2009).

References

FISHER R.A. 1926. The Arrangement of Field Experiments, Journal of the Ministry of Agriculture of Great Britain, Vol. 33, pp. 503–513.

IMBENS G., RUBIN D. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge, Cambridge University Press.

NEYMAN J. 1923. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9, Roczniki Nauk Rolniczych Tom X [in Polish]; translated in Statistical Science, Vol. 5, pp. 465–480.

SPIRTES P., GLYMOUR C., SCHEINES R. 2000. Causation, Prediction and Search. Cambridge, The MIT Press.

PEARL J. 2009. Causality. Models, Reasoning and Inference. Cambridge, Cambridge University Press.

SURVEY RESEARCH CENTER, INSTITUTE FOR SOCIAL RESEARCH, UNIVERSITY OF MICHIGAN. 2022. Panel Study of Income Dynamics, public use dataset. Ann Arbor, MI: University of Michigan.

VERMA T., PEARL J. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence,* pp. 220–227.

WRIGHT S. 1921. Correlation and causation, *Journal of Agricultural Research*, Vol 20, pp. 557–585.

WRIGHT S. 1934. The Method of Path Coefficients, *Annals of Mathematical Statistics*, Vol 5, pp. 161–215.

Giambattista SALINARI, University of Sassari, gsalinari@uniss.it
Gianni CARBONI, University of Sassari, g.carboni21@phd.uniss.it
Gustavo DE SANTIS, University of Florence, gustavo.desantis@unifi.it
Federico BENASSI, University of Naples Federico II, federico.benassi@unina.it