Bayesian networks for causal analysis in socioecological systems

Cabañas, R., Maldonado, A. D., Morales, M., Aguilera, P. A., Salmerón, A.

Bayesian Networks for Causal Analysis in Socioecological Systems

Rafael Cabañas[a,c,*], Ana D. Maldonado[a,c], María Morales[a,c], Pedro A. Aguilera[b], Antonio Salmerón[a,c]

[a] *Department of Mathematics, University of Almería, Ctra. Sacramento s/n, La Cañada, Almería, 04120, Spain*
[b] *Department of Biology and Geology, University of Almería, Ctra. Sacramento s/n, La Cañada, Almería, 04120, Spain*
[c] *Center for the Development and Transfer of Matehmatical Research to Industry (CDTIME), University of Almería, Ctra. Sacramento s/n, La Cañada, Almería, 04120, Spain*

Abstract

Analyzing the influence of socioeconomy on land use is an important task, as socioeconomic factors can drive changes in land use that may ultimately affect human well-being. Recognizing the key factors that induce these changes may help policymakers design more effective strategies for addressing socioeconomic alterations on land-use planning, anticipate potential challenges, and mitigate negative impacts on both the environment and society. While probabilistic graphical models have been employed for this purpose in the past, this paper proposes the application of counterfactual reasoning to enhance the analysis by quantifying the degrees of necessity and sufficiency of various socioeconomic factors influencing land uses and population growth. Specifically, we present a case study using non-experimental data from southern Spain. For this, we propose the use of structural causal models, which are kind probabilistic models for causal analysis that simplify this kind of reasoning due to their graphical representation. They can be regarded as extensions of the so-called Bayesian networks, a well known modeling tool commonly used in environmental and ecological problems. This proposed ap-

*Corresponding author

Email addresses: rcabanas@ual.es (Rafael Cabañas), ana.d.maldonado@ual.es (Ana D. Maldonado), maria.morales@ual.es (María Morales), aguilera@ual.es (Pedro A. Aguilera), antonio.salmeron@ual.es (Antonio Salmerón)

proach is particularly effective for the identification of social and ecological variables that can be used in environmental monitoring and planning, offering key advantages including enhanced interpretability, and ease of adoption by environmental researchers. Our study reveals that immigration is both necessary and sufficient for population growth. In addition, built-up areas and herbaceous crops are favored by non-mountainous terrain and by high population density, whereas natural areas and mixed crops are supported by mountainous terrain and by low population density.

Keywords: Counterfactual analysis, Structural causal models, Structural equations, Bayesian networks, Socioecology, Land uses

---

1. Introduction

Causal (and counterfactual) reasoning (Pearl, 2009) allows to analyze cause-effect relationships, which is of fundamental importance for environmental and ecological practitioners and scientists. It can help in the development of effective strategies to mitigate or adapt to environmental problems, such as designing policies to reduce greenhouse gas emissions. This kind of reasoning can also help to evaluate the impact of different human activities on ecosystems.

Causal reasoning can typically be formally pursued through randomized experiments (a.k.a. randomized control trials), in which the variables of interest are intervened. In doing so, a study sample is randomly divided into one group that will receive the intervention with a given value and another that will be intervened with an alternative value. For example, in the problem of determining if a drug has a significant impact on the recovery from an illness, a group of patients will receive such drug whereas the other will receive a placebo. However, doing a randomized experiment in the field of environmental and ecological sciences might be expensive, unethical or directly impossible. For instance, if we aim to determine the influence of a population (from a specific species) on the structure and functioning of an ecosystem, it cannot be completely removed from it (or introduced in a new one where the population was not originally present). As a consequence, the environmental and ecological data available is usually observational, obtained from non-experimental studies. Using observational data with traditional statistical methods might lead to misleading conclusions when it comes to studying cause-effect relationships. This is because these methods typically rely on

analyzing correlations or associations between variables. This limitation is also present in many methods within the field of explainable AI (Ribeiro et al., 2016; Lundberg and Lee, 2017; Chen et al., 2022; Bach et al., 2015), which aim to identify the most influential inputs for a model's outcome. In contrast, counterfactual reasoning focuses on determining what would have happened if certain inputs had been different.

In line with this idea, this paper aims to analyze the relations of necessity and sufficiency of various socioeconomic factors influencing land uses and population growth, in the conceptual framework of a socioecological system (Anderies et al., 2004). Socioecological systems encompass the intricate interplay between human systems and natural ecosystems (Berkes et al., 2003; Preise et al., 2018). The socioecological system is a complex adaptive system, with some properties, such as: non-linear dynamics, critical thresholds, tipping points, regime shifts (Scheffer et al., 2012; Hughes et al., 2013; Mathias et al., 2020; Arnaiz-Schmitz et al., 2023), system memory, cross-scale linkages (Parrott and Quinn, 2016) and uncertainty (Biggs et al., 2015). All these properties are equally important for characterizing socioecological systems; however, in this work, we focus specifically on uncertainty. In the socioecological context, land-use changes (integrated in a landscape) are primarily driven by socioeconomic processes, influencing the ecological integrity of these landscapes, therefore changes in socioeconomic structures and processes induce an alteration of the landscapes (Schmitz et al., 2003).

In this paper, we provide a coherent overview of the fundamental concepts for applying causal and counterfactual reasoning to data analysis within the domain of environmental and ecological sciences. In particular, we consider the use of *structural causal models* (SCM) (Pearl, 2009; Bareinboim et al., 2022). SCMs are probabilistic graphical models (PGMs), i.e. they are probabilistic models in which the independence structure is encoded by a graph whose vertices are the variables in the model. SCMs are particularly designed for counterfactual reasoning and, like the rest of PGMs, they are suitable for environmental and ecological scientists and practitioners due to their graphical representation. Moreover, the recent method *expectation-maximization for causal computation* (EMCC) (Zaffalon et al., 2024, 2023) is proposed to be used for counterfactual reasoning. A key advantage of this method is its ease of implementation, primarily built upon the widely recognized *expectation-maximization* (EM) (Koller and Friedman, 2009, Ch.19) approach for parameter learning in PGMs. We put these concepts into practical use with an observational dataset, including information about socioeco-

nomic factors and land uses, in different areas of Andalusia (southern Spain). Unlike traditional analysis with other PGMs, the use of SCMs allows to analyze the relations of necessity and sufficiency between the variables in the aforementioned socioecological system.

This paper is structured as follows. Section 2 presents a motivational example; Section 3 reviews the relevant literature regarding the use of PGMs in the analysis of environmental data; Section 4 introduces the fundamentals of causal and counterfactual reasoning, with a specific focus on SCMs and BNs; Section 5 provides details about the case study considered for counterfactual analysis; the analysis of the results is presented in Section 6; Section 7 offers an overview of the main conclusions and policy recommendations drawn from this paper. Finally, the appendices provide a brief introduction to key concepts related to BNs and SCMs.

2. Motivational example

To illustrate the problem of traditional statistic methods, let us consider the observational data from a study that analyzes the relationship between socioeconomic factors and ecosystem services in cultural landscapes (Maldonado et al., 2018). In this illustrative example, only three Boolean variables are considered: *Mountain (M)* indicating whether the topography is mountainous (yes) or flat (no); *Immigration (I)* indicating if there are more people coming into the area (yes) than leaving it (no); and finally *Agricultural-land (A)* indicating if the land is mainly used for agricultural activities (yes) or other activities (no). This data is summarized in Table 1.

From the data presented in the table, it might be possible to build the (discrete) *Bayesian network* (BN) (Pearl, 1988) shown in Figure 1. A BN is formally defined as a tuple $\langle \mathbf{V}, \mathcal{G}, \mathcal{P}_{\mathbf{V}} \rangle$ where $\mathbf{V}$ is a set of variables from the problem being modeled, $\mathcal{G}$ is a directed acyclic graph (DAG), whose nodes are the variables in $\mathbf{V}$ and $\mathcal{P}_{\mathbf{V}}$ is a set containing a conditional probability distribution $P(V|\mathrm{Pa}_V)$ for each $V \in \mathbf{V}$ where Pa_V are the parents of V in $\mathcal{G}$. If all the variables are discrete, the conditional distributions are represented as tables, and we will refer to them as conditional probability tables (CPTs).

Suppose our variable of interest is *Agricultural-land (A)*. When studying the impact of immigration on the land use, one might consider to analyze the distribution $P(A|I) \propto \sum_M P(A|I,M) \cdot P(I|M) \cdot P(M)$. From the CPTs in Figure 1, it follows that $P(A = yes|I = yes) = 0.42$ and $P(A = yes|I = no) = 0.39$. As there is a positive correlation between both

Table 1: Data from an observational study involving three Boolean variables (Maldonado et al., 2018).

Mountain (M)	Immigration (I)	Agricultural-land (A)	Counts
yes	yes	yes	95
yes	yes	no	244
yes	no	yes	80
yes	no	no	183
no	yes	yes	121
no	yes	no	52
no	no	yes	47
no	no	no	8

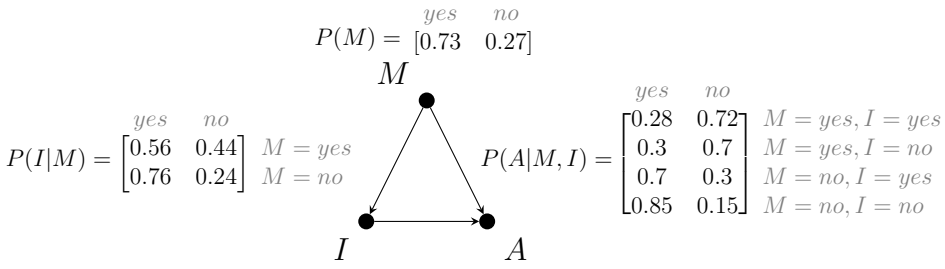

Figure 1: BN obtained from the observational data in Table 1.

variables, one could conclude that immigration has a positive effect on agriculture. However, analyzing separately the data in mountainous and flat areas, the correlation is the opposite as we have: $P(A = yes|M = yes, I = yes) = 0.28 < 0.3 = P(A = yes|M = yes, I = no)$ and $P(A = yes|M = no, I = yes) = 0.7 < 0.85 = P(A = yes|M = no, I = no)$. This is an instance of the so-called Simpson's Paradox (Pearl, 2009, Ch.6), which refers to a phenomenon whereby the association between a pair of variables reverses sign upon conditioning on a third variable (a *confounder*): flat areas are more likely to have immigration and also this topography is more suitable for an agricultural use of the land. A further explanation for this paradoxical situation, is that it is not the same seeing as doing. When calculating $P(A|I = yes)$, we are essentially asking about the probability of the agricultural land use *given that we see* that there is immigration. However, we might be interested in determining if the immigration is a necessary condition for the agricultural land use. In other words, if immigration were to cease in a given area, would it lead to a reduction in agricultural land use? Conversely, it is also valuable to understand if promoting immigration would be advantageous for agriculture, in which case we say that it is sufficient condition. These scenarios involve hypothetical situations that can be effectively addressed through counterfactual reasoning, which involves evaluating how the probability of outcomes would change if certain variables were set to specific values contrary to what actually occurred.

3. Related work

Non-counterfactual causal modeling techniques have been used in environmental and ecological sciences in a wide range of works. Some of these studies rely on causal diagrams. For instance, Byrnes and Dee (2025) address omitted variable bias in causal inference with observational data by using causal diagrams to identify confounders, combined with nested sampling and statistical designs. Arif and MacNeil (2022) emphasize the value of causal diagrams across four additional quasi-experimental approaches-propensity score analysis, BACI studies, regression discontinuity design, and instrumental variables. They demonstrate how causal diagrams clarify and unify variable selection in non-experimental settings, enhance transparency in communicating causal assumptions, and foster more critical and accurate discussions about the conclusions drawn from ecological research. These same authors, in their 2023 work, provide a summary of studies that have used the structural causal

model framework in ecology. They also use simulated ecological examples to examine how the backdoor and front-door criteria can produce accurate causal estimates between key variables, as well as how biases may occur when these criteria are not applied.

Structural equation models (SEMs) (Pearl, 2009) are another effective framework for causal analysis. In this sense, Paul and Anderson (2013) proposed ordination axes arising from multivariate macrobiotic species data in conjunction with SEM approach to analyze the causal effects in the 1978 Amoco Cadiz oil spill. In this work, the conditional independencies are considered by the authors as the only means to test causal structures with observational data (Paul and Anderson, 2013). SEM are also used in (Paul et al., 2016) to assess the risk of wastewater discharge on macro-invertebrate communities, focusing in the adaptation of the causal diagram to a statistical model which allows for computing the effect of an intervention retrospectively. Irvine et al. (2015) combine Bayesian path analysis and SEM to study the how stressors (such as anthropogenic drivers of road density, percent grazing or percent forest within a catchment) affect stream biological condition.

Other relevant works integrate integrate BNs with SEM for causal analysis. For instance, Hatami (2018a) integrate BNs with SEM to infer causal effects of wastewater on the macro-invertebrate community once the effect of natural variation is removed or to analyse the spatiotemporal variations of macrobenthic assemblage caused by leaking from a wastewater treatment plant. Similarly, Hatami (2018b) use an integrated BN-SEM framework to investigate spatiotemporal variations in macrobenthic assemblages resulting from leakage at a wastewater treatment plant.

Carriger et al. (2016) recommend the use of BNs for evidence-based policy in environmental management, on the grounds that these graphical models can look into the evidence for causality through improved measurements, minimizing biases in predicting or diagnosing causal relationships. In their review, the authors propose several guidance works on BN development for environmental problems and, as practical example, use BNs to study the impacts of biological and chemical stressors on a fish population.

Besides causal modeling, counterfactual thinking is essential in environmental policy to draw inferences about program effectiveness as well as to discriminate between program effect and biases (Ferraro, 2009). Siegel and Dee (2025) integrate the potential outcomes framework and structural causal models to create a complementary workflow that guides the process from

formulating a causal question to interpreting the results. Their work also includes resources for self-guided learning and a curated reading list. Andam et al. (2008) apply counterfactual thinking by using matching methods to improve the estimate of the impact of protected areas in Costa Rica on deforestation. In that work, the authors demonstrate that counterfactual thinking let control biases along observable features and check the sensitivity of the estimates to potential hidden biases. Also a statistical matching technique is used by McConnachie et al. (2016) to estimate cost-effectiveness of South Africa's Working for Water program on reducing invasive species.

In relation to our case study, various methodologies have been applied. De Aranzabal et al. (2008) used multiple regression to formalize the landscape–socioeconomic dependence at the local scale, enabling scenario analysis of socioeconomic change and its impact on the landscape. Ropero et al. (2014) developed a regression model based on hybrid BNs to study the landscape–socioeconomic relationship in watersheds, analyzing three landscape change tendencies under two socioeconomic scenarios. Punzo et al. (2022) employed econometric models to identify factors influencing land consumption at the municipal level, highlighting endogenous and exogenous interaction effects and the key role of demographic, socioeconomic, and institutional structures. Chen and Yao (2023) introduced the patch-generating land-use simulation (PLUS) model, combining Markov chains and Cellular Automata to predict land-use change, incorporating various drivers such as socioeconomic factors. Zhai et al. (2021) examined spatio-temporal patterns of land-use/cover change under urbanization in Wuhan, China, using continuous Landsat time series and support vector machines. They further applied the PLUS model to explore future landscape dynamics. Zhou et al. (2022) analyzed differences in arable land change due to urban and rural construction expansion, using geographically weighted regression to detect spatial and temporal patterns of land fragmentation. Wu et al. (2022) proposed a framework combining system dynamics, future land-use simulation, and the InVEST model to assess habitat quality under projected land-use changes, accounting for climate change and development strategies.

To the best of our knowledge, counterfactual reasoning with PGMs has not yet been explored in the context of socioecological systems. This approach could offer a new dimension to the study of socioeconomic influences on land-use changes within socioecological systems. In particular, in this paper we adopt the structural causal model (SCM) formalism (Pearl, 2009). As we will see in Section 4, SCMs can successfully handle causal reasoning

from a semantic point of view. In addition, SCMs can be represented as BNs and therefore it is possible to take advantages of the existing methods for inference and learning over the latter. Another advantage is that practitioners familiarized with BNs are likely to find SCMs as a natural way of dealing with causal reasoning.

4. Background and notation

This section provides an overview of fundamental concepts related to PGMs for causal and counterfactual reasoning that will be later instantiated to socioecological systems. With respect to the general notation, upper-case letters are used to denote random variables and lower-case for their possible values (also called states), i.e. given a variable V, v denotes an element of its domain, denoted by Ω_V. We assume that all the variables are discrete. Similarly, $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$ denotes a set of variables and $\mathbf{v}$ an element of $\Omega_{\mathbf{V}} = \times_{V \in \mathbf{V}} \Omega_V$. The probability mass function of a discrete random variable V will be denoted by $P(v) = P(V = v)$.

Causal reasoning consists of three levels (Pearl and Mackenzie, 2018), namely *association*, *intervention* and *counterfactuals*. The first level, association, accounts for predictions based on past observations. At this level, one can answer questions of the form *"What if I see ...?"*. Such questions are called *observational queries* and can be answered using conditional probabilities (estimated from the observational data) stating how likely is that something happens given that something else has happened. In a general setting where x and y are states of the random variables X and Y respectively, an example of an observational query is the computation of the conditional probability $P(x|y)$. Herein, x and y are the positive states (presence) of X and Y, respectively, while x' and y' are their counterpart negative state (absence).

The second level, intervention, is related to questions of the type *"What if I do ...?"*. Such kind of questions can also be formulated in terms of probabilities using *do calculus* (Pearl, 2009). Let Y_x denote the random variable representing Y under the hypothetical scenario in which X is forced to be equal to x. Then the query $P(Y_x = y)$ stands for the probability that Y takes the value y when X is intervened (i.e., forced) to take the value x. With this in mind, we might be interested in estimating the difference between two interventional queries, which is known as the *average causal effect* (ACE),

defined as

$$\mathrm{ACE}(X, Y) = P(y_x) - P(y_{x'}). \tag{1}$$

Note that ACE takes values between -1 and 1. Positive values of ACE mean that Y is more likely to happen when X also happens, while negative values indicate that Y is more likely to happen when X does not happen.

The last level of causation, *counterfactuals*, aims at queries of the form *"What if I had done ...?"*. In terms of probabilities, counterfactual queries tackle hypothetical scenarios like, *"What would the outcome have been if the variable had taken a different value?"*. For instance, the conditional probability $P(Y_x|X = x')$ represents the probability of Y if X had been x instead of x'. Note that Y_x is a variable related to the hypothetical scenario whereas X (without subindex) denotes a variable in the real scenario. Note how counterfactual queries can give us information telling if it was X that caused Y.

Given this semantics of interventional and counterfactual queries, it is possible to specify some typical queries that can be useful for understanding the model but also for defining policies aimed at solving problems, as we will see in the case study in Section 5. In this context, we might need to measure to what extent an event is a necessary condition for another one (i.e., when one event must occur for another one to happen). This can be achieved using the so-called *probability of necessity* (PN) which can be defined as

$$\mathrm{PN}(X, Y) = P(Y_{x'} = y'|X = x, Y = y). \tag{2}$$

X is said to be a necessary cause for Y if whenever y occurs then x has occurred. Therefore, PN can be interpreted as the probability that X is a necessary cause of Y. In other words, it is the probability that the event y would not have occurred in the absence of event x, given that x and y did in fact occur. In our running example, $\mathrm{PN}(I, A)$ measures the degree of certainty with which we can assert that whenever agriculture is present then immigration is also present.

It might also be useful to consider the probability of necessity, but assuming that $X = x$ did not happen. We call it the *probability of necessity with reverse cause* (PNrc), formally defined as

$$\mathrm{PNrc}(X, Y) = P(Y_x = y'|X = x', Y = y). \tag{3}$$

Analogously, we could also be interested in determining if an event is a sufficient condition for another event to happen. For this we can define the *probability of sufficiency* (PS) as

$$\mathrm{PS}(X, Y) = P(Y_x = y | X = x', Y = y'). \tag{4}$$

X is said to be a sufficient cause for Y if whenever x occurs then y will occur. Therefore, PS can be interpreted as the probability that X is a sufficient cause of Y. In other words, it is the probability that setting x would produce y in a scenario where x and y are in fact absent. In our example, $\mathrm{PS}(I, A)$ measures the degree of certainty with which we can assert that whenever immigration is present, agriculture is also present.

An event could also be, to some extent, both necessary and sufficient. It can be measured by the *probability of necessity and sufficiency* (PNS), defined as

$$\mathrm{PNS}(X, Y) = P(Y_x = y, Y_{x'} = y'). \tag{5}$$

X is said to be a necessary and sufficient cause for Y if whenever x occurs then y will occur and vice-versa. Therefore, PNS can be interpreted as the probability that X is a necessary and sufficient cause of Y. Intuitively, PNS measures how Y reacts to X, hence expressing to what extent $X = x$ is necessary and sufficient for $Y = y$. In the example provided, $\mathrm{PNS}(I, A)$ measures the degree of certainty with which we can assert that whenever immigration is present, agriculture is also present, and whenever immigration is absent, agriculture is also absent.

The first level of causation (association) can be properly handled using probability distributions represented as a Bayesian network, where all the possible observational queries can be answered by computing the relevant conditional probabilities directly on the network. However, handling the other two levels of causation (interventions and counterfactuals) requires going beyond conditional probabilities, so that we are able to handle hypothetical (and not only observed) scenarios. In order to approach these kind of scenarios, Pearl (2009) proposed the so-called structural causal models (SCMs), which are a specific type of probabilistic graphical model used for causal and counterfactual reasoning. SCMs distinguishes between two types of nodes: *endogenous* nodes, which represent the variables within the modeled problem, for which data is available, and *exogenous* nodes, which are associated with external factors for which data is not available. Note that the terms node and variable are used interchangeably. It can be shown that SCMs can be regarded as Bayesian networks that have been extended to accommodate the exogenous variables (see Appendix A for the technical details).

An example of how counterfactual queries are computed over SCMs is given in Appendix B. Note that if all the conditional probability distributions (CPTs) involved in an SCM are known, all the counterfactual queries described above can be answered by using standard inference algorithms for BNs. However, in problems where only observational data is available, and particularly in socioecological systems, the parameters of the CPTs corresponding to the exogenous variables cannot be uniquely determined, because there is no data about those variables. When this problem arises, it is said that the counterfactual query is *unidentifiable* (Correa et al., 2021; Wu et al., 2019).

Nonetheless there are methods able to deal with problems where only observational data is available, but instead of precise probability values, they provide intervals bounding them (Tian and Pearl, 2000; Zaffalon et al., 2020). In our specific case study, we propose the utilization of the innovative technique known as EMCC (*Expectation Maximization for Causal Computation*) as detailed by Zaffalon et al. (2023, 2024).

5. Case study

5.1. Problem and data description

To illustrate the potential of counterfactual reasoning with PGMs, let us consider the ecological and socio-economic data chosen for the case study, which is related to the region of Andalusia, in southern Spain (Figure 2 (a)). This area of study is a socioecological system with a strong relation between the natural and socio-economic components. It also shows high variability regarding elevation, ranging from 0 to 3460 m above the sea level. The main mountain ranges within the study area are the Sierra Morena mountain range in the North and the Baetic Systems in the South, with the Baetic Depression serving as the geological boundary between them. The Guadalquivir River flows through this depression, being the largest river in Andalusia. The flattest areas correspond to the littoral and the Baetic depression, while the steepest ones correspond to the Baetic Systems. Therefore, Andalusia can be divided into 4 main geomorphological units: the Baetic Depression, the Sierra Morena mountain range, the Baetic Systems and the Littoral, as shown in Figure 2 (b).

The Baetic Depression is characterized by its high agricultural production, mainly comprising rain-fed herbaceous crops in the low-lying plains and irrigated herbaceous crops along the Guadalquivir river (Figure 2 (c)). The

Sierra Morena mountain range is characterized by having high emigration and mortality rates and low birth rate, which results in population decline (Figure 2 (d)), and is predominantly covered by rain-fed crops and dehesa, a heterogeneous system exhibiting various states of ecological maturity, with shepherding being the principal economic activity. The Baetic Systems have the highest elevation and steepness in the study area. This area is predominantly cloaked in natural vegetation, with extensive woody crops as a secondary feature. Its rugged terrain discourages the adoption of intensive agricultural practices. Finally, the Littoral, densely populated and characterized by high temperatures, features abundant natural vegetation and serves as the primary location for the majority of greenhouses in the study area, making it a focal point of agricultural activity.

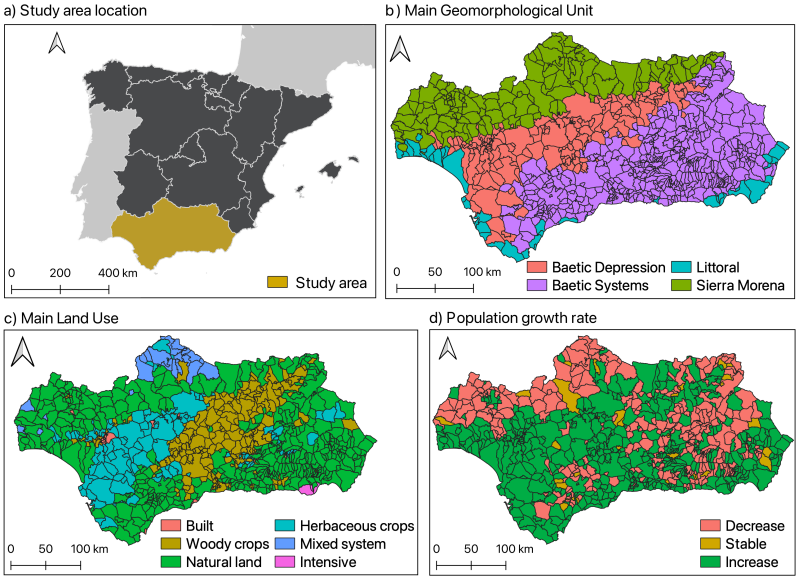

Figure 2: Study area (Andalusia, Spain) (a) and municipalities within the study area, color-coded based on their primary geomorphological unit (b), their main land use (c), and their population growth rate (d). For (b) and (c), in cases where a municipality encompasses more than one geomorphological unit or land use, the color represents the larger or dominant one within that municipality.

In a previous study (Maldonado et al., 2018), 75 different variables representing social, economic and ecological characteristics of the study area were employed to study how socioeconomic changes influence the generation of ecosystem services using BNs with no causal (nor counterfactual) reasoning

conducted at that time. These variables, available in public repositories, were sourced from the Multi-territorial Information System of Andalusia (SIMA) and the Andalusian Environmental Information Network, and municipalities within the study area were taken as the modeling unit. The dataset considered contains 830 instances, one per municipality. In our current study, we narrowed our focus to a subset of 17 variables from the original dataset, to conduct a causal and counterfactual analysis using SCMs. These include land use, social, and economic variables which are detailed in Tables 2, 3, and 4, respectively.

Table 2: Variables representing the ecological dimension used to construct the SCM.

Ecological dimension			
Name	**Description**	**State**	**Threshold**
MGU	The main geomorphological unit a municipality belongs to.	Baetic Depr. Sierra Morena Baetic Sys. Littoral	
Built	Percentage of artificial or built areas in each municipality - including *urban; industrial; mining; freight and technical infrastructures.*	Scarce Fair Abundant	< 5 5 - 30 > 30
GH	Percentage of intensive agriculture (greenhouses) in each municipality.	Scarce Fair Abundant	< 5 5 - 30 > 30
HCrops	Percentage of herbaceous crops in each municipality - including rainfed and irrigated crops.	Scarce Fair Dominant	< 15 15 - 50 > 50
WCrops	Percentage of woody crops in each municipality - including rainfed and irrigated crops.	Scarce Fair Dominant	< 15 15 - 50 > 50
Mixed	Percentage of heterogeneous lands in each municipality - including patches mixing *grassland and forest* and *crops with natural vegetation.*	Scarce Fair Dominant	< 10 10 - 40 > 40
Natural	Percentage of natural areas in each municipality - comprising *bush; grassland; forest; bush and forest; wetlands and naked soil.*	Scarce Fair Dominant	< 25 25 - 60 > 60

Table 3: Variables representing the social dimension used to construct the SCM.

Social dimension			
Name	**Description**	**State**	**Threshold**
Pop	Population density of each municipality in 2011 (inhabitants/Km^2).	Low Moderate High	< 35 35- 150 > 150
SR	Sex ratio. Proportion of males (M) to females (F) in each municipality in 2011 (computed as $SR = \frac{M}{M+F}$).	More females More males	≤ 0.50 > 0.50
EGR	Population growth rate. Exponential growth of the population, computed as $EGR = \frac{\ln(P_t/P_0)}{t}$, where P_0 represents the population in 2001, P_t the population in 2011 and t the 10-year period.	Decrease Stable Increase	< -0.03 -0.03 - 0.03 > 0.03
IME	Index of Migration effectiveness. Percentage of total migration for the period 2001-2011. It ranges from -100 (emigration) to 100 (immigration), with values close to 0 indicating no change in the population dynamic. It is computed as $IME = \frac{Immigration - Emigration}{Immigration + Emigration} \times 100$.	Emigration Balanced Immigration	< -2 -2 - 2 > 2
ODI	Old-age dependency index. Percentage of the older over the younger population in 2011, computed as $ODI = \frac{P_{>65}}{P_{<15}} \times 100$, where $P_{>65}$ is the population older than 65 years old and $P_{<15}$ is the population younger than 15 years old.	Low Moderate High	< 25 25 - 40 > 40
Death	Mortality rate. Number of deaths per 1000 inhabitants in each municipality in 2011.	Low Moderate High	< 9 9 - 15 > 15
Birth	Birth rate. Number of births per 1000 inhabitants in each municipality in 2011.	Low Moderate High	< 5.6 5.6 - 10.3 > 10.3

Table 4: Variables representing the economic dimension used to construct the SCM.

Economic dimension			
Name	**Description**	**State**	**Threshold**
WF	Workforce. Percentage of the municipality's working age population (≥ 16) that are available to work in 2011. It is computed as $WF = \frac{ER+UR}{P_{\geq 16}} \times 100$; where ER is the Employment Rate; UR is the Unemployment Rate and $P_{\geq 16}$ is the population older than 16 years old.	Low Average High	< 0.55 0.55 - 0.62 > 0.62
SSE	Secondary sector employment. Number of employees in the secondary sector per 1000 inhabitants.	Low Moderate High	< 70 70 - 113 > 113
TSE	Tertiary sector employment. Number of employees in the trading, banking and service sectors per 1000 inhabitants.	Low Moderate High	< 78 78 - 122 > 122

The primary goal of this study is to investigate how different variables of interest in this socioecological system are influenced by other variables. In this context, such variables of interest are termed *effect variables*, while the other factors that could potentially influence these effects are referred to as *cause variables*. In connection to Section 4, X and Y represent the cause and effect variables, respectively. Specifically, we consider as effects those variables representing the different land uses and the population growth (i.e., *Built*, *GH*, *HCrops*, *WCrops*, *Mixed*, *Natural* and *EGR*). We will refer to the union of the cause and effect variable sets as the *variables of interest*.

5.2. Data preprocessing and model definition

To conduct any causal or counterfactual analysis, it is necessary to define the cause and effect variables in such a way that their values can be categorized into two groups. In this case, since the variables are categorical (with multiple states), they are transformed into a binary format by collapsing their values into two distinct states: one representing a positive outcome and the other a negative outcome. Table 5 shows the partitions of the states considered for each variable in the dataset. Another reason for the binarization is computational efficiency. Although only the cause and effect variables need to be binarized, leaving the remaining variables in their original multi-state representation would require re-training the model for each possible

pair of cause and effect variables, leading to a substantial increase in computational cost. By binarizing all variables in advance, we can reuse the same learned model across multiple causal queries, thereby reducing redundancy and improving overall efficiency.

Table 5: Categorization into positive and negative values.

Variable	Positive state	Negative state
MGU	Littoral, Baetic Depression	Baetic System, Sierra Morena
Built	Abundant, Fair	Scarce
GH	Abundant, Fair	Scarce
Hcrops	Dominant, Fair	Scarce
Wcrops	Dominant, Fair	Scarce
Mixed	Dominant, Fair	Scarce
Natural	Dominant, Fair	Scarce
Pop	High	Low, Moderate
SR	More females	More males
EGR	Increase	Decrease, Stable
IME	Immigration	Emigration, Balanced
ODI	Low	Moderate, High
Death	Low	Moderate, High
Birth	High	Low, Moderate
WF	High	Low, Average
SSE	High	Low, Moderate
TSE	High	Low, Moderate

Given that the counterfactual analysis relies on SCMs, a causal structure in the form of a DAG is required. To establish this structure, we adopt the BN structure used in the aforementioned prior investigation (Maldonado et al., 2018), initially formulated by domain experts. In that work, the approach used to build the BN was based on the DPSIR framework (European Environment Agency, 2007). In this context, socioeconomic variables are considered the Drivers (as defined in the DPSIR framework) of environmental change. These Drivers produce different Pressures, which are reflected as changes in land use, ultimately altering the State of the ecosystem. This, in turn, affects the supply of ecosystem services and human well-being (Impacts). Finally, Responses are the different actions that governments and society take to control the Drivers. To align with this framework, mean-

ingful variables and their causal relationships were identified through expert knowledge and relevant literature.

Nonetheless, in this new work some modifications are made on the structure of the original BN with the purpose of reducing its complexity, resulting in the graph shown in Figure 3. First, the graph is restricted to our effect variables and their ancestors, i.e. the variables of interest previously detailed in Tables 2, 3, and 4. Note that any descendant from the effect variables in the original graph is irrelevant for a causal analysis. Additionally, the node MGU is a common ancestor of all the cause and effect variables considered, and there exists no alternative path connecting them with the ancestors excluded from the final graph. Consequently, our variables of interest are independent from the rest of ancestors given MGU. Secondly, the number of parents is limited to a maximum of three. The intention of this is to prevent extremely large exogenous variables, whose number of states increases exponentially with the number of parents. Appendix C provides a further discussion about this. The choice of a maximum of three parents was made based on preliminary experiments, which showed that the computation with larger models was extremely time-consuming. Hence, the arc from MGU to Pop was removed. Note that there are alternative directed paths between both nodes, i.e., it is redundant. Moreover, this deletion is the one involving the lowest decrease in the likelihood of the model given the data.

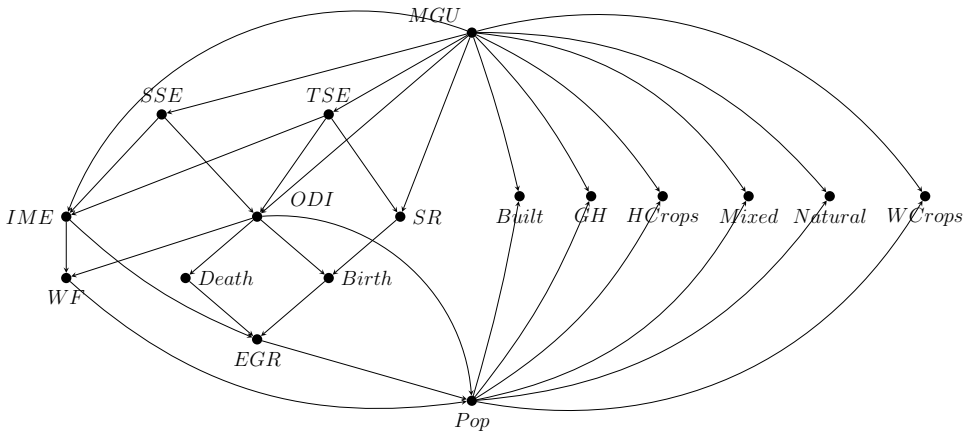

Figure 3: BN obtained from the model in a previous study (Maldonado et al., 2018) by restricting it to our variables of interest and by limiting the number of parents to 3.

After establishing the causal structure, an SCM can be derived by in-

troducing an exogenous parent for each endogenous variable. The resulting graph is illustrated in Figure 4. As for the parameters, the SEs are defined in a canonical form. By contrast, the marginal distributions over the exogenous variables are considered to be unknown and estimated using EMCC. For more details about EMCC, readers can reffer to (Zaffalon et al., 2020, 2024). Specifically, this method is executed with 100 EM runs while capping the EM convergence at 300 iterations and using the 830 data instances available. .

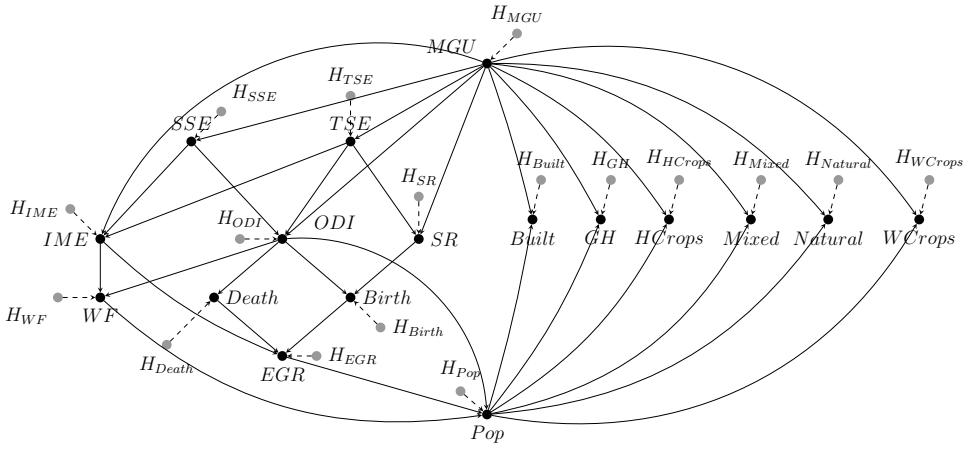

Figure 4: Markovian SCM used for intended counterfactual analysis. All the SEs are asumed to be canonical. Each endogenous variable has only one exogenous cause, and each exogenous variable is cause of only one endogenous one.

Initially, we examine the *difference in conditional probability*, denoted as $P(y|x) - P(y|x')$. This aligns with the conventional BN analysis. Moving to causal (non-counterfactual) reasoning, we investigate the interventional query ACE, defined in Eq. (1). To underscore the advantages of counterfactual reasoning, we evaluate the queries PN, PNrc, PS, and PNS, outlined in eqs. (2) to (5). The variables under study (i.e., effects) are those enumerated in the preceding section, while the causes vary based on the specific effect. This distinction arises from the requirement that causes must be ancestors of the effects. In this sense, in the case of land-use variables, the causes include WF, EGR, MGU, SR, SSE, TSE, ODI, Pop, $Death$, and $Birth$. When EGR is treated as the effect, Pop and the variable itself are excluded as potential causes. It is important to note that the insights are drawn at a regional scale (Andalusia), and are based on information specifically gathered

within this region.

In relation to the implementation, the CREDICI software (Cabañas et al., 2020) was used, a Java library designed for causal reasoning, featuring the implementation of EMCC. For more extensive implementation details and the code to replicate the study, please refer to our GitHub repository[1]. The computation was made in a computer cluster made of 1024 cores (2048 threads). This allowed the parallelization of the execution of each of the 100 EM runs, each of them taking more than 4 hours on average.

6. Results and discussion

Before we present the experimental results, we give a brief explanation of how to interpret them. As indicated at the end of Section 4, the results are given as probability intervals rather than as single values. Note that, a variable is considered to be a necessary or sufficient cause (or both) of another if the associated probability interval is narrow and both the upper and lower limits of the interval are high. For instance, a probability of necessity in the interval [0, 0.99] is too wide to be informative. On the other hand, an interval of [0, 0.1] indicates with high certainty that the variable is not necessary for the given effect. Similarly, an interval of [0.9, 0.99] suggests strong evidence supporting that the variable is necessary for the effect to occur.

To begin with the analysis of the experimental results, we first consider as *effect* the variable *EGR* (population growth rate), whose results are given in Figure 5. The variable with a clearer causal impact on *EGR* is *IME* (index of migration effectiveness). Remarkably, the probability of sufficiency, PS, is bounded between 0.73 and 0.85 (Figure 5 (e)), which means that it is very probable that a positive value of *IME* (immigration) is enough to produce a positive value of *EGR* (increase) despite the values of the other variables. Besides, it is also quite likely that *IME* is a necessary condition for *EGR* to have a high value, since $\mathrm{PN}(IME, EGR) \in [0.60, 0.75]$ (Figure 5 (c)). The probability of necessity and sufficiency is also bounded above 0.5, more precisely in the interval $[0.51, 0.63]$ (Figure 5 (f)).

Regarding the probability of sufficiency of the other variables under consideration, *MGU* (main geomorphological unit) and *TSE* (tertiary sector employment) are likely to be sufficient for *EGR* to take place, reaching values in

[1] `https://github.com/PGM-Lab/2025-counterfactual-land`

the intervals $[0.54, 0.78]$ and $[0.63, 0.82]$, respectively. Conversely, *SSE* (secondary sector employment), *ODI* (old-age dependency index), *Death* (death rate) and *Birth* (birth rate) have a good chance to be sufficient, since their lower bounds are close to 0.4 (Figure 5 (e)).

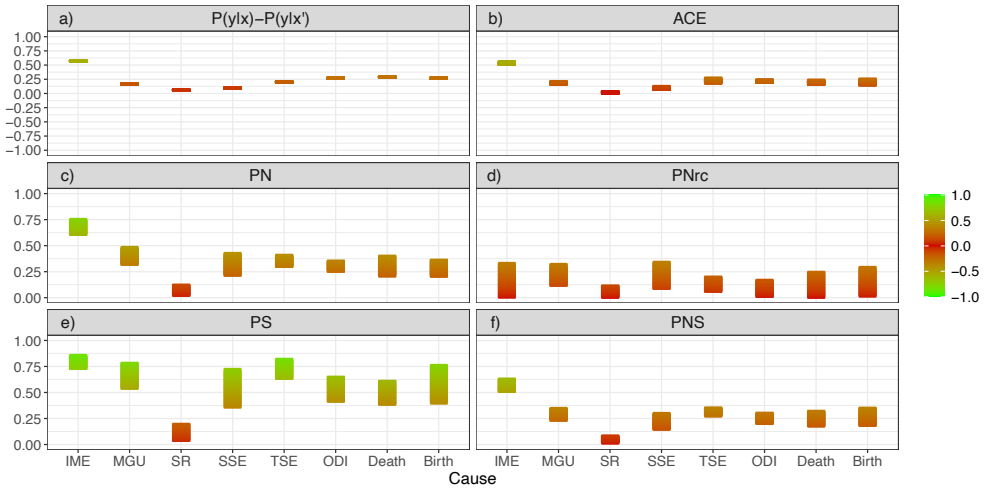

Figure 5: Intervals for the queries with *EGR* as effect variable. The x-axis represents each cause variable, according to the graph in Figure 3, while the y-axis shows the query metric. Panels (a) to (f) depict different types of analysis: (a) conventional BN analysis, (b) causal analysis, and (c-f) counterfactual analysis. The metrics illustrated are (a) the difference in conditional probability, $P(y|x) - P(y|x')$; (b), the average causal effect, ACE; (c) the probability of necessity, PN; (d) the probability of necessity with reverse cause, PNrc; (e) the probability of sufficiency, PS; and (f) the probability of necessity and sufficiency, PNS. Note that metrics in panels (a) and (b) can take negative values, as they are defined as differences of probabilities.

With respect to non-counterfactual queries (Figure 5 (a-b)), the difference in conditional probability (0.58) and the average causal effect in the interval $[0.50, 0.56]$ support the classification of positive *IME* as a cause of positive *EGR*. Note that while non-counterfactual queries enable the identification of variables influencing *EGR*, these queries do not provide information about the nature of the relationship (whether it is one of necessity or sufficiency). This underscores the nuanced insights that counterfactual reasoning can offer in understanding causal relationships within the studied context.

Therefore, a positive immigration rate turns out to be both necessary and sufficient (with high probability) in order to achieve a positive population growth rate (Parsons and Smeeding, 2006; Viñuela et al., 2019; Viñuela,

2022). On the other hand, it is also likely that the location of the municipality and the tertiary sector employment can cause a positive *EGR* value, but this is only expressed in terms of sufficiency. As a matter of fact, Figure 2 (d) indicates a population decline in the majority of municipalities situated in both the Baetic Systems and Sierra Morena regions. Considering the remaining causes, population growth can occur in the absence of all of them, though they might be sufficient on their own to drive population growth. Population growth is a phenomenon influenced by various factors, including immigration, emigration, death, and birth rates (Lutz, 2006; Poston Jr and Bouvier, 2010). The variable *IME* is an index that incorporates both immigration and emigration rates. A high value of *IME* indicates more immigration than emigration, resulting in population increase, assuming that other factors remain constant. In contrast, death and birth rates are considered separately rather than jointly in a single rate of natural population increase. Individually, neither low death rate nor high birth rate is necessary, for population growth.

The remaining *effect* variables are all referred to land uses. For variable *Built* which represents the percentage of built or artificial areas in a municipality, only two variables turn out to have a significant causal effect, namely the location of the municipality, *MGU* and its population density, *Pop*. It is no wonder that *Pop* is the most clear sufficient cause of positive value of *Built*, with $\mathrm{PS}(Pop, Built) > 0.93$ (Figure 6 (e)). However, the probability of necessity for this variable is lower, but still remarkable, between 0.65 and 0.70 (Figure 6 (c)). The same probability reaches higher values for the location of the municipality, reaching the interval $[0.62, 0.82]$. Nonetheless, the probability of sufficiency of *MGU* is in $[0.55, 0.73]$, which is notably lower than that for the other variable. Moreover, both causes have a good chance of being necessary and sufficient for *Built* to be abundant or fairly abundant (Figure 6 (f)). Hence, the results of the queries seem to indicate that the location (flat vs. mountainous areas) is fundamental when determining the percentage of built area, but also in combination with the population density (Ehrlich et al., 2021; Thornton et al., 2022).

Figure 7 shows the results of the queries related to variable *HCrops*, which is the percentage of herbaceous crops in the municipality. In this case, only variable *MGU* reaches a value clearly above 0.5, and only for one query, the probability of necessity, which is estimated to be in the interval $[0.59, 0.99]$ (Figure 7 (c)). It means that it is quite likely that a positive value of *MGU* (littoral, Baetic depression) is necessary for a positive value of *HCrops* (dom-

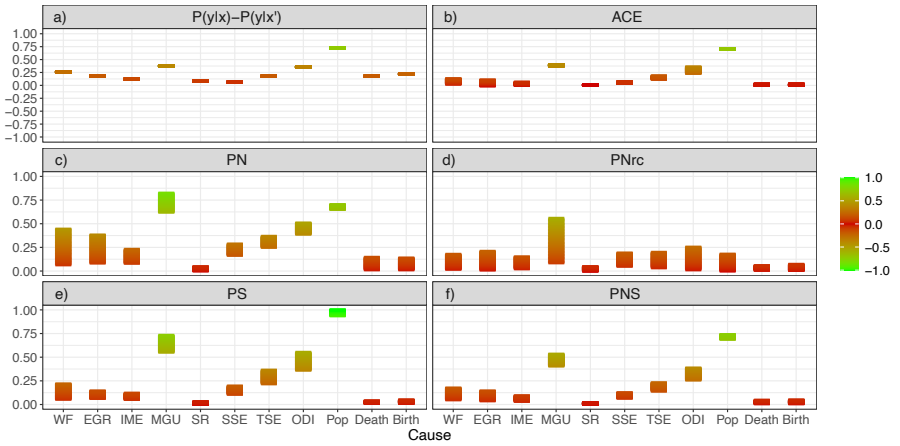

Figure 6: Intervals for the queries with *Built* as effect variable.

inant, fair) to be observed, but it is not so clearly rendered as sufficient too, since PS($MGU, HCrops$) $\in [0.44, 0.74]$ (Figure 7 (e)), however there is still some chance that it is sufficient. On the other hand, *WF* and *Pop* show a high uncertainty with respect to the probability of necessity, since their intervals are considerably wide (Figure 7 (c)). The remaining variables are clearly not necessary, nor sufficient, for *HCrops*. The findings align coherently with the insights derived from Figure 2 (c). Herbaceous crops emerge as the predominant land use across a significant portion of the Baetic Depression, whereas they do not hold the same prevalence in other geomorphological units (Molero and Marfil, 2017).

In the case of the percentage of greenhouses (variable *GH*), the results are displayed in Figure 8. It is apparent from the plots that only three variables, *MGU*, *Pop* and *ODI*, have a causal impact on the target variable *GH* (Figure 8 (c)). The most remarkable effect is observed in terms of the probability of necessity, with values in the intervals $[0.72, 1]$, $[0.88, 1]$ and $[0.69, 0.91]$ for *MGU*, *Pop* and *ODI*, respectively, which indicates that it is considered almost certain that all three variables must have a positive value for *GH* to have a positive value as well (Aznar-Sánchez et al., 2011; Mendoza-Fernández et al., 2021). In addition, variable *TSE*, with the upper bound for PN above 0.5 and the lower one close to 0.5, has a good chance to be necessary (Galdeano-Gómez et al., 2013). It is also quite significant that none of the variables is sufficient condition by its own (Figure 8 (e)). With respect to the non-counterfactual queries, none of the insights previously

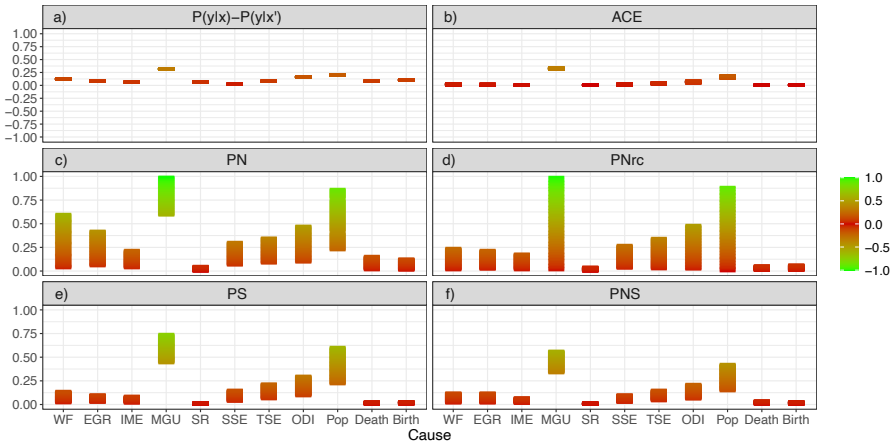

Figure 7: Intervals for the queries with *HCrops* as effect variable.

mentioned are reflected in such queries (Figure 8 (a-b)), which showcases the advantage of counterfactual reasoning. The results coherently reflect the predominant location of greenhouses in the Littoral (Figure 2 (c)). However, it is important to note that the positive state of *MGU* encompasses both the Littoral and the Baetic Depression, with the latter not exhibiting a notably high density of greenhouses. Therefore, *MGU* is necessary but not sufficient condition for *GH* (Wolosin, 2008).

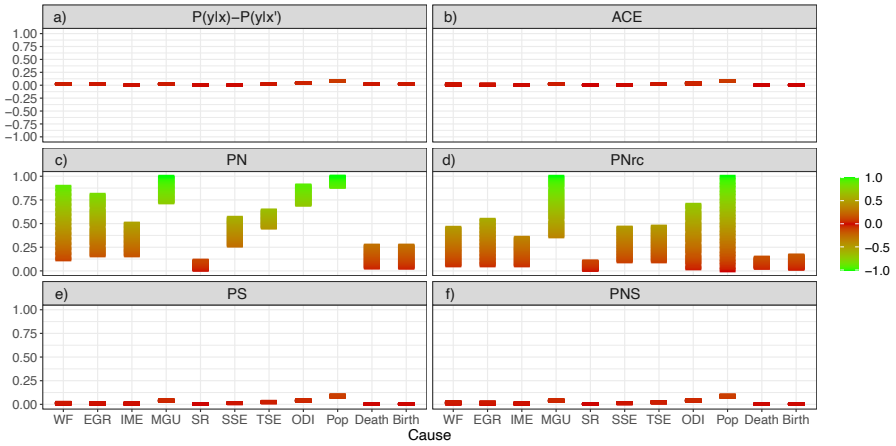

Figure 8: Intervals for the queries with *GH* as effect variable.

The results for variable *Natural*, representing the percentage of natural areas, can be seen in Figure 9. In terms of the difference in conditional proba-

bility and average causal effect, only the location (MGU) shows a remarkable variation (Figure 9 (a-b)). Regarding the probability of necessity, all variables except MGU and Pop are very unlikely to be necessary, due to the low value of the upper bounds of their intervals. Even the two mentioned variables are not clearly necessary, since most of their intervals are below 0.5, and their widths indicate uncertainty about the probability value (Figure 9 (c)). It is also clear that all variables except MGU and Pop are very unlikely

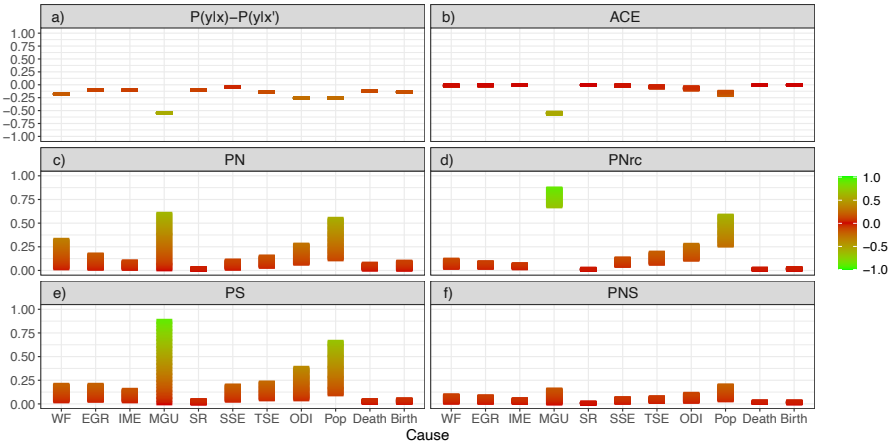

Figure 9: Intervals for the queries with $Natural$ as effect variable.

to be sufficient to increase the natural spaces. The intervals for MGU and Pop are compatible with them being sufficient conditions, but their amplitude indicate that there is considerable uncertainty (Figure 9 (e)). The most remarkable insight about the natural areas can be obtained from the high value in the query PNrc with MGU as cause variable (Figure 9 (d)). This shows that not being in the littoral nor in the Baetic depression is necessary to have mainly a natural land use, since $\text{PNrc}(MGU, Natural) \in [0.67, 0.87]$. In other words, mountainous areas (Sierra Morena and Baetic Systems) are necessary for the natural areas to be dominant over the other land uses. The results align with the fact that natural lands are predominantly located in the Baetic Systems and Sierra Morena mountain ranges, as shown in Figure 2 (c) (Gratzer and Keeton, 2017; Snethlage et al., 2022).

Figure 10 shows the results for variable $WCrops$, which measures the percentage of woody crops in each municipality. It can be seen that none of the variables shows a significant variation in conditional probability or remarkable values of ACE (Figure 10 (a-b)). Considering the queries related

to necessity and sufficiency, *MGU* and *Pop* might have a significant influence on *WCrops*, but the width of the corresponding intervals poses a high level of uncertainty on such statement (Figure 10 (c-f)). The rest of possible causes are clearly not relevant.

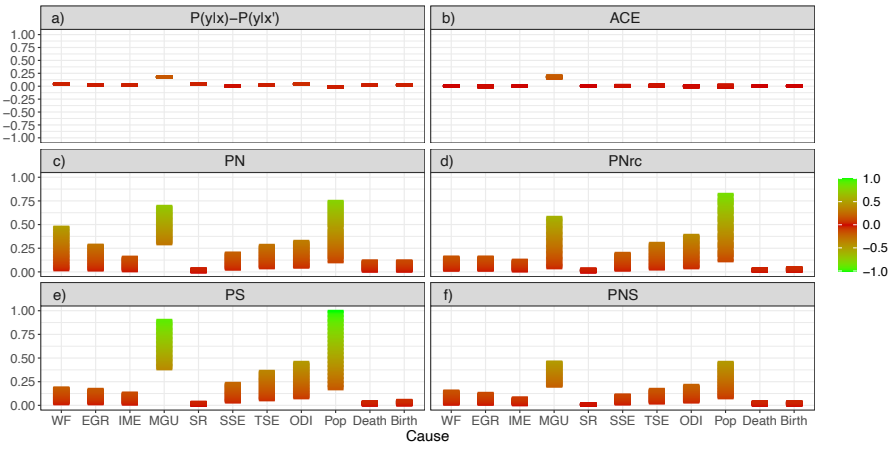

Figure 10: Intervals for the queries with *WCrops* as effect variable.

Finally, Figure 11 displays the results for variable *Mixed* (percentage of heterogeneous lands). The only causal effect in terms of difference in conditional probability and ACE is provided by variables *Pop* and *MGU* (Figure 11 (a-b)). It is highly unlikely that all variables except *MGU* and *Pop* are necessary (see PN, PNrc and PNS in Figure 11 (c), (d) and (f), respectively), but even for these two variables there is paramount uncertainty in terms of PN. However, PNrc clearly points towards the facts that not being in the littoral nor in the Baetic depression and having a low population density are both necessary conditions. On the other hand, the values of PS and PNS clearly show that none of the variables are sufficient for *Mixed* to have a positive value (Figure 11 (e-f)). The results are consistent with the fact that heterogeneous lands are the main land use in some areas of Sierra Morena (Muñoz-Rojas et al., 2011; Plieninger et al., 2021), coinciding with a concurrent trend of population decline in those regions, as shown in Figure 2 (c,d) (Plieninger and Wilbrand, 2001).

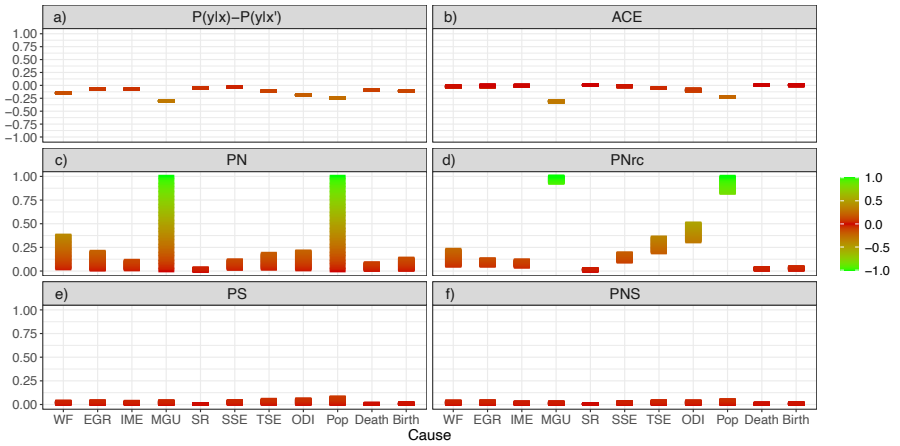

Figure 11: Intervals for the queries with *Mixed* as effect variable.

7. Conclusions and policy recommendations

7.1. Conclusions

This paper proposes the application of counterfactual reasoning with BN for analyzing socioecological systems. This approach tackles a limitation of traditional probabilistic analysis, which cannot determine the nature of the relation between variables (whether it is one of necessity or sufficiency). To address this, we suggest employing a recently developed technique based on the well-known EM algorithm for parametric learning in BNs with latent variables. The advantages of this method are, first, that it allows the analysis of observational data alone, unlike other methods that also require interventional data, which may not be feasible to obtain. Secondly, since it is based on well-established methods for BNs, it can be easily adopted by researchers from diverse fields.

Note that the primary requirement for adopting this framework is the definition of the causal graph. In the current study, this structure is specified by domain experts, ensuring interpretability. However, expert-derived structures may be constrained by the availability of expert knowledge and may not generalize well across domains. As future work, we will explore hybrid approaches that integrate expert knowledge with data-driven structure learning techniques.

To demonstrate the utility of counterfactual reasoning, we have presented a case study using an observational dataset containing information on socioeconomic factors and land uses in southern Spain. Concerning population

dynamics, our study indicates that immigration not only is necessary but also sufficient for population growth. With respect to land uses, location (flat vs. mountainous areas) and population density are the most influential factors. In particular, a non-mountainous area and high population density are likely to be necessary conditions for the presence of built areas, greenhouses, herbaceous crops, and woody crops. These factors are also likely to be sufficient conditions for built areas and herbaceous crops. Conversely, a mountainous area and low population density are likely to be necessary conditions for the presence of natural areas and mixed crops. All these findings underscores the power of counterfactual reasoning in uncovering relationships within socioecological systems, in contrast to the plain use of BNs for which only observational queries can be solved.

While this study provides valuable insights, further research could be carried out to consider counterfactual queries with multiple cause variables, for instance to determine whether simultaneous causes are jointly sufficient. It could also be of interest to analyze such queries conditioned on other relevant variables. Additionally, this methodology could be of considerable interest for other topics within the environmental and ecological areas, such as the study of species distribution or risk assessments.

7.2. Policy recommendations

Land-use change is a major driver of the distribution and functioning of ecosystems. In rural areas, these changes have mainly occurred in two opposing directions: agricultural intensification in highly productive, typically flat regions; and the abandonment of traditional practices in less productive, often mountainous areas. Both trends are primarily driven by socioeconomic factors. Therefore, intensification and abandonment are critical issues on the political agenda of the European Union.

European environmental policy should be designed considering the inherent uncertainty of socio-ecological systems. The application of counterfactual reasoning using BNs at a regional scale (Andalusia) enables the identification of variables that can be considered socio-ecological indicators. These indicators provide valuable information for policymakers: environmental planning in flat and mountainous areas must follow different strategies. Flat areas, characterized by high population density, urban development, and intensive agriculture, should consider proposals for sustainable intensification. Mountainous areas, associated with low population density, natural landscapes,

and heterogeneous crops, should prioritize policies aimed at preventing rural abandonment. Such policies should be embedded within holistic rural development programs.

Code and data availability

The code and data-set used in the current study is available at the repository at `https://github.com/PGM-Lab/2025-counterfactual-land`.

CRediT authorship contribution statement

Conceptualization, R.C., A.D.M. and P.A.A.; methodology R.C; software R.C. and M.M.; data curation A.D.M. and M.M.;formal analysis R.C., A.D.M. and A.S.; writing–original draft preparation, R.C., A.D.M., M.M., P.A.A. and A.S.; funding acquisition, A.S.; supervision, P.A.A. and A.S. All authors have read and agreed to the published version of the manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors did not use any form of generative AI during the preparation of this work. The writing and editing process was done by human hand.

Funding

Grant PID2022-139293NB-C31 funded by MICIU/AEI/10.13039/501100011033 and by ERDF A way of making Europe. R.C. acknowledges the support by Spanish Ministry of Science, Innovation and Universities through the "María Zambrano" grant (RR_C_2021_01) funded with NextGenerationEU funds. Rafael Cabñas was also supported by "Plan Propio de Investigación y Transferencia 2024-2025" from University of Almería under the project P_LANZ_2024/003. Partly supported by the University of Almería Research and Transfer Programme funded by "Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía" through the European Regional Development Fund (ERDF), Operation Programme 2021-2027. Programme: Research and Innovation 54.A.

Declaration of competing interest

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

Appendix A. Structural causal models and Bayesian networks

Structural causal models (SCMs) (Pearl, 2009) are a specific type of probabilistic graphical model used for causal and counterfactual reasoning. SCMs can be formally defined as follows (Bareinboim et al., 2022).

Definition 1 (Structural causal model (SCM)). *A structural causal model* $\mathcal{M}$ *is a 4-tuple* $\langle \mathbf{U}, \mathbf{X}, \mathcal{F}_{\mathbf{X}}, \mathcal{P}_{\mathbf{U}} \rangle$*, where*

- $\mathbf{U}$ *is a set of exogenous variables that are determined by factors outside the model;*
- $\mathbf{X}$ *is a set of variables* $\{X_1, X_2, \ldots, X_n\}$*, called endogenous, that are determined by other (exogenous and endogenous) variables in the model, i.e. by variables in* $\mathbf{U} \cup \mathbf{X}$*.*
- $\mathcal{F}_{\mathbf{X}}$ *is a set of functions* $\{f_{X_1}, f_{X_2}, ..., f_{X_n}\}$ *called structural equations (SE), such that each of them is a function* $f_{X_i} : \Omega_{\mathbf{U}_i} \cup \Omega_{\mathrm{Pa}_{X_i}} \rightarrow \Omega_{X_i}$*, where* $\mathrm{Pa}_{X_i} \subseteq \mathbf{X}$ *are the endogenous variables directly determining* X_i *and* $\mathbf{U}_i \subseteq \mathbf{U}$ *are the exogenous variables directly determining* X_i*.*
- $\mathcal{P}_{\mathbf{U}}$ *is a set containing a probability distribution* $P(U)$ *for each* $U \in \mathbf{U}$*.*

Note that the structural equations $\mathcal{F}_{\mathbf{X}}$ actually define a directed acyclic graph (DAG) $\mathcal{G}$ called the *causal graph* of the model, whose nodes correspond to the variables in $\mathbf{U} \cup \mathbf{X}$ and containing a link from each variable in $\mathbf{U}_i \cup \mathrm{Pa}_{X_i}$ to X_i, $i = 1, \ldots, n$.

As an illustrative example, we will begin by examining Figure A.12, which depicts two potential causal graphs for SCMs extending the BN presented in Figure 1. The endogenous variables, represented as black nodes, correspond to the variables originally found in the initial BN, specifically $\mathbf{X} = \{M, I, A\}$. These variables retain their original domains. In addition to the endogenous variables, the causal graphs also incorporate exogenous variables, which are

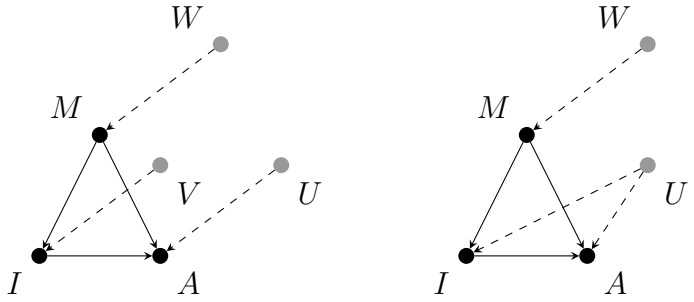

Figure A.12: Examples of two possible causal graphs for the problem shown in Figure 1. U, V and W are exogenous variables.

depicted as gray nodes. In the left graph, the set of exogenous variables is $\mathbf{U} = \{U, V, W\}$, while in the right graph, it is $\mathbf{U} = \{U, W\}$.

The distinction between the two causal graphs is as follows: the graph on the left assumes the absence of any exogenous (and hidden) confounder between any pair of variables. In contrast, in the graph on the right, variables I and A are both influenced by the exogenous variable U. According to the classification given by Avin et al. (2005), a SCM with a graph like the one on the left is termed *Markovian*, meaning that each exogenous variable has only one endogenous child. Conversely, a SCM with a graph like the one on the right is termed *semi-Markovian*, indicating that any of the exogenous variables can have more than one endogenous child[2]. It is important to note that in both cases, all endogenous variables must have exactly one exogenous parent. In other words, given the endogenous variables, each exogenous variable is independent of the others, which is denoted as $U_i \perp\!\!\!\perp U_j | \mathbf{X}$ for all $U_i, U_j \in \mathbf{U} \times \mathbf{U}$ with $i \neq j$. If this condition is not satisfied the model will be classified as *non-Markovian*.

A parameterization for the previously mentioned SCM is illustrated in Figure A.13. The set of marginal distributions associated with the exogenous variables is $\mathcal{P}_{\mathbf{U}} = \{P(U), P(V), P(W)\}$. Conversely, the set of SEs is represented by $\mathcal{F}_{\mathbf{X}} = \{f_A(U, I, M), f_I(V, M), f_M(W)\}$.

In the formalism of SCMs, SEs are typically assumed to be provided, often derived from expert knowledge. Alternatively, SEs can be automatically

[2]Some authors consider a less general definition and limit exogenous variables in semi-Markovian models to have no more than 2 children (Huang and Valtorta, 2006).

$$P(W) = \begin{matrix} w_1 & w_2 \\ [0.7253 & 0.2747] \end{matrix} \qquad P(V) = \begin{matrix} v_1 & v_2 & v_3 & v_4 \\ [0.56312 & 0 & 0.19565 & 0.24123] \end{matrix}$$

$$P(U) = \begin{matrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} \\ [0 & 0.24979 & 0 & 0 & 0.03045 & 0.30418 & 0 & 0.14545 & 0 & 0.27013] \end{matrix}$$

$$f_M(W) = \begin{matrix} w_1 & w_2 \\ [yes & no] \end{matrix} \qquad f_I(M,V) = \begin{matrix} v_1 & v_2 & v_3 & v_4 \\ \begin{bmatrix} yes & yes & no & no \\ yes & no & yes & no \end{bmatrix} & \begin{matrix} M = yes \\ M = no \end{matrix} \end{matrix}$$

$$f_A(M,I,U) = \begin{matrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} \\ \begin{bmatrix} yes & yes & yes & yes & yes & no & no & no & no & no \\ yes & no & yes & yes & no & yes & no & no & yes & no \\ yes & yes & no & no & no & yes & yes & yes & no & no \\ yes & yes & yes & no & yes & yes & yes & no & yes & yes \end{bmatrix} & \begin{matrix} M = yes, I = yes \\ M = yes, I = no \\ M = no, I = yes \\ M = no, I = no \end{matrix} \end{matrix}$$

Figure A.13: SEs and marginal distributions for the Markovian SCM in Figure A.12 (left).

inferred from the causal graph, without any loss of generality, via *canonical specification* (Zhang et al., 2022). The states of an exogenous variable will then represent all possible function mappings between its children domains from their respective endogenous parents domains. In this sense, an exogenous variable in a Markovian model, with X as its child, would require the number of states given by the following expression.

$$|\Omega_U| = \begin{cases} |\Omega_X|, & \text{if } Pa_X = \emptyset, \\ |\Omega_X|^{|\Omega_{Pa_X}|}, & \text{otherwise.} \end{cases} \tag{A.1}$$

In Figure A.13, the SEs associated with variables M and I are already canonical, whereas the one for A is not, since the number of possible values of U (10 in this case) does not match $|\Omega_A|^{|\Omega_{Pa_A}|}$ (16 in this case). The SE under the canonical specification for A and the corresponding distribution $P(U)$ are shown in Figure A.14.

As already considered in some works from the literature (Zaffalon et al., 2020, 2024), a SCM can be specified as a BN as follows. First, the graphical component is the same: the causal graph in the SCM is the DAG $\mathcal{G}$ in the BN. In this way, the BN is defined over the union of the exogenous and endogenous sets of nodes, i.e. $\mathbf{V} = \mathbf{U} \cup \mathbf{X}$. Regarding the distributions in $\mathcal{P}_{\mathbf{V}}$, each exogenous variable is associated with the corresponding marginal distribution in $\mathcal{P}_{\mathbf{U}}$, whereas each SE $f_X(Pa_X)$ induces a CPT of the form

$$P(U) = \begin{array}{c} \begin{matrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} & u_{11} & u_{12} & u_{13} & u_{14} & u_{15} & u_{16} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0.16 & 0.12 & 0 & 0 & 0 & 0 & 0.02 & 0.68 & 0 & 0 & 0 & 0.02 \end{bmatrix} \end{array}$$

$f_A(M, I, U) =$

$$\begin{array}{c} \begin{matrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} & u_{11} & u_{12} & u_{13} & u_{14} & u_{15} & u_{16} \end{matrix} \\ \begin{bmatrix} yes & yes & yes & yes & yes & yes & yes & yes & no & no & no & no & no & no & no & no \\ yes & yes & no & no & yes & yes & no & no & yes & yes & no & no & yes & yes & no & no \\ yes & yes & yes & yes & no & no & no & no & yes & yes & yes & yes & no & no & no & no \\ yes & no & yes & no & yes & no & yes & no & yes & no & yes & no & yes & no & yes & no \end{bmatrix} \end{array} \begin{array}{l} M = yes, I = yes \\ M = yes, I = no \\ M = no, I = yes \\ M = no, I = no \end{array}$$

Figure A.14: Canonical SE for A and marginal distribution for exogenous variable U in the Markovian SCM in Figure A.12.

$P(X|Pa_X)$, characterized by containing only ones and zeros. More precisely, for each $(x, \pi_X) \in \Omega_X \times \Omega_{Pa_X}$, $P(x|\pi_X)$ takes the value 1 if $f_X(\pi_X) = x$ and 0 otherwise. For instance, Figure A.15 shows the SEs from the running example represented as CPTs.

$$P(M|W) = \begin{array}{c} \begin{matrix} w_1 & w_2 \end{matrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} M = yes \\ M = no \end{array} \quad P(I|M, V) = \begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} M = yes, I = yes \\ M = yes, I = no \\ M = no, I = yes \\ M = no, I = no \end{array}$$

$$P(A|M, I, U) = \begin{array}{c} \begin{matrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{l} M = yes, I = yes, A = yes \\ M = yes, I = yes, A = no \\ M = yes, I = no, A = yes \\ M = yes, I = no, A = no \\ M = no, I = yes, A = yes \\ M = no, I = yes, A = no \\ M = no, I = no, A = yes \\ M = no, I = no, A = no \end{array}$$

Figure A.15: SEs from Figure A.13 represented as CPTs. $P(M|W), P(I|M, V)$ and $P(A|M, I, U)$ represent the same information as SEs f_M, f_I and f_A, respectively.

Appendix B. Computing counterfactual queries in SCMs

While observational queries can be calculated directly in the original model, interventional and counterfactual queries require applying graphical

operations in the causal graph. Figure B.16 depicts the modified graphs for various queries in the motivational example. Interventional queries are calculated in the so-called *post-intervention model*, which is the result of applying a graphical operation involving the removal of incoming arcs into the intervened variable and the replacement of its SE with a constant function that always returns the intervened value. Denoting by $P_{\mathcal{G}_i}$ the probability calculated in a model with the causal graph $\mathcal{G}_i$, and taking into account for instance the case of $\mathcal{G}_1$, that depicts the post-interventional model after forcing $I = yes$, the corresponding interventional query is $P(A_{I=yes}) = P_{\mathcal{G}_1}(A|I = yes)$.

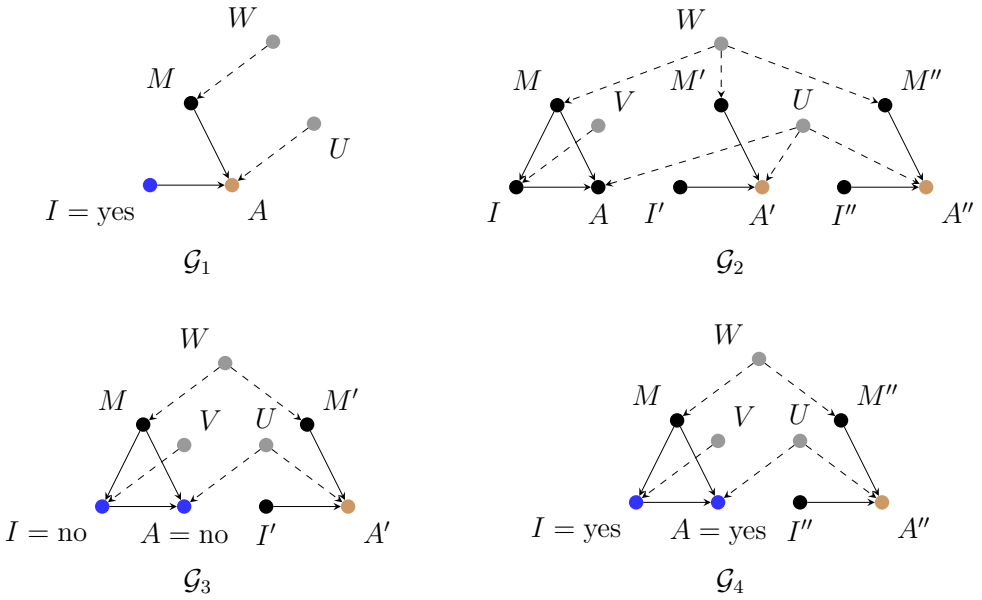

Figure B.16: Graphs of the post-interventional and counterfactual models for calculating various interventional and counterfactual queries in the SCM from Figure A.12 (left). Observed variables and target variables are shown in blue and yellow, respectively.

Counterfactual queries can be computed using an extended model called the *counterfactual model* (also known as the *twin model*). The twin model is an SCM that includes endogenous variables from both the real and hypothetical scenarios. This is achieved by duplicating the subgraph composed of the endogenous nodes for the real scenario and then applying the intervention.

In the counterfactual model, the endogenous nodes in both the real and hypothetical scenarios share the same exogenous parents, except for the intervened variables. In Figure A.12, $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ are the graphs corresponding

to the twin models for various counterfactual queries in the running example.

In $\mathcal{G}_2, \mathcal{G}_3$ and $\mathcal{G}_4$, variables from the real scenario are M,I and A; variables from the hypothetical scenario in which immigration is always present are denoted with a single apostrophe, i.e. M',I' and A'; those from the hypothetical scenario in which immigration is never present are denoted with double apostrophe, i.e. M'',I'' and A''. With this in mind, the query $\text{PNS}(I, A)$ can be calculated as $P_{\mathcal{G}_2}(A' = yes, A'' = no)$; $\text{PS}(I, A)$ as $P_{\mathcal{G}_3}(A' = yes | I = no, A = no)$ and $\text{PN}(I, A)$ as $P_{\mathcal{G}_4}(A'' = no | I = yes, A = yes)$.

Appendix C. Complexity discussion

The complexity of an SCM increases compared to the original BN. Each variable gains an additional parent, the corresponding exogenous variable, which increases the maximum in-degree by one. For the learning process, the number of variables doubles: the original endogenous variables and the exogenous ones. However, only the exogenous variables are trained. During inference, the number of variables can triple in the worst case, as the twin graph must also be considered.

The main limitation of the method is related to the exponential growth of the cardinality of U, which is given by equation (A.1). For example, if all variables are binary ($|\Omega_X| = 2$), with 5 parents, the cardinality becomes $|\Omega_U| = 2^{2^5} = 2^{32}$, resulting in over 4000 million possible outcomes. This implies that, to store only the probability values with 64-bit numbers, 32 GB would be required. With 6 parents, this grows to $2^{2^6} = 2^{64}$, exceeding 18 quintillion outcomes. Such massive spaces make storing probability tables or computing exact inference infeasible.

References

Andam, K.S., Ferraro, P.J., Pfaff, A., Sanchez-Azofeifa, G.A., Robalino, J.A., 2008. Measuring the effectiveness of protected area networks in reducing deforestation. Proceedings of the national academy of sciences 105, 16089–16094.

Anderies, J.M., Janssen, M.A., Ostrom, E., 2004. A framework to analyze the robustness of social-ecological systems from an institutional perspective. Ecology and Society 9, 18. URL: `http://www.ecologyandsociety.org/vol9/iss1/art18/`.

Arif, S., MacNeil, M.A., 2022. Utilizing causal diagrams across quasi-experimental approaches. Ecosphere 13, e4009.

Arnaiz-Schmitz, C., Aguilera, P.A., Ropero, R.F., Schmitz, M.F., 2023. Detecting social-ecological resilience thresholds of cultural landscapes along an urban–rural gradient: a methodological approach based on Bayesian networks. Landscape Ecology 38, 3589–3604. doi:10.1007/s10980-023-01732-9.

Avin, C., Shpitser, I., Pearl, J., 2005. Identifiability of path-specific effects, in: IJCAI International Joint Conference on Artificial Intelligence, pp. 357–363.

Aznar-Sánchez, J.A., Galdeano-Gómez, E., Pérez-Mesa, J.C., 2011. Intensive horticulture in Almería (Spain): A counterpoint to current European rural policy strategies. Journal of Agrarian Change 11, 241–261.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10, e0130140.

Bareinboim, E., Correa, J.D., Ibeling, D., Icard, T., 2022. On Pearl's hierarchy and the foundations of causal inference, in: Probabilistic and causal inference: the works of judea pearl. ACM, pp. 507–556.

Berkes, F., Colding, J., Folke, C., 2003. Navigating social-ecological systems: Building resilience for complexity and change. Cambridge University Press, Cambridge, UK.

Biggs, R., Rhode, C., Archibald, S., Kunene, L.M., Mutanga, S.S., Nkuna, N., Ocholla, P.O., Phadima, L.J., 2015. Strategies for managing complex social-ecological systems in the face of uncertainty: examples from South Africa and beyond. Ecology and Society 20, 52. doi:10.5751/ES-07380-200152.

Byrnes, J.E.K., Dee, L.E., 2025. Causal inference with observational data and unobserved confounding variables. Ecology Letters 28, e70023. E70023 ELE-00244-2024.R3.

Cabañas, R., Antonucci, A., Huber, D., Zaffalon, M., 2020. CREDICI: a Java library for causal inference by credal networks, in: International Conference on Probabilistic Graphical Models, PMLR. pp. 597–600.

Carriger, J.F., Barron, M.G., Newman, M.C., 2016. Bayesian networks improve causal environmental assessments for evidence-based policy. Environmental Science & Technology 50, 13195–13205.

Chen, H., Lundberg, S.M., Lee, S.I., 2022. Explaining a series of models by propagating shapley values. Nature communications 13, 4512.

Chen, S., Yao, S., 2023. Identifying the drivers of land expansion and evaluating multi-scenario simulation of land use: A case study of mashan county, china. Ecological Informatics 77, 102201.

Correa, J., Lee, S., Bareinboim, E., 2021. Nested counterfactual identification from arbitrary surrogate experiments. Advances in Neural Information Processing Systems 34, 6856–6867.

De Aranzabal, I., Schmitz, M.F., Aguilera, P., Pineda, F.D., 2008. Modelling of landscape changes derived from the dynamics of socio-ecological systems: A case of study in a semiarid Mediterranean landscape. Ecological Indicators 8, 672–685.

Ehrlich, D., Melchiorri, M., Capitani, C., 2021. Population trends and urbanisation in mountain ranges of the world. Land 10, 255.

European Environment Agency, 2007. Europe's environment: the fourth assessment. Reports, Copenhagen: EEA.

Ferraro, P.J., 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation 2009, 75–84.

Galdeano-Gómez, E., Aznar-Sánchez, J.A., Pérez-Mesa, J.C., 2013. Sustainability dimensions related to agricultural-based development: the experience of 50 years of intensive farming in Almería (Spain). International Journal of Agricultural Sustainability 11, 125–143.

Gratzer, G., Keeton, W.S., 2017. Mountain forests and sustainable development: The potential for achieving the United Nations' 2030 Agenda. Mountain Research and Development 37, 246–253.

Hatami, R., 2018a. Development of a protocol for environmental impact studies using causal modelling. Water Research 138, 206–223.

Hatami, R., 2018b. A practical method to control spatiotemporal confounding in environmental impact studies. MethodsX 5, 710–716.

Huang, Y., Valtorta, M., 2006. Identifiability in causal Bayesian networks: A sound and complete algorithm, in: Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. pp. 1149–1154.

Hughes, T.P., Carpenter, S.R., Rockström, J., Scheffer, M., Walker, B., 2013. Multiscale regime shifts and planetary boundaries. Trends in Ecology & Evolution 28, 389–395. doi:10.1016/J.TREE.2013.05.019.

Irvine, K.M., Miller, S.W., Al-Chokhachy, R.K., Archer, E.K., Roper, B.B., Kershner, J.L., 2015. Empirical evaluation of the conceptual model underpinning a regional aquatic long-term monitoring program using causal modelling. Ecological Indicators 50, 8–23.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press, Cambridge, Massachusetts.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Lutz, W., 2006. Fertility rates and future population trends: will Europe's birth rate recover or continue to decline? International Journal of Andrology 29, 25–33.

Maldonado, A.D., Aguilera, P.A., Salmerón, A., Nicholson, A.E., 2018. Probabilistic modeling of the relationship between socioeconomy and ecosystem services in cultural landscapes. Ecosystems Services 33, 146–164.

Mathias, J., Anderies, J., Baggio, J., Hodbod, J., Huet, S., Janssen, M., Milkoreit, M., Schoon, M., 2020. Exploring non-linear transition pathways in social-ecological systems. Scientific Reports 10, 4136. doi:10.1038/s41598-020-59713-w.

McConnachie, M.M., van Wilgen, B.W., Ferraro, P.J., Forsyth, A.T., Richardson, D.M., Gaertner, M., Cowling, R.M., 2016. Using counterfactuals to evaluate the cost-effectiveness of controlling biological invasions. Ecological Applications 26, 475–483.

Mendoza-Fernández, A.J., Peña-Fernández, A., Molina, L., Aguilera, P.A., 2021. The role of technology in greenhouse agriculture: Towards a sustainable intensification in Campo de Dalías (Almería, Spain). Agronomy 11, 101.

Molero, J., Marfil, J.M., 2017. Betic and southwest Andalusia. The Vegetation of the Iberian Peninsula: Volume 2 , 143–247.

Muñoz-Rojas, M., De la Rosa, D., Zavala, L., Jordán, A., Anaya-Romero, M., 2011. Changes in land cover and vegetation carbon stocks in Andalusia, Southern Spain (1956–2007). Science of the Total Environment 409, 2796–2806.

Parrott, L., Quinn, N., 2016. A complex systems approach for multiobjective water quality regulation on managed wetland landscapes. Ecosphere 7, e01363.

Parsons, C.A., Smeeding, T.M., 2006. Immigration and the Transformation of Europe. Cambridge University Press.

Paul, W.L., Anderson, M.J., 2013. Causal modeling with multivariate species data. Journal of Experimental Marine Biology and Ecology 448, 72–84.

Paul, W.L., Rokahr, P.A., Webb, J.M., Rees, G.N., Clune, T.S., 2016. Causal modelling applied to the risk assessment of a wastewater discharge. Environmental Monitoring and Assessment 188, 1–20.

Pearl, J., 1988. Probabilistic reasoning in intelligent systems. Morgan-Kaufmann, San Mateo, CA.

Pearl, J., 2009. Causality. Models, inference and reasoning. Second edition. Cambridge University Press, New York.

Pearl, J., Mackenzie, D., 2018. The book of why. Penguin Random House, UK.

Plieninger, T., Flinzberger, L., Hetman, M., Horstmannshoff, I., Reinhard-Kolempas, M., Topp, E., Moreno, G., Huntsinger, L., 2021. Dehesas as high nature value farming systems: a social-ecological synthesis of drivers, pressures, state, impacts, and responses. Ecology and Society 26, 23.

Plieninger, T., Wilbrand, C., 2001. Land use, biodiversity conservation, and rural development in the dehesas of Cuatro Lugares, Spain. Agroforestry Systems 51, 23–34.

Poston Jr, D.L., Bouvier, L.F., 2010. Population and society: An introduction to demography. Cambridge University Press.

Preise, R., Biggs, R., Vos, A.D., Folke, C., 2018. Social-ecological systems as complex adaptive systems: organizing principles for advancing research methods and approaches. Ecology and Society 23, 46. doi:10.5751/ES-10558-230446.

Punzo, G., Castellano, R., Bruno, E., 2022. Exploring land use determinants in Italian municipalities: comparison of spatial econometric models. Environmental and Ecological Statistics 29, 727–753. doi:10.1007/s10651-022-00541-8.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Ropero, R.F., Aguilera, P.A., Fernández, A., Rumí, R., 2014. Regression using hybrid bayesian networks: Modelling landscape–socioeconomy relationships. Environmental modelling & software 57, 127–137.

Scheffer, M., Carpenter, S., Lenton, T., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I.A., Levin, S., van Nes, E., Pascual, M., Vandermeer, J., 2012. Anticipating critical transitions. Science 338, 344–348. doi:10.1126/science.1225244.

Schmitz, M.F., De Aranzabal, I., Aguilera, P.A., Rescia, A., Pineda, F.D., 2003. Relationship between landscape typology and socioeconomic structure: Scenarios of change in Spanish cultural landscapes. Ecological Modelling 168, 343–356.

Siegel, K., Dee, L.E., 2025. Foundations and future directions for causal inference in ecological research. Ecology Letters 28, e70053. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.70053, doi:https://doi.org/10.1111/ele.70053, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.70053. e70053 ELE-00830-2024.R2.

Snethlage, M.A., Geschke, J., Ranipeta, A., Jetz, W., Yoccoz, N.G., Körner, C., Spehn, E.M., Fischer, M., Urbach, D., 2022. A hierarchical inventory of the world's mountains for global comparative mountain science. Scientific Data 9, 149.

Thornton, J.M., Snethlage, M.A., Sayre, R., Urbach, D.R., Viviroli, D., Ehrlich, D., Muccione, V., Wester, P., Insarov, G., Adler, C., 2022. Human populations in the world's mountains: Spatio-temporal patterns and potential controls. Plos One 17, e0271466.

Tian, J., Pearl, J., 2000. Probabilities of causation: Bounds and identification. Annals of Mathematics and Artificial Intelligence 28, 287–313.

Viñuela, A., 2022. Immigrants' spatial concentration: Region or locality attractiveness? Population, Space and Place 28, e2530.

Viñuela, A., Gutiérrez Posada, D., Rubiera Morollón, F., 2019. Determinants of immigrants' concentration at local level in Spain: Why size and position still matter. Population, Space and Place 25, e2247.

Wolosin, R.T., 2008. El Milagro de Almeria, España: a political ecology of landscape change and greenhouse agriculture. Ph.D. thesis.

Wu, J., Luo, J., Zhang, H., Qin, S., Yu, M., 2022. Projections of land use change and habitat quality assessment by coupling climate change and development patterns. Science of the Total Environment 847, 157491.

Wu, Y., Zhang, L., Wu, X., 2019. Counterfactual fairness: Unidentification, bound and algorithm, in: Proceedings of the twenty-eighth International Joint Conference on Artificial Intelligence.

Zaffalon, M., Antonucci, A., Cabañas, R., 2020. Structural causal models are (solvable by) credal networks, in: International Conference on Probabilistic Graphical Models, PMLR. pp. 581–592.

Zaffalon, M., Antonucci, A., Cabañas, R., Huber, D., 2023. Approximating counterfactual bounds while fusing observational, biased and randomised data sources. International Journal of Approximate Reasoning 162, 109023.

Zaffalon, M., Antonucci, A., Cabañas, R., Huber, D., Azzimonti, D., 2024. Efficient computation of counterfactual bounds. International Journal of Approximate Reasoning , 109111.

Zhai, H., Lv, C., Liu, W., Yang, C., Fan, D., Wang, Z., Guan, Q., 2021. Understanding spatio-temporal patterns of land use/land cover change under urbanization in wuhan, china, 2000–2019. Remote Sensing 13, 3331.

Zhang, J., Tian, J., Bareinboim, E., 2022. Partial counterfactual identification from observational and experimental data, in: International Conference on Machine Learning, PMLR. pp. 26548–26558.

Zhou, Y., Chen, T., Feng, Z., Wu, K., 2022. Identifying the contradiction between the cultivated land fragmentation and the construction land expansion from the perspective of urban-rural differences. Ecological informatics 71, 101826.