STAT34-Dissertation

UCL Student ID: 20150879

September 2021
Wordcount: 12,544

Abstract

This dissertation aims to provide an exhaustive introduction to causal inference and spatio-temporal causality and propose extensions to the “latent Space Causal Model” (LSCM) model proposed by Christiansen et al. (2020) to include spillover (or treatment interference across the space) and time lagged effects.

Causal inference is first introduced through Pearl’s *ladder of causality* and explains the use causal graphs and probability distributions towards estimating the causal average effect of a treatment from the observational data, as opposed to experiments.

Second, it reviews the Latent Structural Causal Model (LSCM), a novel framework proposed by Christiansen et al.(2020) that addresses the specific characteristics of spatio-temporal problems, when incomplete causal knowledge impedes their estimation of the true underlying effects. This work introduces the reader to the intuition and theory at the core of the LSCM framework. We propose, as an extension, eight new estimators that account for the causal effect of lagged treatments and that of the potential spatial interference of treatments on the response variable.

These estimators were tested using the original data set studied by Christiansen et al.(2020), to provide a second assessment of the causal lagged and spatial interference effects of the conflict on deforestation during the Colombia-FARC’s conflict. While the extensions presented here did not lead to conclude that there is indeed a causal relation between these two phenomenon, the results are consistent with the findings in Christiansen et al.(2020). Still, this approach should now be a more comprehensive modelling which widens the applications of the LSCM framework to general research designs seeking to establish and quantify causality in spatio-temporal data considering both, spillovers and/or time series effects.

Contents

1 Introduction

In most introductory courses to statistics and machine learning we are taught that correlation alone does not necessarily imply causation. However, students are often left to wonder why that is. In fact, we are given a plethora of examples which illustrate this claim by *reductio ad absurdum*[1], such as the unexpected correlation between chocolate consumption and the relative number of nobel laureates per inhabitant [10]. However how do we answer the big question: could consuming more chocolate actually increase our likelihood to win a nobel prize?

The route to formally establishing causality is rarely explained in our undergraduate courses. Despite being able to control everything of interest to the best of our knowledge and/or ability throughout advanced research designs, any modelling is limited by potential unobservable variables that could be responsible for the apparent relationship that we find empirically, which impedes our claim to causality. In particular, a common hidden factor driving both explanatory and outcome variables at the same time is a concern for all researchers, as discussed by Reichenbach's common cause principle (1956)[21]. It states that, when two variables X and Y are correlated, if neither causes changes in the other, then there must exist a third variable Z coming into play and driving them both[2]. The next section will discuss how we can determine when causality exists.

When it comes to building predictive models, correlation is often simply good enough to achieve highly reliable and effective predictions, especially with an increasingly abundant source of data[3]. Still, some situations are so peculiar and unprecedented that having access to causal knowledge is essential for us to identify the true underlying mechanisms of the phenomenon we observe. This is extremely relevant now to understand and predict the impact of the current COVID-19 pandemic on the different aspects of our life during and after this health crisis, where past data is not a good predictor of the future anymore, as we are witnessing an unprecedented event[4]. Such research questions might encompass the effect on a specific treatment on a aged patient or the impact of the Federal Reserve raising US interest rates by 1% on the global economy.

While the literature in social science is plentiful of controlled experiments which explore said issues, it is not always possible for the researcher to design a study which fully answers these, for ethical or logistics reasons. Think of the societal cost to force patients to smoke for example. In these cases the researcher can only infer

[1] i.e. argument to absurdity

[2] "If two random variables X and Y are statistically dependent ($X \not\perp Y$), then either (a) X causes Y, (b) Y causes X, or (c) there exists a third variable Z that causes both X and Y. Further, X and Y become independent given Z, i.e., $X \perp Y|Z''$. [21]

[3] high dimensional data: easily gathered, stored and accessed; find source

[4] speak of non-stationary data not relevant to analyze abrupt changes

causation under a set of assumptions applied to observational data in two ways. The first way is the bottom-up approach suggesting that causal relation structure can be derived, i.e learnt, directly from the data, which is known as *causal discovery* [18]. The second way is positing models based on previously established theory and then testing them in practice. This report will focus solely on the first way, applied to the field of spatio-temporal datasets and the inclusion of temporally-lagged and treatment interaction effects. Causality for spatio-temporal data is a fast growing research area [20]. Still as past data ceases to be valid or even relevant, this complicates the task of understanding already complex correlation structures in spatio-temporal datasets even more. While there are models capable of establishing accurate spatio-temporal predictions, they mainly rely on past data and are not meant to address questions such as *what would happen if something exterior to the model, or anything else changes?*.

This works relies on the foundation paper by Christiansen et al. 2020, "Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia". Throughout this work, the authors propose the Latent Structural Causal Model (LSCM), which is causal framework that address the specific characteristics of spatio-temporal problems, when incomplete causal knowledge impedes their estimation of the true underlying effects [3]. And applied this framework to estimate the real effect of the FARC conflict on deforestation in Colombia.

The main contributions of this report are twofold. First this work provides an exhaustive introduction to causal inference and spatio-temporal causality and second it proposes extension to the LSCM model proposed by Christiansen et al. (2020) to include spillover (or treatment interference across the space) and time lagged effects.

This dissertation is structured in the following five chapters. Chapter 2 summarises the ladder of causality proposed by Pearl and Mackenzie (2020) and introduces several methods presented in the existing literature which do consider spillovers. Next chapter 3 provides a summary and analysis of the main outcomes of the foundation paper "Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia". Based on the data from this paper, chapter 4 presents extensions to include spillovers and lagged effects and results of their implementation are discussed in chapter 5. Finally, chapter 6 concludes and proposes avenues for future research work.

2 Literature Review

This section introduces the basics of causality inference and the associated methods to establish causality in spatio temporal datasets. It then presents the main theoretical concepts required for the understanding of the foundation paper of this dissertaion Christiansen et al. (2020, that will be described in further details in chapter 3) of this report.

2.1 Causal Inference

Pearl and Mackenzie (2020) argues that the main the reason why causality has become increasingly studied only recently, i.e. throughout the late decade, is due to the relatively new development of a modern language that is able to express causality in a scientific way, called the *calculus of causation*. The formal calculus of causation is formed by two languages: (i) the causal diagrams representing the structure of what we know, i.e. a symbolic language and its associated semantics, and (ii) algebra to express what we are looking to know [17]. In this chapter we will introduce the reader to this calculus, by summarizing the paper by Dablander (2020), which explains the graphical approach to causation as presented in Pearl's *ladder of causality* [17].

The ladder of causality proposes three levels of causal inference associated to our cognitive abilities. The most basic, *association*, corresponds to detecting regularities in our environment, so it is limited to *see* some variables that are statistically related. The next level, *intervention*, is about predicting (at the *population* level) the effect of making changes to the environment, such as forcing every patient to take a treatment. Finally, the highest level, *counterfactual*, entails to question what should have happened if we had done something different. This is related with *imagining* and, together with a set of relevant assumptions, will allow us to answer questions at an *individual*-level.

This section is formed by five subsections, each one of the first three is dedicated to the explaining one step of the ladder of causality: Association, Intervention and Imagining. Subsection 2.1.4 introduces the problems of working with hidden variables and explains the extent to which instrumental variables can help deal with endogenous covariates. Finally, in section 2.1.5, the learning previously introduced are applied to analyse an example of Simpson's paradox, where hidden variables might be obscuring the results.

2.1.1 Association

The idea of association is to see if two or more things are related. Here it is important to distinguish between (i) the *marginal association*, which describes the relation of two variables X and Y without looking at others variables, noted $Y \perp X$ (respectively $Y \not\perp X$), which means the response variable Y and observed variable X are dependent (respectively independent); and (ii) *conditional associations*, which take into account other variables, e.g. $Y \perp X|Z$, which is interpreted as Y and X are independent given a third variable Z.

The aforementioned associations between variables can be easily visualized using graphs. Graphs are mathematical objects formed by nodes, that represent variables, and edges, that represent the relationship between them. Directed Acyclic Graphs (DAG) are one of the most commonly used graphical representations. Here the direction of the edges, represented as arrows, indicates which variable causes the other, and in the case of DAGs, these arrows cannot form a cycle (hence the name "Acyclic"). Figure 1 shows three DAGs depicting different relationships between the variabes X, Y and Z. For instance the left one is read as X causes Z and Z causes Y, which also means that both, X and Z, are the ancestors of Y (while only X is the ancestor of Z) and that Y is the descendant of Z (and both Y and Z are descendants of X). Interestingly enough, all three DAGs shown in the figure entail the same conditional independence $Y \perp X|Z$, which illustrate the fact that it is not possible to identify the true causal graph from only seeing the observational distribution. To address this issue, it is necessary to impose a stricter and more suitable restriction to the models. Several methods have been proposed to do so. Among the main ones are constraint-based methods, score-based methods, methods based on restricted Structured Causal Models, and methods based on the independence of causal mechanisms [2]. A good summary of these can be found in the paper by Guo et al.(2020). In this report we will focus solely on the constraint-based methods and throughout this chapter we will introduce the reader to the relevant concepts in order to consolidate a comprehensive introduction to causal discovery.

First, in order to interpret these graphs as causal we need to introduce two critical assumptions: the Markov property and the assumption of faithfulness. These ensure that graphical relations found for graph $\mathcal{G}$ over the variables X implies specific properties concerning a generic distribution P over a random vector X and vice-versa. Formally this is defined as:

Definition 2.1 (Markov and Faithfulness properties)**.** Definition. P is said to satisfy the Markov property and Faithfulness, with respect to $\mathcal{G}$, if for all disjoints sets $A, B, C \subseteq X$, it holds that:

- $A \perp_{\mathcal{G}} B|C \Rightarrow A \perp_{P} B|C$. (Markov)
- $A \perp_{P} B|C \Rightarrow A \perp_{\mathcal{G}} B|C$. (Faithfulness)

However, these two properties do not have to be assumed simultaneously. In fact, if our intention is to infer independence constraints implied by a given graphical structure $\mathcal{G}$ then only the Markov property is necessary. Conversely, if our intention is to infer the graph structure form the data, only Faithfulness is required. This work focuses on the former.

Second, once a theoretical base for causality with graphs has been posited, and by assuming the Markov property holds, it is then possible to state the concepts of DAG analysis:

- A **path** from X to Y is a sequence of nodes and edges such that the start and end nodes are X and Y, respectively.

- A **conditioning set** $\mathcal{L}$ is the set of nodes we condition on (note that the set can be empty, implying independence).

- Conditioning on nodes that are not descendent of both X and Y along a path from X to Y will **block** that path, and given that the Markov property holds, these variables will be considered **independent** probabilistically.

- Some nodes are called *colliders* in a path. A collider is a node that has two or more edges pointing to it. Colliders have the opposite effect than other nodes as they block a path by not conditioning on them. Still, when a conditional association is forced on the colliders, the path becomes unblocked and the variables separated are now dependent [5].

As a consequence of this formulation, we define the d-separated tool: "we call two nodes X and Y d-separated by $\mathcal{L}$ if conditioning on all members in $\mathcal{L}$ blocks all paths between the two nodes"[5]. Together with the Markov property, it implies that $X \perp Y|L$ holds in the probability distribution. The d-separation is a powerful tool that can be used to visualize and calculate conditional independence and will be further explained in the next subsection.

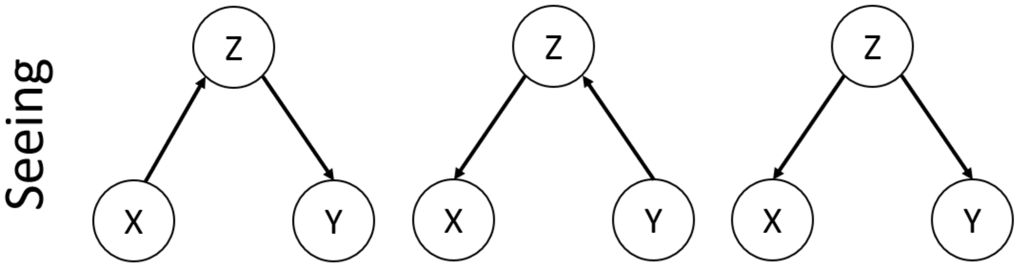


Figure 1: DAGs Examples. These three DAGs encodes the same conditional independence structure $X \perp Y|Z$ [5]. In the image A is the treatment, Y is the outcome, and X is the observed confounder. (source: Dablander, 2020)

2.1.2 Intervention

With the background information from the section Association, DAGs have now causal meaning, and we can interpret the arrows as direct causal effects between two variables. But what does it mean to be causal? There are many definitions and it has been the source of many debates across centuries. Hume in 1784 [6] was one of, if not the first, to propose a definition that is widely in use today: "We may define a cause to be an object, followed by another, [...] where, if the first object had not been, the second had never existed". Pearl and others, work form this definition and proposed to

[5] In this work we will not discuss this type further, but if the reader wants to learn more about it, we encourage him to read the work of Dablander (2020).

take an “interventionist position” to prove causality by saying that a variable X has a causal influence on Y if changing X leads to changes in (the distribution of) Y. The do-operator was then defined as a tool to impose these changes on the observational data.

Formally the do-operator is noted as $p(Y|do(X = x))$. It describes the values that Y is likely to take when the interventionist sets X to be x [5]. Comparatively, the conditional distribution $p(Y|X = x)$ describes the values that Y is likely to take when X happens to be x. The latter corresponds to ‘seeing‘ the data and the former to a ‘do‘ intervention. Figure 2 shows the effect of the intervening on the variable X for the three DAGs that encodes the distribution of seeing $X \perp Y|Z$. In effect, the do-operator means that we are intervening the experiment and forcing a specific variable to take an specific value, breaking all connections (DAG’s edges) of this variable with its ancestor. Therefore, all the direct descendant variables of the intervened variable will change its distribution, but its ancestors would not be disturbed.

We now have a notion of what an intervention is. Thus, in order to describe the causal effect and find the true DAG, we need a way to compute the do-operation. One way would be to actually design an investigation, define and allocate treatments and then performing the actual investigation [6]. Nevertheless implementing the intervention might be expensive, slow and even unfeasible or unethical (such as forcing someone to smoke intensely). Another way is to learn the causality structure directly form the observed data. To do so we need to consider additional assumptions, which the researcher must assess how likely they are to hold true in each project/case, that allow to link the observational DAG and the manipulated DAG[7]:

- A1: Interventions are *local*: when we set a variable to a value, we do not affect other variables, we are like surgeons with a bistoury, only cutting what we want to affect.

- A2: *Autonomy*: the mechanism through the variables interact are not affected (i.e. do not change) due to the interaction.

Together with these two assumptions and the Markov property, the do-operator for the original DAG becomes the marginal observed distribution of the manipulated DAG, then $P(Y|do(X = x)) = P_m(Y|X = x)$ where m denotes the intervened DAG. Looking again at the figure 2, the left intervened DAG remains the same as the observational DAG, therefore $P(Y = y|do(X = x)) = p_m(Y = y|X = x) = P(Y = y|X = x)$. In the other two cases, X is detached from the rest of the DAG and so $X \perp Z$, therefore we have the following equation development:

[6] For more on Design of investigations, we recommend the course notes of course STAT029 UCL [13]

[7] also called “DAG intervened”, i.e. cutting all the causal arrows that point to the variable intervened

$$
\begin{aligned}
P(Y=y|do(X=x)) &= p_m(Y=y|X=x) && (1)\\
&= \sum_z p_m(Y=y, Z=z|X=x) && (2)\\
&= \sum_z p_m(Y=y|X=x, Z=z)p_m(Z=z|X=x) && (3)\\
&= \sum_z p(Y=y|X=x, Z=z)p(Z=z) && (4)
\end{aligned}
$$

Where line (1) comes from the definition of the do operation, line (2) from total probabilities, line (3) from the product rule of probability. Finally, line (4) comes from the assumption of autonomy, that the intervention does not affect the mechanism that makes the other variables change so $P_m(Y=y|X=x, Z=z) = P(Y=y|X=x, Z=z)$ and that the intervention are local so $X \perp Z$ and $p_m(Z=z) = p(Z=z|X=x) = P(Z=z)$.

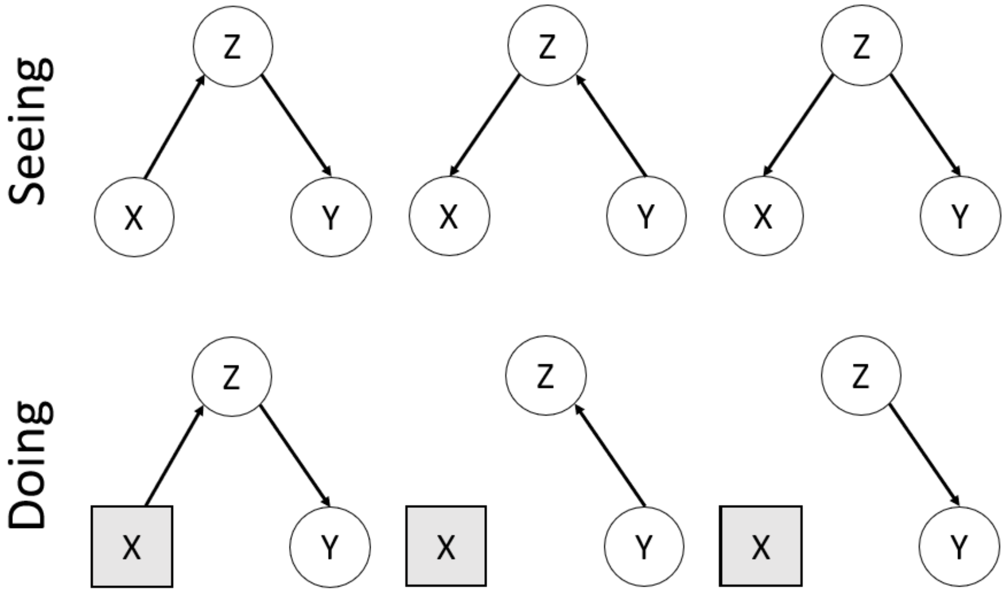


Figure 2: Observational DAG vs manipulated DAG though intervention on X. (source: Dablander, 2020)

Now, let us consider the case of confounding variables. Figure 3 shows a three-variable DAG where $X \not\perp Y|Z$, and in this case we are facing confounding variables. Confounding can be defined as "the situation where a (possibly unobserved) common cause obscures the causal relationship between two or more variables" [2]. Formally, this can be defined as $P(Y|X=x) \neq P(Y|do(X=x))$. Therefore there are also confoundings in the right and middle DAGs of Figure 2. In these cases, a tool that is used to see the effect of the X on Y is called the "Backdoor Adjustment". Pearl [14] defined it as "given two nodes X and Y, an adjustment set L fulfils the backdoor criterion if: (i) no member in L is a descendant of X and (ii) L block all paths between X and Y that contain an arrow into X (they enter X 'through the backdoor'). Adjusting for L thus results in the unconfounded causal effect of X on Y". Formally for all x, y, it holds that : $p_m(y|x) = \int_{z \in L} P(y|x, z)P(z)\ dz$ [2] as proven before. In summary, (i) this tool allows us to compute the intervention model's distribution from the observational distribution of (X, Y, Z), (ii) we are not required to know all the variables, just the one that block the path, and (iii) if the backdoor criterion is satisfied by conditioning on L, the causal effect of X over Y is now unconfounded.

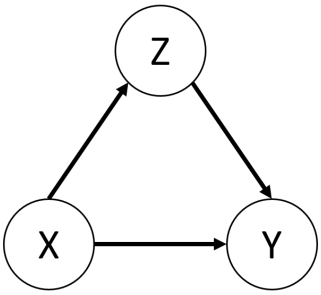


Figure 3: confounded DAG. (source: Dablander, 2020)

2.1.3 Counterfactual

The final level of the ladder of causation is *imagining*. Here we look for the counterfactuals, which are *what would have happened if the another treatment was used instead?*. This is a normal way we humans think and learn, i.e. by a retrospective analysis of the situations we live through. But for years this has been the "fundamental problem of causal inference", since we simply cannot observe counterfactuals as we forfait their very existence once we move forward with the treatment implementaion. To solve this, we need a way to estimate, find elsewhere, or even construct, the counterfactuals for our interventions. Many ways have been proposed such as matching and propensity scores. In this report we show how to compute the Average Causal Effect (ACE) using the do-operator. The ACE is formally defined for a binary variable $X \in \{0, 1\}$ over Y as:

$$ACE(X \to Y) = E[Y|do(X = 1)] - E[Y|do(X = 0)]$$

Nevertheless, ACE estimates the causal effect of the treatment over the population, not for an individual. To estimate the effect for a specific individual, one way is by using a Statistical Causal Model (SCM) which describes the mechanism processes using equations. Formally the SCM is defined as:

Definition 2.2 (Structural causal model(SCM))**.** A structural causal model (SCM) over variables $X_1, ..., X_p$ is a pair $M = (\mathcal{S}, \mathcal{Q})$ consisting of:

- a family $\mathcal{S}$ of structural assignments

$$X_j := f_j(PA_j, \epsilon_j), j = 1, ..., p\epsilon_j \sim \mathcal{Q}$$

SCMs are very similar to Structural Equation Models, but in them the causal assignment is now made explicit by using the symbol ':='. In SCMs, every equation (or structural assignment) describes the relation of a node with its parent, specifying the causal structure that are shown in a graph ($PA_j \rightarrow X_j$) and a function f_j that quantifies the causal effect of each parent on an specific node X_j. Assuming that we are able to find the SCM, we can calculate the ACE for the data-population by applying the do-operator directly on an SCM equation, or we can calculate the expected effect for a specific individual called the Individual Causal Effect (ICE). The ICE is formally defined as:

$$ICE(X \rightarrow Y) = Y_1(X = x) - Y_0(X = x')$$

It is worth noticing that behind the ICE, there is a prediction and therefore it has a fundamental error term associated to it since we cannot observe counterfactuals as we only observe the effect of taking one treatment at a time, so the counterfactuals must be estimated somehow. For SCM, the counterfactual is calculated as particular expected values (just as the ACE) using the information available from the rest of observations. There are alternatives methods such as matching. The main idea is to match observations between them using a set of criterion, e.g. a share of common traits define a similarity score, so observations that were treated differently become the counterfactual of its associated/matched pair. [8]

2.1.4 Hidden variables and Instrumental variables

Until now we have been assuming that our DAGs consider all the relevant variables. Nevertheless, it not always possible to ensure that all the relevant confounders variables are observed, so there might not be a valid observed set of nodes $\mathcal{L}$ that blocks all the different paths from Y to X (i.e. the 'backdoor' of X is not blocked). In these cases, we say that we are in presence of *hidden variables*, which constitutes a major threat to establishing causality as it is impossible to distinguish which variable is responsible for the causal relation. In fact, at least one of the regressors has become endogenous, i.e they are correlated with the error term, thus introducing bias in the causal estimators . We can see this mathematically using a linear regression of the structural model noted $y = \alpha + \beta x + \epsilon$,where α is the intercept, the slope β captures

[8] A recommended video about matching methods and a comprehensive explanation on causal inference in general can be found in the MIT open course ware webpage. [16].

the true effect, and ϵ is the error term. The ordinary least square estimator can be written as following:

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}XY = \frac{cov(X,Y)}{Var(X)} \\ &= \frac{cov(X,\alpha+\beta x)}{Var(X)} + \frac{cov(X,\epsilon)}{Var(X)} \\ &= \beta + \frac{cov(X,\epsilon)}{Var(X)}\end{aligned}$$

Where β would be the unbiased estimator, but since our regressor X is correlated with the error ϵ the OLS estimator $\hat{\beta}$ is biased by the now non-zero second term,leading to either an under or over-estimation of the true effect. [9]

Therefore, when have one or more hidden variables that are blocking one or multiples path $X \rightarrow Y$, we cannot correctly measure the causal effect of X on Y, and one solution is to use *Instrumental Variables(IV).* To illustrate the idea of IV, let us consider the following case when the analysis of the demand for cars Y is confounded with the price of the cars X, as per the microeconomics law of demand-and-offer, demand depends on the price and vice-versa. In this case we say that X and Y are endogenous. Then, to block all the backdoor paths, we will need to observe several variables that are not easily found in real world conditions such as, salaries, the state of the world economy, global and local inflation in countries throughout the supply chain, an so on. Another solution is to generate an IV that predict the car price X using a variable Z, that is exogenous to Y. This IV would then be unconfounded with the response variable Y and therefore it can safely be used to estimate the causal effect of X on Y as there is no needed node to be blocked to fulfill the backdoor criterion. Here, a good IV in our example case,would be the price of steel, as it only affects the demand for cars through the price of cars and it is a relevant predictor for the car's price as steel is one of the main materials used to build them.

In summary, in order for a variable Z to be a good IV, it must be: (i) relevant, which means that has to be related with X, i.e. $corr(Z,X) \neq 0$ and; (ii) Exogenous, so Z is related with Y . Together this gives the exclusion criteria that Z affects Y only through X. Graphically this means that the only open path form Z to Y is $Z \rightarrow X \rightarrow Y$, therefore $Z \perp Y|do(X)$.

A common method used to calculate the IV estimator is the two-stage least squares (2SLS) [24]. In the first stage of this method, the explanatory endogenous variables X is regressed on all of the exogenous variables Z, so we have an estimated X, noted $\hat{X}$, that is exogenous, in particular to the response variable Y in our example above. In the second stage Y is regressed on the $\hat{X}$ estimated in the previous stage.

[9] This is one of the reasons our lecturers on the statistics 101 courses stressed the issues of having "heteroscedasticity" in the error when fitting a linear regressions

The estimators of this last regression can then be interpreted as a causal effect. Mathematically the IV estimator, β_{IV}, can be calculated as following:

$$\beta_{IV} = (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top Y = (X^\top P_z^\top P_z X)^{-1} X^\top P_z^\top Y = (Z^\top X)^{-1} Z^\top Y$$

Where, P_z refers to the projection matrix of the first stage $P_z = Z(Z^\top Z)^{-1} Z^\top$ that specifies $\hat{X} = P_z X$.

It is worth noticing that good covariables are those that block the backdoor of X. Therefore they are not good as IV as they are not exogenous, and the opposite is also true as exogenous variables that are good IVs also happen to be bad covariables because they are only correlated to the response variable through X and their inclusion adds collinearity to the model. Nevertheless, it can be argued that no variable is completely exogenous of another, as depending on the modelling assumptions, we can find a mechanism through which variables could be related. For example, rainfall could be a good predictor of economic output from farming activities in hot and less developed countries, but only to the extent that climate change is not man-made. [10] With regards to our example on the demand for cars discussed above, it can be argued that the steel price is not exogenous from the demand for cars in a country where the main source of income is steel production, as it affects the mean earnings of the population, and so their purchasing power and their capacity to buy cars. In both cases the researcher must assess the extent to which this assumption holds depending on both the research objective and their scope of study.

2.1.5 Simpson Paradox

In this final subsection, the Simpson's paradox is presented as an illustration on how to causal modelling for data analyse, that can be summarized into first understand our causal graph and from there analyze it using the function of the data that the graph represents. The Simpson's paradox is a well studied analysis problem that is related with the presence of hidden variables. Let us consider the example of Table 1 taken from Dablander(2020). This case is (wrongly) called a paradox because we can see that for the two blood pressure levels, low and high, the treatment effect has a better recovery rate than in the absence of treatment, but when looked at the aggregate level ("Low & High Blood Pressure") the treatment's recovery rate is lower than the no-treatment's rates; how can this be? which one is correct? These questions have motivated many studies in statistics, and causal inference can provide light on how to analyse it.

In order to understand the causal graph, the important question is which variable is causing the other? Looking at figure 4, DAG (a) proposes the case when the variable Blood Pressure (BP) causes Treatment (T). This could be justified by a mechanism

[10] which the recent IPCC 2021 report disproves

such as a hospital policy that conditions doctors to give a specific treatment to patients that exhibit high blood pressure. Then, when estimating the $ACE = E[Y|do(T = 1)] - E[Y|do(T = 0)]$, it can be noticed that the interventions on T would erase the causal link $BP \rightarrow T$, so from the implications of the Markov property holding true, in the intervened case we can assume that $T \perp_{\mathcal{G}} BP \Rightarrow p(BP|T) = p(BP)$. Finally, using our do-operator calculus we can compute the distribution from the observed data as following:

$$\begin{aligned}
E[Y|do(T = t)] &= E[Y|T = t] \\
&= \sum_b E[Y|T = t, BP = b]p(BP = b|T = t) \\
&= \sum_b E[Y|T = t, BP = b]p(BP = b) \\
&= E[Y|T = t, BP = high]\ p(BP = high) \\
&\qquad + E[Y|T = t, BP = low]\ p(BP = low)
\end{aligned}$$

then,

$$\begin{aligned}
ACE =& E[Y|do(T = 1)] - E[Y|do(T = 0)] \\
=& 93\% * \frac{357}{700} + 73\% * \frac{343}{700} - (87\% * \frac{357}{700} + 69\% * \frac{343}{700}) \\
=& 83\% - 77\% = 5\%
\end{aligned}$$

So we can now conclude that the treatment has a positive average causal effect by a +5% in the recovery rate of patients.

On the other hand, considering the right DAG of the figure 4, it is proposed that the Treatment(T) causes Blood pressure (BP). This could be justified by a (potentially unexpected) body response to the treatment patients receive depending on their individual characteristics, such as stress and fatigue levels. In this case the intervention is not eliminating the causal effect $T \rightarrow BP$ according to the Markov property. The treatment and blood pressure are not independent due to the intervention itself. Thus when we calculate ACE using do-calculus we obtain the conditional distribution over the treatment:

$$\begin{aligned}
E[Y|do(T = t)] &= \sum_b E[Y|T = t, BP = b]\ p(BP = b|T = t) \\
&= E[Y|T = t]
\end{aligned}$$

Then,

$$\begin{aligned}
ACE &= E[Y|do(T = 1)] - E[Y|do(T = 0)] \\
&= 78\% - 83\% = -5\%
\end{aligned}$$

In this case the opposite conclusion is reached. The treatment is actually bad for the patient's health. These two examples shed light on the necessity to understand the true causal structure of our data to solve this conundrum. With this in mind, the researcher can define the appropriate research questions and their associated design to reveal the true direction of the edges of the DAG really by implementing targeted controlled experiments for instance. Once the causal DAG is correctly specified, the researcher can define the function of the data and it will estimate the wanted causal effect accordingly.

Table 1: Example Simpson Paradox (source: Dablander, 2020)

	Treatment	No Treatment
Low Blood pressure	81 out 87 recovered (93%)	234 out 270 recovered (87%)
High Blood pressure	192 out 263 recovered (73%)	55 out 80 recovered (69%)
TOTAL: Low & High	273 out 350 recovered (78%)	289 out 350 recovered (83%)

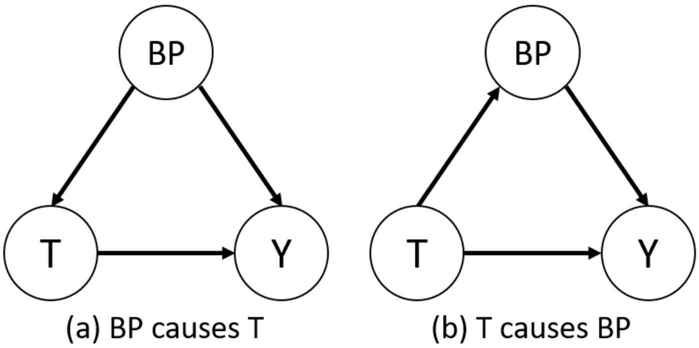


Figure 4: Proposed DAGs for Simpson paradox example (source: Own development from the work of Dablander, 2020)

2.2 Review of Spatial Causal Inference Methods

This subsection is based mainly in the work of Reich et al. (2021) [20], that reviewed different spatial causal methods and proposed that spatio-temporal method could be grouped in three types:

1. Methods to adjust for missing spatial confounding variables (unmeasured variables correlated with treatment and response variables), such as case control matching, spatial smoothing and propensity-score methods;

2. Method to adjust for spatial causal inference interference or spillovers, where the treatment of one location could then affect the outcome at another location; and

3. Causal methods based on the potential outcomes framework with Granger causality. Since we are looking to extend the model introduced in the subsection 3 to include time lags and spillovers we will introduce the reader to the second groupof methods.

In this section we will first introduce the implications for the model assumptions when we consider spillovers, then the next subsection will describe how to calculate summaries of the ACE. Finally subsection 2.2.3 is dedicated to introducing spatial structures used to reduce the interference patterns, thus rendering the computation of the ACE possible.

2.2.1 Spillovers

Spillovers, or interference, refer to the unwanted effect of a treatment done at certain unit (or location) on the outcome measured at another unit, possibly untreated too. Examples of interference can be found in many research areas such as: the global effects of vaccination campaign which reduce the probability of contracting the virus for both, vaccinated and unvaccinated persons.

Including spillovers in a causal model is a direct violation of the assumption of non-interaction between units that is part of Stable Unit Treatment Value Assumption (STUVA) [20], which is a fundamental element of mainstream causal approaches. For instance, a randomized controlled trial, i.e. an experiment which randomizes the treatment in a population to ensure it is exogenous, will make the distribution of the spillovers random too and probably biased, thus complicating the research design while spillovers are out of the researcher's control already [4]. For example, consider an experiment in which units exhibit different density zones, defined by a proximity criteria. In such a setup, unless a stratification strategy is correctly implemented, a randomized treatment assignment will, by design, attribute higher chances of being affected by spillovers to the units that are closer together, thus inducing biases to the results.

On the other side, non-interference is considered a very strong assumption [4], and the inclusion of spillovers replaces it with a slightly weaker one. Still, we need additional assumptions that define the reach of the interference to reduce the complexity of the assumed data generation process, otherwise the computation of the the average causal effect might be impossible. While their exclusion will lead to a biased causal effect estimator, their inclusion with yet a set wrong assumptions, such as the radius effect or aggregation function used to summarize the spillover, can also yield incorrect estimates[4]. Throughout this dissertation, we are considering the following two core assumptions in regards to spillovers (i) the proximity of the spillover is considered in geography (in contrast to commerce or internet relation among tiles) and (ii) group variables, later discussed in section 2.2.3

2.2.2 Potential Outcomes Summaries

This section presents some existing methods which adjust for spatial causal inference. To do so, we first explain the potential outcome framework and then several estimators for the causal effect are introduced. Starting with the potential outcome framework review, we first define the following key notation:

- let S be the set regions, and s or s_i the subindex that specify the location
- let M denote the observations (which differentiate on time) and t the subindex of that observation
- Y_s^t the response variable in location s and time t
- X_s^t the treatment variables in location s and time t
- $X_{s'}^t$ is the collection of treatments at time t in the regions s', in a neighbourhood to s
- $W_s^t = ((W_1)_s^t, ..., (W_p)_s^t)$ a set of potential confounding variables
- H_s unobserved time-invariant (or spatial) confounding variables in region s
- We introduce the potential outcome notation as $Y_s^t(X_s^t)$

The main goal is the estimation of the average treatment effect (ACE) which, as seen in section 2.1, can be estimated as

$$\mathbb{E}[Y|do(X = x_1)] - \mathbb{E}[Y|do(X = x_0)]$$

but now we will also consider the effect of spillovers. To do so, let us first introduce the potential outcome considering spillover notation as $Y_s^t(x_s^t, x_{s'}^t) = \mathbb{E}[Y_s^t|x_s^t, x_{s'}^t, W_s^t, H_s]$ where s' represents a neighborhood to s.

Hudgens and Halloran(2008) [8] described four estimates of the potential outcome assuming a binary treatment at individual level:

1. The direct effect (DE), which compares the potential outcomes difference of the local treatment holding the spillover treatments fixed:

$$DE(x_{s'}^t) = \mathbb{E}[Y_s^t(1, x_{s'}^t) - Y_s^t(0, x_{s'}^t)]$$

2. The spillover effect or interaction effect (IE), which measures the contribution of the treatment in other locations:

$$IE(x_{s'}^t, (x')_{s'}^t) = \mathbb{E}[Y_s^t(0, x_{s'}^t) - Y_s^t(0, (x')_{s'}^t)]$$

3. The total effect (TE), which is the sum of both effects:

$$\begin{aligned} TE(x^t_{s'}, (x')^t_{s'}) &= \mathbb{E}[Y^t_s(1, x^t_{s'}) - Y^t_s(0, (x')^t_{s'})] \\ &= \mathbb{E}[Y^t_s(1, x^t_{s'}) - Y^t_s(0, x^t_{s'})] + \mathbb{E}[]Y^t_s(0, x^t_{s'}) - Y^t_s(0, (x')^t_{s'})] \\ &= DE^t_s + IE^t_s \end{aligned}$$

4. The Overall effect (OE), which is similar to the Total effect but allowing the local treatment to be the same in both cases.

In a linear regression model, such as the one presented bellow, the direct and indirect effects are noted by β_1 and $\beta_2$respectively.

$$Y^t_s(x^t_s, x^t_{s'}) = x^t_s\beta_1 + x^t_{s'}\beta_2 + H_s\gamma + \epsilon^t_s$$

The previous effects are useful to estimate the implications at individuals locations. To asses the effect at the population level we then average the potential outcomes across the locations. To do so, the direct, spillover and total effect can be calculated using a policy-averaged effect. With a probability distribution for the potential actions of the neighbours defined as $(\psi(x') = p(X^t_{s'} = x'|X^t_s))$, the policy-averaged expected counterfactual outcome is given by:

$$\bar{Y}^t_s(x^t_s, \psi) = \sum_{x' \in X^t_{s'}} \mathbb{E}[Y^t_s(x^t_s, x')] \times \psi(x') \tag{5}$$

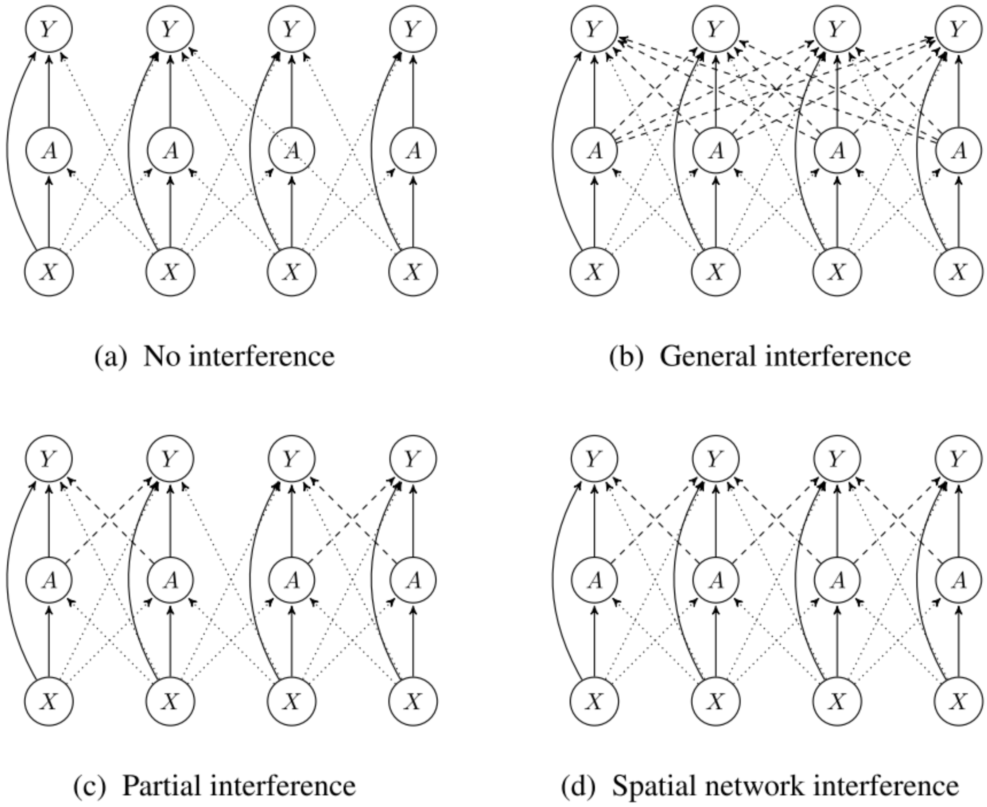
If we are in a position to assume independence between treatments we obtain that $\psi(x') = p(X^t_{s'} = x'|x^t_s) = p(X^t_{s'} = x')$. Then the policy-averaged direct effect at (s, t) can be computed as $DE(\psi) = \bar{Y}^t_s(1, \psi) - \bar{Y}^t_s(0, \psi)$.

2.2.3 Spatial Structures Interference Assumption

To compute the summary measures of the treatment effect, we need to define the neighbour that is going to be considered in the subscript s'. If we allow it to be the entire space, which is called “General Interference”, we will have to consider $2^{|S|}$ potential outcomes in a binary treatment framework, one for each possible combination. This can often render the computation of the average effects intractable. To address this icomplication, several modeling methods has been proposed to reduce the possible number of combinations by taking advantage of the spatial structure of each unit and proposing partial and network interference. Both interference are summarized and compared in Figure 5.

Partial inference was proposed by Sobel (2006) [7] and suggests that the units or locations can be grouped so the model only considers within-group interference. For example, the allocation of spillovers among municipalities of the same city is a possible application of this framework. In the literature, different group definitions have been used, in particular Perez-Heydrich et al. (2014) and Zigler and Papadogeorgou (2018) defined groups by spatial proximity while Zigler et al.(2012) grouped sites according to their attainment status.

Network interference modelling originates in extensions of existing methods developed specially for social network data, where models assume interference along their connected users. In the literature, several research streams have suggested a definition for these networks: Forastiere et al. (2016) allowed interference between an observation and its immediate neighbors, creating a local interference neighborhood around each observation; Aronow et al. (2017) defined an exposure mapping function; Tchetgen et al. (2017) examined interference subject only to a local Markov property that observations are conditionally independent after taking into account the nodes between them; and Giffin et al. (2020) use the distance between units themselves.



(a) No interference

(b) General interference

(c) Partial interference

(d) Spatial network interference

Figure 5: Variable dependencies under different forms of interference.
In the image A is the treatment, Y is the outcome, and X is the observed confounder, dashed lines represent the interference effect and solid lines the confounding relationships (source: Reich et al., 2021).

Finally, a common way to reduce the complexity of the model is to add assumption about the form of the interference so it becomes possible to aggregate the spillover effect in one value [20] and simplify the overall model. The general form of a regression model with this kind of interference would be specified as follow:

$$\begin{aligned} Y_s^t(x_s^t, x_{s'}^t) =& x_s^t\beta_1 + \phi x_{s'}^t\beta_2 + H_s\gamma + H_s + \epsilon_s^t \\ \phi(\{x_{s'}^t\}) =& \int_{k\in S} w(s,k)x_k^t dk \end{aligned}$$

So $w(s, k)$ can be used to estimate average, limit the distance or state any assumption of the spillovers effect. Additionally. the β_1 and β_2 has has the causal meaning of direct and spillover effect.

3 Foundation paper summary: "Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia"

3.1 Introduction

The causal approaches to learn from observational data (as opposed to randomized trials) shown in the previous subsection relies on assumptions to hold true in order to consider the relations between variables as causal relations. These assumptions are rarely easy to argue for real problems, and this is certainly true for spatio-temporal data, where there are complex dependency structures and the generation mechanism is hardly known, making mainstream causal inference theory and tools not directly applicable for this type of data [20]. Indeed, the theory and methods for causal inference for spatio-temporal data did not develop as fast as it did for data with independent generative processes ([1]; [12]). Reich et al. (2021) [20] identified three analytical challenges for spatio-temporal data that explain its late development: 1) that randomization is often infeasible due to logistic and/or ethical issues and so studies rely on observational data; 2) exposure and response variables exhibit spatial correlation, which complicates the modelling and computations as several variables would be required; and 3) those treatments might have spillovers or interference, where the treatments at one location may influence the outcomes of other locations.

In this subsection we will list the considered challenges and the assumptions proposed in the foundation paper (Chrisiansen et al.,2020 [3]) and introduce the proposed model: "Latent spatial confounder model" that formally encompasses these assumptions.

First, assuming that the data is generated from an i.i.d process is not direct, and in fact, i.d.d. data is not quite common [2]. In this way, it is more plausible to consider any two different observations as a realization from different distributions, and then attempt to formulate a justification for the i.i.d. assumption from a negligible difference between these distribution (yet it is still not i.i.d.). Nevertheless, we will make a more sensible assumption, which posits that the entire data set is a single outcome of some underlying joint distribution, This then allows us to acknowledge that the data is heterogeneous and/or exists on dependent measurements.

This last assumption will entail a second and equally essential assumption: *the homogeneity of the structural assignments of the variables*, which means that the functional dependence of each variable on its spatial neighbourhood remains the same across space, in other words the marginal distribution of the variables is the same for each tile, which is called being weakly stationary. This assumption is necessary in order to reduce the degrees of freedom so it is possible to compute the quantification of the causal relationships. For instance, without this condition to model the spatial relation between $|S|$ tiles using SCM with d processes, one would be required to specify $d|S|$ assignment, and each possible depended on $d|S| - 1$ variables [2]. Also, it is also very rare to bring support to the assumption that the data generating mechanism is fully specified [3], thus complicating the assumption for backdoor adjustment (or other adjustments). To solve this issue, the authors [3] proposed to consider hidden variables in their model specification. The main idea is that under certain assumptions, by taking the expectation over the hidden variable process, it is possible to compute the causal expectation of the treatment variables over the response without estimating these hidden variables directly. In order for this model to be consistent though, the authors propose a framework that restricts both, the causal structure and the type of hidden variables considered. This framework is introduced in subsection 3.2. Section 3.3 presents the authors' [3] procedure to estimate the causal effect from the data while section 3.4 describes the associated non-parametrical method for hypothesis testing.

3.2 Proposed framework for spatio-temporal causal analysis

First of all, in this framework the "confounders" are not independent variables, instead they have to be considered as spatio-temporal coordinate processes, where each has some spatial relationship with itself, other coordinates and/or time, and/or other processes even. These processes can be decomposed into disjoint 'bundles' of processes. In this framework "we are interested in specifying causal relations among these 'bundles' while leaving the causal structure among variables within each bundle unspecified". In this way the proposed model to predict the response process is similar to a Fourier transformation process, as it models processes separately and combines them to propose a final estimation. The proposed approach relies on the extension of the autonomy assumption, introduced in the subsection 2.1.2. At its core, the model is factorizing the joint distribution of these process Z into a number of conditional distributions given its causal parents processes.

Throughout this work the following notation is used: while a p-dimensional spatio-temporal process Z is a random variable, Z_s^t to denote the random vector obtained from marginalizing Z at spatial location s and temporal instance t. We use Z_s for the time series $(Z_s^t)_{(t\in\mathbb{N})}$, Z^t for the spatial process$(Z_s^t)_{(s\in\mathbb{R}^2)}$, and $Z^{(\mathcal{S})}$ for the spatio-temporal process $\mathcal{S} \subseteq 1, ..., p$. Then the following definition of a causal graph and model for spatial graph are presented as shown in the paper [3].

Definition 3.1 (Causal graphical models for spatio-temporal processes)**.** A causal graphical model for a p-dimensional spatio-temporal process Z is a triplet $(\mathcal{S}, \mathcal{G}, \mathcal{P})$ consisting of

- a family $\mathcal{S} = (S_j)_{j=1}^k$ of non-empty, disjoint sets $S_1, ..., S_k \subseteq \{1, ..., p\}$ with $\bigcup_{j=1}^k S_j = \{1, ..., p\}$
- a directed acyclic graph $\mathcal{G}$ with vertices (or nodes) $S1, ..., Sk$, and
- a family $\mathcal{P} = (\mathcal{P}^j)_{j=1}^k$ of collections $\mathcal{P}^j = \mathbb{P}^j z_{z\in Z_{|PA_j|}}$ of distributions on $(\mathcal{Z}_{|S_j|}, \mathcal{F}_{|S_j|})$, where for every j, $PA_j := \bigcup_{i:S_i\to S_j\in\mathcal{G}} S_i$. Whenever $PA_j = \emptyset, \mathcal{P}^j$ consist only of a single distribution which we denote by $\mathbb{P}^j$

To ensure that these relations have a causal interpretation we need to specify the distribution of Z under certain interventions. Assuming that the Markov property holds, we can infer some properties of the distribution and the causal graph. Then, since $\mathcal{G}$ is acyclical, w.l.o.g. $S_i \to S_j$ whenever $i > j$, we can induce the joint distribution P over Z for a set of confounders F as:

$$\mathbb{P}(F) = \int_{F1} \cdots \int_{F_k} \mathbb{P}^k_{z^{PA_k}}(dz^{(S_k)}) \ldots \mathbb{P}^1(dz^{(S_1)}) \tag{6}$$

Then, assuming the property of autonomy of the distribution holds, we can define an intervention on the process $Z^{(S_j)}$ as replacing the conditional distribution, $\mathbb{P}^j = \mathbb{P}(Z^{(S_j)}|PA_j)$ in (6) by the corresponding probability distribution of the intervened graph. As seen in the previous section 2.1.2, this graph differs from the observational conditional distribution as it does not depend on its parents processes PA_j.

Before moving to the estimation of the average causal effect of a treatment vector X over a response variable Y, with both variables being also affected by some latent (hidden) process H; it is required to introduce the proposed SCM that specifies a general "homogeneity" (i.e. structure) in the generation process with hidden variables. Next follows a discussion on the Latent spatial confounder model, as proposed by the authors:

Definition 3.2 (Latent spatial confounder model (LSCM))**.** Consider a spatio-temporal process $(X, Y, H) = (X_s^t, Y_s^t, H_s^t)_{(s,t)\in\mathbb{R}^2\times\mathbb{N}}$ over a real-valued response Y, a vector of covariates $X \in \mathbb{R}^d$ and a vector of latent variables $H \in \mathbb{R}^l$. We call a causal graphical model over (X, Y, H) with causal structure $[Y|X, H][X|H][H]$ a latent spatial confounder model (LSCM) if both of the following conditions hold true for the observational distribution:

- The latent process H is weakly stationary and time-invariant
- There exists a function $f : \mathbb{R}^{d+l+1} \rightarrow \mathbb{R}$ and i.i.d. sequence $\epsilon^1, \epsilon^2, ...$ of weakly-stationary spatial errors process, independent of (X, H), such that:

$$Y_s^t = f(X_s^t, H_s^t, \epsilon_s^t), \text{ for all } (s,t) \in \mathbb{R}^2 \times \mathbb{N} \tag{7}$$

The proposed LSCM model states the assumptions that specify a general structure from which it is possible to compute causal effects estimations. These assumptions are commented below:

1. The first important point concerns the causal graph $\mathcal{G}$, which is specified in the notation $[Y|X,H][X|H][H]$. It is defining the same graph for each location s and time t, where the treatment X_s^t causes Y_s^t and the hidden variables H_s^t causes both X and Y. It worth noting the space to which belong the dimensions of the variables $X \in \mathbb{R}^d, \quad H \in \mathbb{R}^l$. Thus the models are not restricting any introduction of multiples variables and confounders. Later in section 4.2, we explore how to extend the causal structure in order for it to consider multiples spaces acorss multiple time periods.

2. Two properties of the latent process H are assumed: a) *Time-invariance.* Mathematically this is defined as $P(H^1 = H^2 = ...) = 1$, so for each location s: $H_s^1 = H_s^2 = ... = H_s^m$. This assumption is crucial for computing the proposed causal estimator method, as it can be seen as the effect in a location s of X_s on Y_s is confounded only by H_s^1, a *fixed hidden effect* over time ;b) *Weakly stationary.* This means that its marginal distribution H_s^t is the same for all (s,t) [11], which is an important property that allow us to propose a generic estimator for $p(H = h)$

3. Finally, it is assumed that the same function f describes the relation amongst the spatio-temporal variables of the process $(X, Y, H) = (X_s^t, Y_s^t, H_s^t)$. In other words, this means we assume that Y_s^t depends on past realisations of (X, H), (X_s^t, H_s^t).

Combining these assumptions is essential for the average causal effect to be computed without the need of observing the hidden variables H and the mathematical proof will be further developed in the next section 3.3.

We have now all the pieces to introduce the method to quantify the causal effect proposed in the foundation paper, called “Average Causal Outcome”[12]

[11] in the original paper the notation H_0^1 encountered in several definitions, refers to this unique marginal distribution across the space

[12] Please note that this semantics refers to what was formally defined as “Average Causal Effect” in the foundation paper. This has been changed here to “Average Causal Outcome” to avoid confusion with the ACE introduced in Section 2.1.3 where ACE was understood as a contrast. The wording “outcome” seems to fit better the expected value of the intervention $do(X = x)$

Definition 3.3 (Average Causal Outcome)**.** The *average causal outcome* of X on Y is defined as the function $f_{AVE(X \to Y)} : \mathbb{R}^d \to \mathbb{R}$ for every $x \in \mathbb{R}^d$ given by:

$$f_{AVE(X \to Y)}(x) := \mathbb{E}[f(x, H_0^1, \epsilon_0^1)] \tag{8}$$

Follows next the demonstration that the previous definition 3.3 has causal interpretation, i.e. that $f_{AVE(X \to Y)}(x)$ is equal to the expected value of Y_s^t under any intervention that enforces $X_s^t = x$

$$\begin{aligned}
\mathbb{E}[Y_s^t | do(X_s^t = x] &= \mathbb{E}[f(x, H_s^t, \epsilon_s^t) | do(X_s^t = x] \\
&= \sum_h \mathbb{E}[f(x, h, \epsilon_s^t)] \; \mathbb{P}(H_s^t = h | X_s^t = x) \\
&= \sum_h \mathbb{E}[f(x, h, \epsilon_s^t)] \; \mathbb{P}(H_s^1 = h) \\
&= \mathbb{E}[f(x, H_0^1, \epsilon_0^1)] = f_{AVE(X \to Y)}(x)
\end{aligned}$$

The logic in the equation development above is similar to what was done in section 2.1. The first line is derived from the definition of intervention and equation (7), the second line from the law of total probabilities and the third line due to the intervention on X $H \perp_{\mathcal{G}} X$. Finally the fourth line relies on the assumption that the marginal distribution H_s^t is unaffected by the intervention.

3.3 Estimation of the average causal outcome

Since we do not observe the hidden variables, how do we estimate the average causal outcome (ACO) from the observed process (X,Y)? As a preliminary step to answer this question, we need to consider the author's proposal as a "spatial varying regression model", $f(.)$, that is allowed to change arbitrarily from one location to the other, within a model class given by the user $f(., H_s)$. So by allowing some degree of flexibility, the model is able to capture the time-invariant hidden effects h in each declination of the function $f(., H_s)$ (e.g. one per location). The ACO is then summarizing all these regressions by taking the expectation across space, and so over the hidden variables as describes visually in Figure 6 below.

Formally, the authors method "requires as input [from the user] a model class for the regressions $f_{Y|(X,H)}(., h), h \in \mathbb{R}^l$ alongside with its suitable estimator $\hat{f}_{Y|X} = (\hat{f}^m_{Y|X})_{m \in \mathbb{N}}$ and returns the estimator of the average outcome defined in (8) within the model class" [3]:

$$\hat{f}^{nm}_{AVE(X \to Y)}(X_n^m, Y_n^m)(x) := \frac{1}{n} \sum_{i=1}^{n} \hat{f}^m_{Y|X}(X_{s_i}^m, Y_{s_i}^m)(x) \tag{9}$$

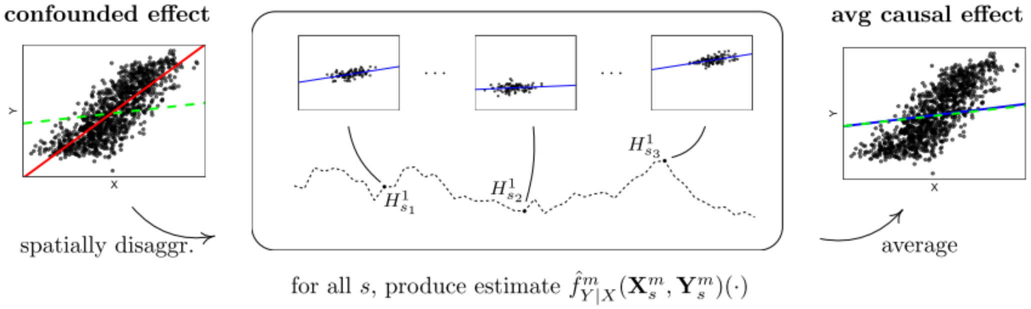


Figure 6: Conceptual idea for estimating the average causal effect (source: Christiansen et al., 2020

The last estimator in (9) has also a causal interpretation as proven in the next series of equations:

$$\begin{aligned}\mathbb{E}(Y|do(X=x)) =& \sum_y y\ P(Y=y|X=x)\\ =& \sum_h\ \mathbb{E}(Y|X=x,H=h)\ (H=h|X=x)\\ =& \sum_h\ \mathbb{E}(Y|X=x,H=H_s)\ p(H=h)\\ =& \sum_{s\in S} \hat{f}^m_{Y|X}(X^m_{s_i},Y^m_{s_i})(x)\ p(H=H_s)\\ =& \frac{1}{n}\sum_{s\in S} \hat{f}^m_{Y|X}(X^m_{s_i},Y^m_{s_i})(x)\end{aligned}$$

The logic behind the last equation development is as follow. In line (1) is applied the definition of expectation over Y and the assumption of autonomy (intervention doesn't affect the mechanism that triggers Y, so $P(Y|do(X=x) = P(Y|X=x))$. In line (2) the law of total probabilities holds; interventions induces that $X \perp H$ in line (3). In line (4), since H is time-invariant we can change the subscript h to location s, and we can use the definition regression estimator for a location. Finally in line (5), since H is weakly stationary we can assume the same distribution H_0^1 for any location s, and furthermore estimate it as: $\hat{p}(h) = \{\frac{1}{n}, \text{if h in the data}; 0, \text{otherwise}\}$

Finally, it worth noticing that the choice of the levels of regularization for the treatments x impacts the credible interval which yields the causal estimator $f_{AVEX \to Y)}$. Silva(2016) studied the same estimator (9) and constructed a dose-response curve that is learned non parametrically from observational data. The paper argues that even if the levels are derived from the data, it is probably not optimal for the causal estimator $(\hat{f}^{nm}_{AVE(x\to Y)})$. The reason is that since this strategy splits the data by locations s to compute each estimator, too many levels in X might very likely fragment the data-set even further in data-segments that are noticeably different from the actual population distribution of Y. This can potentially lead to poorly calibrated credible intervals. Data levels are discussed further in section 4.1.

3.4 Testing for the existence of causal effects

We seek to demonstrate that the causal effect of all the treatment X on Y is relevant, i.e. is $ACE(X \to Y) = 0$? Under the LSCM framework the null hypothesis can be formulated as:

$$H_0 : (X, Y) \text{ comes from an LSCM with a function } f(Y, X, H) \text{ constant w.r.t } X_s^t$$

Conversely, we could claim that this alternative hypothesis, X has a causal effect on Y, holds instead if the f function were to be significantly not constant for X. Next, we define the test statistic T that has power against the alternative hypothesis that we seek to disprove. In line with the previous H_0, it is logical to use a test statistic estimator that depends on the ACO $(f_{AVE(x \to y)})$. Therefore we must use a plug-in estimator: $\hat{T} = \psi(\hat{f}_{AVE(x \to y)})$.

Once a statistic T has been selected, Christiansen et al. (2020) [3] propose to test this hypothesis by constructing a resampling test following the setup presented by Pfister et al.(2018) [15]. This resampling procedure is based on the fact that X has no effect on Y under H_0, so the sequence of pairs (X_n^m, Y_n^m) in each location s must be exchangeable across time. The formal definition of excitability can be summarized as the property join distribution $p(X, Y)$ that remains equal after any permutation $\sigma_{k_b}(x, y)$ of any of the sequences, in our case noted Y_n^m. Therefore, we could estimate the distribution of the statistic $\hat{T}$ under H_0 by calculating it in B different permutations $\sigma_{k_b}(x, y)$, and comparing the distribution to the actual data so we can estimate an overall *p-value* that represents *"how likely it is to observe the actual data we are observing under H_0"*.

Formally the methodology for the test is as follow. Let $B \in \mathbb{N}$ and $k_1, ..., k_B$ be independent uniform draws from $\{1, ..., M\}$, with M the number of periods of time. Then we can define

- One-sided test p-value:

$$p_{\hat{T}}(x, y) := \frac{1 + |\{b \in \{1, .., B\} : \hat{T}(\sigma_{k_b}(x, y) \geq \hat{T}(x, y)\}|}{1 + B}$$

  Which can be explained as the proportion of the permuted statistics $\hat{T}(\sigma_{k_b}(x, y))$ that are greater or equal than the statistics found for the actual data $\hat{T}(x, y)$.

- Two-sided test p-value (of been equal to 0):

$$p_{\hat{T},2-sided}(x, y) := min(1, 2min(p_{\hat{T}}(x, y), p_{-\hat{T}}(x, y)))$$

  Where, $p_{-\hat{T}}(x, y)$ is the opposite one-sided test, so the proportion of permuted statistics that are less or equal than the negative statistic of the actual observed data $(-1 \times \hat{T})$.

The main advantage of the resampling method is being a non-parametric test, so there is no need for modeling restrictions, nor assumption on the data distribution, making it particularly useful for spatio temporal data as this class of assumption is widely considered as strong. On the other hand, this method limits the possibilities of this analysis, by restricting the statistics T that can be used. Particularly, so exchangeability can be assumed, the null hypothesis must state that the response variable Y is completely detached of **all** considered treatment variables X. Then, no T can be defined to test two treatment variables both, separately and simultaneously. Nevertheless it is possible to compute the causal impact of X on Y while considering any confounder Z by averaging across space, as it is presented in section 4.2.1.

3.5 Discussion on the estimator

Christiansen et al. [3] introduced a new methodology for establishing causal inference in spatio-temporal data, and suggested both a procedure to estimate causal effect from the data and a non-parametrical method for hypothesis testing of the causal effect. This section summarises the assumptions behind this complex procedure and discusses both, its impacts and limitations.

In addition to the assumptions previously introduced, Christiansen et al. [3] enforced the following two assumptions in order to ensure that the estimator of the causal average outcome is consistent, i.e. that the estimator in (9) tends to its expectation in (8) as the number of observations increases. These assumptions are: (A1) the “Law of Large Number for the latent process” and (A2) the“Consistent estimator of the conditional expectation”. Together, assumptions A1 and A2 imply additional conditions on the estimators $f(.)$ and T, which, while not strong, still limit the type of variables and functions this procedure covers.

First, we will focus on the function $f(.)$ in (7).

1. It is assumed that the same function f describe the process for the response variable Y. This is one of the stronger assumptions of the paper as its hard to prove since there might be locations s where this is not true.

2. Additionally, in order to enforce the assumption A2, the following structural assumption on the function f is made, in that $f(.)$ follows a simple regression model, so it can be separated into two sums given by functions such as:

$$Y_s^t = \phi(X_s^t)^\top f_1(H_s^t) + f_2(H_s^t, \epsilon_s^t), \text{ for all } (s,t) \in \mathbb{R}^2 \times \mathbb{N} \tag{10}$$

   Where, $\phi(X_s^t)$ is a known basis of continuous functions on $\mathbb{R}^d$, and $\phi_1 = 1$ an intercept term; $f_1 : \mathbb{R}^l \to \mathbb{R}^p$; and $f_2 : \mathbb{R}^{l+1} \to \mathbb{R}$

3. The regression model can be estimated via OLS:

$$\hat{f}_{Y|X}^m(X_s^m, Y_s^m)(x) = \phi(x)^\top \hat{\gamma}_s^m$$

where γ is the OLS estimator.

Conditions 2 and 3 are necessary only if we are using regression estimators. Still, theoretically the framework allows for other types of model, such as splines, random forest, etc., that fulfill both assumptions A1 and A2. These two conditions are listed here however since the foundation paper demonstrated that the OLS estimator is consistent in the LSCM framework, and that by enforcing them we ensure that our conclusions are correct.

Second, while the hidden variable H is not observed, three assumptions restrict the range of the variable H:

1. H is assumed to be weakly stationary and a time-invariant process. Therefore H does not consider variables with spatial trends nor those which change over time.

2. For assumption A1 to hold the authors demonstrated that this framework is consistent under its "proposition 8" (see the foundation paper [3]), which assumes that H is a multivariate Gaussian process with a covariance function. Including different forms of H would require further mathematical support first before they can be used in the LSCM this framework.

3. The process must be in a regular grid equally spaced, so the expectation over H across the space is correctly defined. However, note that weighted expectations could remedy the situation in which grid units do not have the same size.

Finally, three conditions on the predictors X are also enforced so the estimator and the causal effect are consistent:

1. The hidden variables only account for the time-invariant variables, so in order to ensure that the estimator in (9) describes a causal relation, all time-variant confounders of the relation X on Y must be included so the backdoor adjustment is satisfied.

2. X and ϕ must be such that $\phi(X)$ is independent from the error term: $\phi(X_s^i) \perp f_2(H_s^t, \epsilon_s^t)$. However this is not immediate to prove for all s as there might exist some unobserved time-variant variables that bias it.

3. if OLS regression is used, X and ϕ must be such that $\phi(X_{s_i}^t)^\top \phi(X_{s_i}^t)$ is invertible, so the standard OLS is well defined.

4 Methodology

The main outcome of this dissertation is to propose extensions to the foundation paper by Christiansen et al. [3] "Towards Causal Inference for Spatio-Temporal Data:

Conflict and Forest Loss in Colombia" by also considering spillovers and lagged causal effects to their overall approach which is summarized in section 3. By doing so, we seek to revisit their findings regarding the causal patterns between armed conflicts and deforestation in Colombia using their original dataset.

In this chapter, we will first introduce the data-set, provide descriptive statistics and justify the data transformations. We will then focus on explaining the causal models proposed, and the method requirements associated with them, i.e. a new functions estimator and a $\hat{T}$ statistic to test the hypothesis in the extensions.

4.1 Data and prepossessing

The dataset used in Christianse et al.(2020) [3] observed the effect of conflict on the deforestation in Colombia over both, space and time. This data originates mainly from two public available data-sets from 2000-2018, that were crossed and aggregated to grid-map of $10km \times 10km$ of Colombia: 1) forest loss data-set developed by Hansen et al.(2013) [11], and 2) conflict events from the Uppsala Conflict Data Program (UCDP) [23] [25]. Following the procedure and codes by Christiansen et al.(2020) the data set was downloaded and cross using python. The final data-set contains, among others, the following fields:

- *PolygonID*:id that identified the $10 \times 10km$ polygon in the grid. There are in total 11,350 different polygons but only 584 have a conflict in the analysed period.
- *Year*: the year of the observations.
- *Forest Cover*: "Tree cover in the year 2000, defined as canopy closure for all vegetation taller than 5m in height. Encoded as a percentage per output grid cell, in the range 0–100"[25].
- *Forest Loss*: yearly change on *Forest Cover*.
- *Conflict*: binary data that describes whether "armed force was used by an organized actor against another organized actor, or against civilians, resulting in at least one direct death at a specif location and a specific date".

Additionally the data contains some other possible confounders, specifically: distance to road, population and latitude-longitude. The last was used to defined to define the neighbourhood of the spillover effect, the other two were not used in this work, but could be used as confounder in posterior researches or to simulate data to test the power of the test in this same data.

Figure 7 summarises the distribution of the data. We can see that in a general view there is no evident correlation between conflict and forest lost variable, suggesting that there is no evidence for a causal relation, at least without considering the spatial

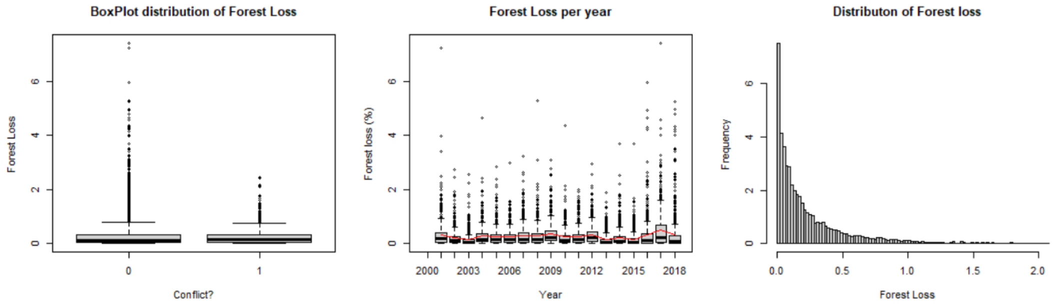


Figure 7: Data Exploration

The data analysed was filtered to only consider polygons that had at least one conflict and one not conflict years. Because the proposed methodology require to filter out points without variance in the treatment variable (See section 4.2.1 for more details) *The graphs describes the response variables forest loss: 1) The left panel boxplots compares the distribution of forest loss in conflict (mean=0.266, variance=0.115) and non conflict (mean=0.283, variance=0.2). There is limited visual and statistical association between these variables; 2) The central panel shows the distribution of the Forest Loss per year, the red line present the mean forest loss per year; 3) The right panel presents the distribution of the forest loss per year (skewness= 4.68128)*

relation and other confounder variables. The left panel suggest that the distribution of forest loss for both conflict and without conflict are similar, indeed the paired t-test to compare both forest loss means gives a p-value 0.1739. The central panel shows that there is no trend for the mean forest lost per year, suggesting that the data is actually stationary. Finally the right graph, shows that the distribution of the forest loss per year is skewed to the left, concentrating 95% of the cases that have a loss bellow than 1.0

On the other hand, Figure 8 analyses the correlation between the lags and the response variable. This analysis was done to correctly define the new variables that were used to consider lag and spillovers. The left panel shows that closer tiles' conflict is more correlated than for distance ones, up to 40km, still in average none relations appear to be strong. For that spillover were tested considering radius of 15km, 25km, 35km. Particularly interestingly, the middle graph shows that the forest loss is more correlated with past values (yet all the averages correlation are bellow 20%). Due to the limited observations (18 per location) and that number of observations that can be considered in a regression model decreases by the number of lags considered, which in turn reduces the power of any test, in this report only one lag is consider for our analysis. Finally the right panel shows that there is no significant correlation in the data for the treatment variable, so we should observe "colinearity" by including

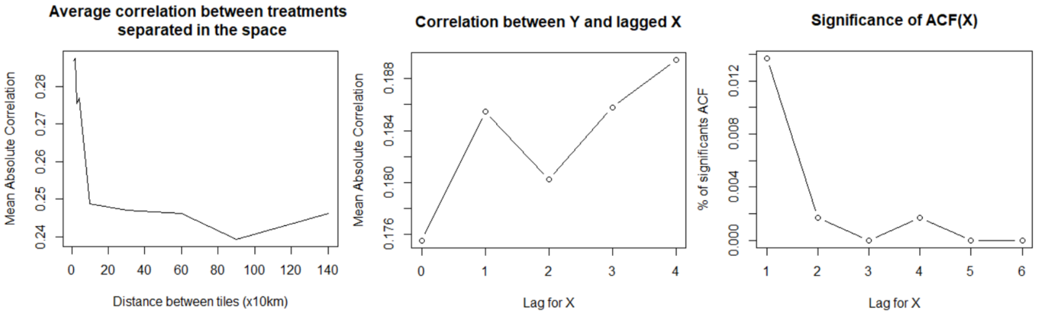


Figure 8: Lag and spillover data Analysis

The data analysed was filtered to only consider polygons that had at least one conflict and one not conflict years. Because the proposed methodology require to filter out points without variance in the treatment variable (See section 4.2.1 for more details). *The plots analyse the behavior of the lags and spillover for the treatment variable, conflict: 1) The left panel graph compares how the correlation between treatment variable in different locations changes as they are further apart in distance. 2) The middle panel study the correlation between Y_s^t (Forest Loss) and X_s^{t-k} (Conflict) as k grows, implying a possible lagged effect over an instantaneous effect; 3) The right panel summarizes the number of significant Auto-Correlation Function values for the treatment variable (i.e. $ACF(X)$), separated by the lag considered. Less than 2% of locations have a significant autocorrelation for X.*

any lag. Due to this is was decided that the spillover variable affecting tile s should also consider the information of the treatment recived.

Then, in order to account for spillover and lagged effects, we define the following two new variables:

- X_{t_1}: conflict variable lagged by 1 time period (years).
- I, I_1: Interference or spillovers summary variables, not lagged and lagged by one period. It is calculated as: $I_s^i = max(\{X_{s'}^{t-i}\})$. Where s' correspond to the neighborhood defined, in this case we tested with radius of $15km$, $25km$, $35km$, leading to similar results.

4.2 Proposed Extensions

This section is dedicated to explaining the models class and statistics test proposed associated to the extensions which aim to estimate the time series and spillovers causal effects for LSCM models. First, in subsection 4.2.1 the challenges to the LSCM estimators are described, alongside the models and statistic developed by Christiansen et al. (2020) which are reuiqred for the extensions. Then subsection 4.2.2 and section 4.2.3 explain the intricacies of considering both, causal time-lags and causal spillovers.

4.2.1 Challenges of the LSCM base estimator

The difficulties associated to the implementation of the framework developed by Christiansen et al. (2020) [3] are twofold: 1) developing estimators for the regression model f that capture the effect of the wanted variable X on the response variable Y; and more indirectly, 2) developing statistics $\hat{T}$, constructed from the $f_{AVE(X \to Y)}$, that can be used to test whether **all** treatments variables are relevant or not, and therefore test if they have causal meaning. This subsection discusses the challenges associated with proposing a new estimator under the LSCM framework and summaries the direct implications for the dataset under study.

As a first step towards establishing the extension, we discuss the original estimator used for this data by Christiansen et al. (2020). This statistic takes advantage of the treatment variable X_s^t being binary and suggests we estimate the causal influence of X on Y using the following estimator:

- Variables: Only treatment $X_s^t :=$ whether or not there was a conflict during year t in location s.
- Model estimator $\hat{f}_{AVE}$: the sample averages of the response variable for a treatment x

$$\hat{f}^{mn}_{AVE(X \to Y)}(x) = \frac{1}{|\mathcal{R}_n^m|} \sum_{i \in \mathcal{R}_n^m} \frac{1}{|\{t : X_{s_i}^t = x\}|} \sum_{t : X_{s_i}^t = x} Y_{s_i}^t, \qquad x \in \{0, 1\}$$

Where $\mathcal{R}_n^m$ is the set of locations after the filtering mechanism

- Data filtering mechanism: keep only the set of tiles where there are variance in the conflict variable, i.e at least one conflict and one peace period is observed in the data.

- Test statistic T: the contrast of AOC between the two possible values of X

$$T : f_{AVE(X \to Y)}(1) - f_{AVE(X \to Y)}(0)$$

The first challenge lies in the **definition of the treatment variables** and this task can be separated mainly in two parts. First, the sized of the dataset limits the *number of treatments variables* X that can be used for the function model class regression $f_{Y|(X,H)}(., h)$, as including more variables implies more combination-groups, which in turn reduces the number of observations of each group and the power of the test-statistic. Second, for the same reasons, the size of the dataset also limits the number of *levels that a treatment variable* X can have. This issue must be addressed by examining the variability in the data and selecting which variable combinations have sufficient observations. For this report models with only one or two variables and two levels are studied to avoid reducing the observed dataset bellow 300 locations.

The second challenge lies in the **definition of the model class**. The A1 and A2 assumptions are discussed in subsection 3.2, and we impose key restrictions on the data generating process, in particular that any estimator $\hat{f}^m_{Y|X}$ converges to the true ACO 8, at least in some area of X. Additionally, the size of the data-set also limits the options of model selection: for instance using modern regression models such as splines, random forest and gradient boosting with only 18 data-times and/or only one or two regressors is insufficient, as explained below. In this report, due to data limitations, i.e. there are only 19 observations per location, only counting and regression methods are explored as alternative testing methods would have insufficient power in this context.

The third challenge lies in the **definition of a data filtering mechanism**. In the LSCM base estimator introduced above, the computation requires filtering all locations s by excluding those which exhibit a null-variance over the treatment variables X. This would normally result in a biased estimator T. Nevertheless, Christiansen et al. argues that by making one additional assumption, even with a biased model $\hat{f}$ we can estimate an unbiased $\hat{T}$ statistic. In fact, if a causal mechanism between X and H does indeed exist, filtering locations s with regards values of X would result in a biased estimator $\hat{f}$, since the distribution of the latent variable H would now be distorted throughout the filtering process. Therefore, in the LSCM causal DAG intervened on X, the relationship $[H \to X]$ would disappear, making X and H independent. Then, since the estimator for f is assumed to adopt a linear form in 10, the equation 9 can be separated in the following sum of two terms:

$$f_{AVE(X\to Y)}(x) = \mathbb{E}[f(x,H,\epsilon)] = \mathbb{E}[f_1(x,\epsilon)] + \mathbb{E}[f_2(H,\epsilon)]$$

Where the second sum expectation contains the causal effect of H on Y and is constant for all values of X. So to compute $\hat{T}$, the effect of H (either biased or not) is canceled in the subtraction, i.e. $\hat{T} := f_{AVE(X\to Y)}(1) - f_{AVE(X\to Y)}(0) = \mathbb{E}[f_1(1,\epsilon)] - \mathbb{E}[f_1(0,\epsilon)]$. Conversely, in the absence of the filtering mechanism which allowed for X-invariant locations, both $\hat{f}$ and $\hat{T}$ would have been biased since X and H could now be considered co-linear in such points for they both are time-invariant. We can easily address this by defining a filtering mechanism that ensures variability in X^t_s for all locations s.

This very problem can also be addressed using the Ordinary Least Square (OLS) linear model as the model class estimators:

$$\hat{f}_{Y|X}(x) = \hat{\beta_0}^{OLS}_{s_i} + x\hat{\beta_1}^{OLS}_{s_i}$$

where $\hat{\beta_0}^{OLS}_{s_i}$ and $\hat{\beta_1}^{OLS}_{s_i}$ are respectively the intercept and the slope OLS coefficients of each location s_i. Within, this model class, all locations s_i without variance in X_{s_i} will have a constant $\hat{f}_{Y_{s_i}|X_{s_i}}$ for any value of x as $\hat{\beta_1}^{OLS}_{s_i} = 0$ (or =NA in R-project). Then, it is possible to see that by considering the AOC as:

$$\hat{f}_{AVE}(x) := \frac{1}{|i : \hat{\beta_1}^{OLS}_{s_i} \neq 0|} \sum_{i=1}^{n} (1,x)(\hat{\beta_0}^{OLS}_{s_i}, \hat{\beta_1}^{OLS}_{s_i})^\top$$

and the statistic test as:

$$\hat{T} := \hat{f}_{AVE}(1) - \hat{f}_{AVE}(0)$$

The effect of locations with invariant X is cancelled while computing $\hat{T}$, even though $\hat{f}$ is still biased.

Finally, the fourth challenge lies in the **definition of the test-statistics**. As explained in subsection 3.4, two criteria for the test-statistic $\hat{t}$ have to be fulfilled so the causal effect correctly estimates the effect of X on Y. First, it must be a plug-in estimator of the AOC, $\hat{T} = \psi(\hat{f}_{AVE(x\to Y)})$ and second, it must be a unique test for all the treatment variables considered in X. One solution for this could be a T that contrasts with the AOC evaluated for different values for X.

It is worth noting that using the F-statistics here instead is not straightforward. While the F-test is a test hypothesis which investigates wether or not any (or some) coefficients are jointly significant (i.e. $H_0 : \hat{\beta_1}^{OLS} = .. = \hat{\beta_n}^{OLS} = 0$), once correctly defined it would test if at least one of the treatment variable X has a causal effect on Y. Nevertheless, its application is not immediately applicable to the LSCM framework since the formal calculation of the F-statistic is based on the prediction error of the linear model, which does not make sense in this case. In fact, the LSCM has latent variables that are not estimated, and therefore there is no evident way to propose a

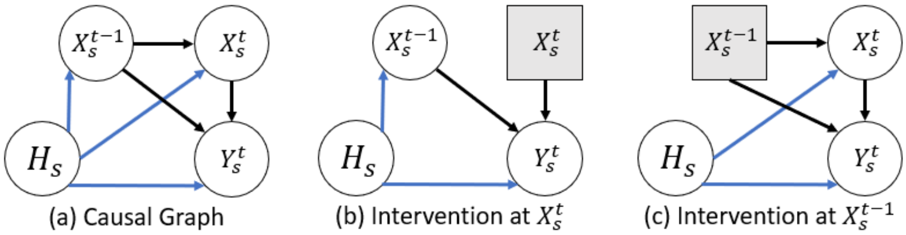


Figure 9: Lagged Causal and intervened Graph

Variables X_s^t, Y_s^t, H_s are present in the original model by Christiansen et al. (2020), while the additional node corresponds to the lagged treatment variable X_s^{t-1}. *This summarises the assumptions of our causal graph and panels (b) and (c) show the intervened graphs on* X_s^t *and* X_s^{t-1}

T that is function of $f_{AVE(X \to Y)}$ and that is based on its prediction error. In section 4.2.3 however, we suggest an alternative test and propose a new estimator, based on the total effect of spillovers, that is testing for the causal effect of two variables on the response variable simultaneously.

4.2.2 Extension 1: Time Series Model and relevance Hypothesis Test

The framework used to analyze the effect of lag time treatments is an extension to the time-series modelling presented in the foundation paper, which now allows f to depend on past values of the predictors. Therefore given a number of lags $k \geq 1$, equation 7 from the LSCM framework can be replaced by: $Y_s^t = f(X_s^{t-k+1}, ..., X_s^t, H_s^t, \epsilon_s^t)$ [3].

Here we are only considering one lag to avoid reducing the data-set too much (see figure 8 analysis in previous subsection 4.2.1), then $k = 1$. Panel (a) in Figure 9 summarises the assumptions of our causal graph and panels (b) and (c) show the intervened graphs on X_s^t and X_s^{t-1}. Additional lags could be added but the data size would be reduced for each location by the same number of lags considered, which in turn can lead to reduce the power and significance of the tests in hand. Note that the causal graph assumes a forward temporal causal relation between the treatment variables, $X_s^{t-1} \to X_s^t$, therefore this causal effect will be still in place when the lagged treatment is intervened, as shown in (c), but will not when X_s^t is intervened in (b). In this report the two treatments variables effects, X_s^t and X_s^{t-1}, will be respectively be addressed as instantaneous effect and lagged effect.

The extension estimators for $\hat{f}_{(AVEX \to Y)}$ and $\hat{T}$ are therefore defined as follow:

1. Instantaneous Direct Effect (DE-Inst):

   - Variables: X^t as treatment and X^{t-1} as confounder.

- Model class for a generic location s: $\hat{f}_{Y|X}(X_s^m, Y_s^m)(x^t, x^{t-1}) = (1, x^t, x^{t-1})^\top \beta^{OLS}(X_s^m, Y_s^m)$
- Data filtering process: Only polygons s with non-zero $Var(X_s)$
- T-statistic:

$$T = DE(f_{AVE(X \to Y)}) = f_{AVE(X \to Y)}(1, 0) - f_{AVE(X \to Y)}(0, 0)$$
$$\Rightarrow \hat{T} = \frac{1}{n} \sum_s^n (\beta_1^{OLS})_s = \lambda_1$$

- Hypothesis test $H_0 : (X^t) \perp Y | H \iff \lambda_1 = 0$

2. Lagged Direct Effect (DE-Lag1):

- Variables: X^{t-1} as treatment and X^t as confounder.
- Model class for a generic location s: $\hat{f}_{Y|X}(X_s^m, Y_s^m)(x^t, x^{t-1}) = (1, x^t, x^{t-1})^\top \beta^{OLS}(X_s^m, Y_s^m)$
- Data filtering process: Only polygons s with $Var((X^{t-1})_s) > 0$
- T-statistic:

$$T = DE(f_{AVE(X \to Y)}) = f_{AVE(X \to Y)}(0, 1) - f_{AVE(X \to Y)}(0, 0)$$
$$\Rightarrow \hat{T} = \frac{1}{n} \sum_s^n (\beta_2^{OLS})_s = \lambda_2$$

- Hypothesis test $H_0 : (X^{t-1}) \perp Y | H \iff \lambda_2 = 0$

3. Average instantaneous Direct Effect (do-Inst):

- Variables: X^t as treatment and X^{t-1} as confounder.
- Model class for a generic location s: $\hat{f}_{Y|X}(X_s^m, Y_s^m)(x) = \sum_{k \in dom((X^{t-1})_s)} \frac{|t: X_s^{t-1} = k|}{|\{t: X_s^t = x\}|} \sum_{\{t: X_s^t = x\}} Y_s^t$
- Data filtering process: All polygons s where the four possible combinations between X_s and $(X^{t-1})_s$ (i.e. $\{(0,0), (0,1), (1,0), (1,1)\}$) are observed
- T-statistic:

$$T = f_{AVE(X \to Y)}(1) - f_{AVE(X \to Y)}(0)$$

- Hypothesis test $H_0 : X^t \perp Y | H \iff T = 0$

4. Average Lagged Direct Effect (do-Lag):

- Variables: X_s^{t-1} as treatment and X_s^t as confounder.
- Model class for a generic location s: $\hat{f}_{Y|X}(X_s^m, Y_s^m)(x) = \sum_{k \in dom(X_s^t)} \frac{|R_s^{x,k}|}{|\{t: X_s^{t-1} = x\}|} \sum_{l \in R_s^{x,k}} \frac{Y_s^t}{|R_s^{x,k}|}$
with, $R_s^{x,k} = \{t : X_s^{t-1} = x, X_s^t = k\}$

- Data filtering process: All polygons s where the four possible combinations between X_s and $(X^{t-1})_s$ (i.e. $\{(0,0),(0,1),(1,0),(1,1)\}$) are observed.
- T-statistic:

$$T = f_{AVE(X \to Y)}(1) - f_{AVE(X \to Y)}(0)$$

- Hypothesis test $H_0 : X^{t-1} \perp Y|H \iff T = 0$

4.2.3 Extension 2: Spillover and lagged-spillover Models and relevance Hypothesis Test

Let us focus now on the **Spillovers Models**. Figure 10 shows the causal graph for spillovers and its representation after performing an intervention on X_s^t. Beside the original I_s^t and $H_{s'}$ variables, we now consider I_s^t which represent the interference or spillovers. In the most general case it is defined as follow:

$$I_s^t = \phi(\{X_{s'}^{t-i}\})$$

where s' represents a neighbourhood around position s and ϕ is a function that aggregates the variables $X_{s'}$ to one value: in this case we present the results using $max(.)$ [13]. The $t-i$ index stands for the lagged variable, i time periods before t. By symmetry, we also include the hidden variables $H_{s'}$ in the neighborhood s'.

Panel (a) in Figure 10 shows the causal graph and all potential relations assumed thus far between the different covariates, while panels (b) and (c) show the causal graphs when intervened in X_s^t and I_s^t respectively.

For us to simplify this analysis we need to make a few assumptions regarding the general causal graph shown in panel (a) of figure 10 with respect to the causal relations depicted in orange solid line. First, that I_s^t causes X_s^t (orange). We can justify this by the fact that I_S^t can be lagged in time and it makes sense to have temporal causality. If I_s^t is not lagged though, we could then assume direct independence of treatment between I and X. Second, we will also assume that $H_{s'}$ causally affects Y_s^t only through I_s^t and H_s^t. Therefore $H_{s'}^t$ is not directly affecting Y_s^t and so by conditioning on these two variables, all the paths are blocked and the backdoor criteria is satisfied. The assumption is not strong since logically $H_{s'}$ is time-invariant as H_s, and the framework assumes that H_s consider all the temporal variables that affect the location s. Finally, the symbol between $H_{s'}^t$ and H_s^t represents that the relation is not clear between these variables, even though given the previous assumption, we do not need to specify it to compute the causal effect of I_s^t and X_s^t on Y_s^t since we are conditioning on H_s the effect of $H_{s'}$ is blocked.

From the discussion above follows a series of combinations of estimators for $\hat{f}_{(AVEX \to Y)}$ and $\hat{T}$ based on the different potential outcomes which were proposed by Hudgens and Halloran (2008) [8] and reviewed in subsection 2.2.

[13] we also replicate the same analysis with $Avg(.)$ which leads to similar results

1. Direct effect (DE):

    - Variables: X_s^t as treatment and I_s^t as confounder.
    - Model class for a generic location s:
      $\hat{f}_{Y|X}(X_s^m, I_s^m, Y_s^m)(x, i) = (1, x, i)^\top \beta^{OLS}(X_s^m, I_s^m, Y_s^m)$
    - Data filtering process: Only polygons s with $Var(X_s) > 0$
    - T-statistic:

$$T = DE(f_{AVE(X \to Y)}) = f_{AVE(X \to Y)}(1, i) - f_{AVE(X \to Y)}(0, i)$$
$$\Rightarrow \ \hat{T} = \frac{1}{n} \sum_s^n (\beta_1^{OLS})_s = \lambda_1$$

    - Hypothesis test $H_0 : X \perp Y | H, I \iff \lambda_1 = 0$

2. Based on interference (indirect) effect (IE).

    - Variables: I_s^t as treatment and X_s^t as confounder. Note that the difference between the contrast values of X is 1 ($|X_1 - X_0| = |x - x'| = 1$)
    - Model class for a generic location s:
      $\hat{f}_{Y|X,I}(X_s^m, I_s^m, Y_s^m)(x, i) = (1, x, i)^\top \beta^{OLS}(X_s^m, I_s^m, Y_s^m)$
    - Data filtering process: Only polygons s with $Var(I_s) > 0$
    - T-statistic:

$$T = IE(f_{AVE(X \to Y)}) = f_{AVE(X \to Y)}(x, 1) - f_{AVE(X \to Y)}(x, 0)$$
$$\Rightarrow \ \hat{T} = \frac{1}{n} \sum_s^n (\beta_2^{OLS})_s = \lambda_2$$

    - Hypothesis test $H_0 : I \perp Y | H, X \iff \lambda_2 = 0$

3. Based on total effect (TE). This is the only extension-test that is testing simultaneously wether or not there is a causal effect of both variables X and I on Y. Therefore, this test could be considered similar to the application of an F-test for the LSCM model.

    - Variables: both I_s^t and X_s^t as treatment. Note that the difference between the contrast values of I is 1 ($|I_1 - I_0| = |i - i'| = 1$)
    - Model class for a generic location s:
      $\hat{f}_{Y|X,I}(X_s^m, I_s^m, Y_s^m)(x, i) = (1, x, i)^\top \beta^{OLS}(X_s^m, I_s^m, Y_s^m)$
    - Data filtering process: All polygons s with $Var(X_s) > 0$

- T-statistic:

$$\begin{aligned} T = TE(f_{AVE(X\to Y)}) &= f_{AVE(X\to Y)}(1,i) - f_{AVE(X\to Y)}(0,i') \\ &= f_{AVE(X\to Y)}(1,i) - f_{AVE(X\to Y)}(0,i) + \\ &\quad f_{AVE(X\to Y)}(0,i) - f_{AVE(X\to Y)}(0,i') \\ &= DE(f_{AVE(X\to Y)}) + IE(f_{AVE(X\to Y)}) \\ \Rightarrow \; \hat{T} &= \lambda_1 + \lambda_2 \end{aligned}$$

- Hypothesis test $H_0 : (X + I) \perp Y|H \iff \lambda_1 + \lambda_2 = 0$

4. Based on concept spatial average direct effect (SDE),[14].

   - Variables: Treatment X_s^t and is averaged over I_s^t
   - Model class for a generic location s:
     $\hat{f}_{Y|X,I}(X_s^m, I_s^m, Y_s^m)(x) = \sum_{k\in I_s} \frac{1}{|R_s^{(x,k)}|} \sum_{l\in R_s^{(x,k)}} (y_s^l) \frac{|I_s=k|}{|T|}$
     with $R_s^{(x,k=)}\{t : X_s^t = x, I_s^t = k\}$.
     So, this estimator can be described as the average of the averages of Y filtered and weighted by the probability of finding $I = i$ in the location s. (See demonstration of this estimator in appendix A).
   - Data filtering process: All polygons s where the four possible combinations between X_s and I_s (i.e. $\{(0,0),(0,1),(1,0),(1,1)\}$) are observed.
   - T-statistic:

$$\hat{T} = f_{AVE(X\to Y)}(1) - f_{AVE(X\to Y)}(0)$$

   - Hypothesis test $H_0 : X \perp Y|H, I$

It worth noticing that the assumption imposed, that the hidden spillover $H_{s'}$ causally affects Y_s^t only through I_s^t and H_s^t, could be studied further by proposing new estimators for $\hat{p}(h_s|h_{s'})$ that are required to advance in causal structure with modeled hidden spillovers. Nevertheless, it is possible that it could also take flexibility from the LSCM framework capability to model hidden variables.

5 Results and Discussion

This section presents the results obtained after the implementation of the the extensions suggested in section 4.2 to the Colombia's deforestation in response to armed conflicts. Section 5.1 analyse the results for the Time series model and section 5.2 for the interference models.

[14] note that the possible values for I are forced to be low dimensional (binary)

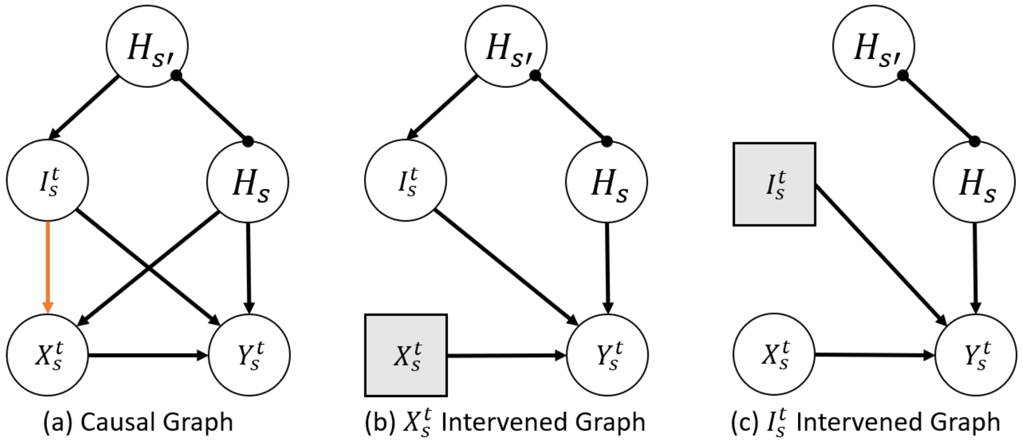


Figure 10: Spillover/Interference Causal Graph

Variables X_s^t, Y_s^t, H_s are the same variables proposed by the original model. We add I_s^t which is the spillover summary variable and $H_{s'}$ the hidden variable of tiles s'. The assumptions behind this DAG are discussed in main text

5.1 Extension 1: Time-series Model

Figure 11 reports the box plots of the distribution of p-values associated to the test statistic for the time-series models proposed in section 4.2.2. Each box-plot shows the empirical distribution of $\hat{T}$ under the null hypothesis, computed with the permutation test. It is also used to compute the significance of the observed $\hat{T}_{obs}$ that is summarised in the p-value.

In line with the foundation paper, we can conclude from these results that none of the statistic shown provide enough information to ensure that any of the causal effect are relevant, therefore we can not state if the causal effect exists or not and further studies must be undertaken.

Interestingly though, we can observe that the results for the Direct effect (DE-) and Average Direct effect (do-) are very close. In fact, DE-inst and do-Inst are virtually identical, while DE-Lag1 and do-Lag1 have minor differences, specifically the latter has a lower variance (specifically, $var(DE - Lag1) - var(do - Lag1) = 1.5918e - 05$). This difference might be explained using the causal graphs for the time series models shown in Figure 9. Indeed, for the intervention estimators, graphs (b) and (c) represent the intervened graphs used for the estimators do-Inst and do-Lag1 respectively. We can observe that the relation $X_s^{t-1} \to X_s^t$ in (b) is dropped while in (c) is still active. In comparison with the estimators that use linear regression as model class (DE-Inst and DE-Lag1) the causal graph is actually intervened in both variables therefore the causal relation $X_s^{t-1} \to X_s^t$ is dropped, making the causal graph equivalent to (b) but still different from (c).

5.2 Extension 2: Spillover and Lagged-Spillover Models

Figure 12 reports the results of computing the different tests statistic for spillovers proposed previously in section 4.2.3.

In line with the conclusion in Christiansen et al.(2020), results show that none of the different two-sided test indicates that any of the considered variables, being either conflict, spillover and lagged-spillover, is relevant. In a similar vein to the original paper, all the statistics with the exception of SDE for the non-lagged spillover statistic show negative effects on the response variable, implying that the conflict is actually reducing deforestation in average.

Additionally, we can see that the distributions of the statistics that consider the same variable (i.e. same color in the graph) are very similar. Furthermore, as expected the p-value of the TE statistic, which is defined as TE=DE+IE and test for both variables X and I, is in between the p-values obtained for DE and IE while the variance of TE is also larger that the other two.

5.3 Discussion

Overall, these results are in line with the foundation paper: the effect of conflict on deforestation is negative but not significant. Therefore our study is not conclusive, as we cannot reject the hypothesis $H_0 : (Y, X, I)$ *is originated from a LSCM with a constant function f with respect to* (X_s^t, I_s^t). In fact, it might be possible that the power of the statistics under study here is too small, and this conclusion is simply a false positive. There are several actions we could undertake to increase the power of our test. Sampling additional data is always a good strategy. Still, in our case this is not possible as this would mean collecting data for additional years, which does not make sense as the peace was signed already. An alternative would be to increase the granularity of our data-set by changing the period between observations from year to semesters. An other option to increase power is to make stricter assumptions to the model which would then require deeper knowledge of the problem at hand for us to propose more suitable statistics tests T and functions f. Also, one could manipulate the data-set by transforming the variables or the definition of spillovers accordingly,

In any case, it is possible to perform a power analysis for our statistics in order to asses if our conclusion in the extensions could be a false positive, i.e due to a low power. [15]

[15] Progress was made during this dissertation towards proposing a methodology that performs power analysis assessment using simulations based on the work by [19]. Nevertheless due to time constraints across summer term, the code was not finished, this can be shared upon request still.

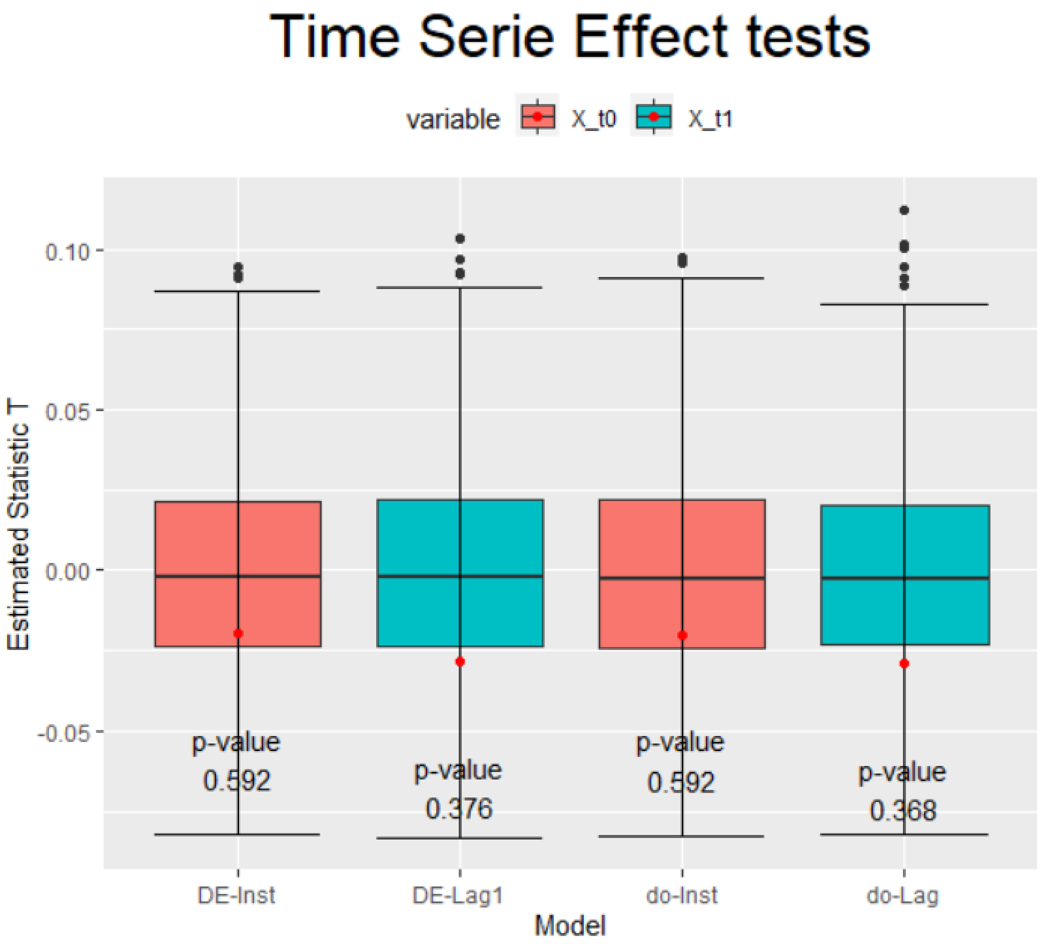


Figure 11: Time-series causal effect results

Each box plot shows the empirical distribution of $\hat{T}$ under the null hypothesis, calculated from 999 resampled data-sets. The red dots indicates the value observed actually in the data, which is used to calculate the shown two sided test p-value.
The horizontal axe indicate the estimator for $\hat{T}$ used. The meaning of the acronyms are: DE-Inst: Instantaneous Direct Effect; DE-Lag1: Lagged Direct Effect; do-Inst: Average instantaneous Direct Effect; do-Lag: Average Lagged Direct Effect.
The analysis is in line with the findings of Christiansen et al. (2020) as there is no evidence that the causal effect of any of the variables considered are significant. Additionally, the effect of the treatment, if not null, are probably negative, as observed in the foundation paper.

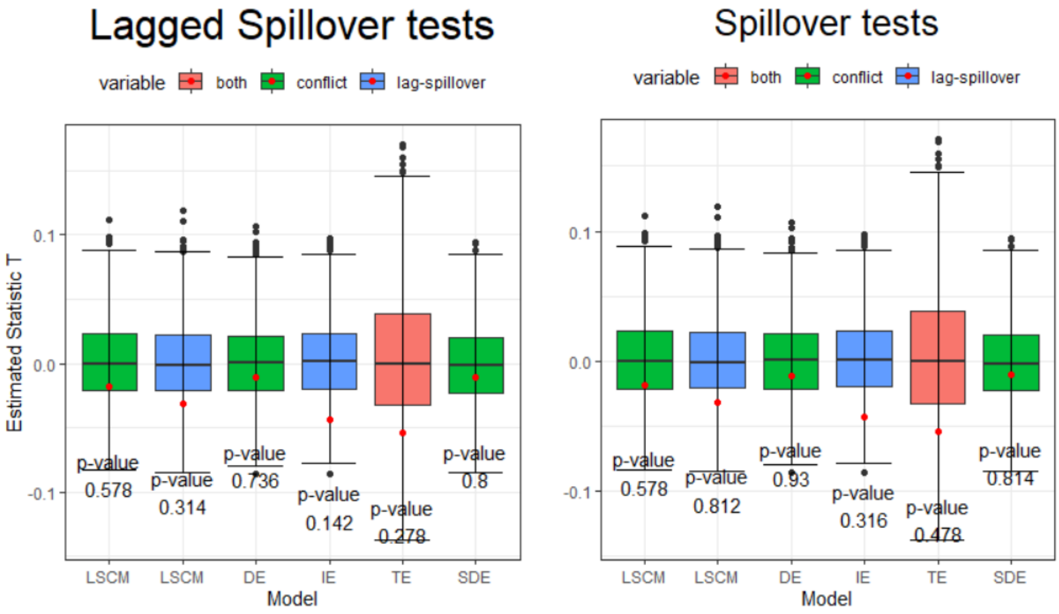


Figure 12: Spillover Interference results

The left plot shows the results considering instantaneous and lagged variables $X_s^t = conflict, I_s^t = max(\{X_{s'}^{t-1}\})$. On the right plot, it is shown the results considering the variables in the same time $X_s^t = conflict, I_s^t = max(\{X_{s'}^t\})$. Both graphs are read in the same fashion, each box plot shows the empirical distribution of T under the null hypothesis, calculated from 999 resampled data-sets. The red dots show the values found in the data, which were used to calculate the presented two sided test p-value. The horizontal axe indicate the statistic $\hat{T}$ model used: 1) LSCM statistic, as proposed by Christiansen et al. (2020); DE: Direct effect, the estimate effect of X_S^t over Y when I_s^t is considered as confounder, IE: Interference or spillover Effect, the estimate effect of I_S^t over Y when X_s^t is considered as confounder; TE: Total effect, the sum of both effect over Y; and SDE: The Spatial Direct Effect, the weighted average effect of X_s^t on Y_s^t weighted by the proportion of spillovers values.

6 Conclusion

6.1 Summary of findings

Throughout this report we presented an exhaustive introduction to causal analysis of spatio-temporal data. Starting with the basics of causal data discovery, we explained the ladder of causality and how to use causal graphs and probability distributions from the observational probabilities (as opposed to designed controlled experiments) to discover the causal relations and/or estimate the causal average effect. Then, the initial method for causal inference for spatio-temporal data proposed by Christiansen et al.(2020) is explained, introducing the reader to the use of the "latent Space Causal Model" (LSCM) framework. After having explained the intuition and theory behind the proposed framework with the goal of improving the model proposed in that paper, we describe the requirements associted with the investigation of Spillover effects, which intends to model the effect that a treatment at one location has on the response variable at another location, possibly untreated. In the last two chapters of this report, the suggested models for the extensions are described and a discussion on the results follows. While the models are not significant enough to conclude that there is a causal relation, the results are in line with the findings in the original paper findings, in that "conflicts have only weak explanatory power for predicting forest loss, and the potential causal effect is therefore likely to be small" [3]. Nevertheless, this should now be a comprehensive modelling that extended the applications of the LSCM framework towards estimating causality in spatio temporal data considering spillovers and/or time series effects.

6.2 Avenues for future research

A few questions emerge from this work. How can we formally compare the performance of these tests? This will require us to develop simulation tests to study the power (β) of the extension tests. Furthermore, the analysis of the effect of armed conflict on deforestation could be partly addressed by developing a power analysis via a simulation process to find out if, using the extension tests, the study is still underpowered to reach the correct conclusion. Progress towards that end was made based on the work by Quand(2020) but still remains unfinished.

In that regard, it would also be interesting to study the effects of confounders that are only temporal, such as the negotiation status between the FARC and the Colombian's government. To do so Christiansen et al.(2020) ensure that the LSCM model could be applied by averaging across time instead of space, therefore the analysis would be analog to the one above, by simply interchanging the time and space variables in the expectations. Finally, this framework does not evaluate if all the covariates required by the backdoor adjustment are considered in the model, thus leaving a procedure to be developed to that end. This is usually addressed by studying the distribution of the error terms in the regression model. Still in the case of spatio-temporal datasets, this is no direct, nor easy, task since it is averaged over space.

References

[1] M.-A. Bind. “Causal modeling in environmental health”. In: *Annual Review of Public Health* (2019).

[2] Rune Christiansen. “Causal Inference in the Presence of Latent Variables: Structure Learning, Effect Estimation and Distribution Generalization”. English. PhD thesis. 2020.

[3] Rune Christiansen et al. “Towards Causal Inference for Spatio-Temporal Data: Conflict and Forest Loss in Colombia”. In: (2020). arXiv: `2005.08639 [stat.ME]`.

[4] Alexander Coppock. *10 Things to Know about Spillovers.* URL: `https://egap.org/resource/10-things-to-know-about-spillovers/`. (accessed: 10.8.2021).

[5] Fabian Dablander. *An Introduction to Causal Inference.* Feb. 2020. DOI: `10.31234/osf.io/b3fkw`. URL: `psyarxiv.com/b3fkw`.

[6] Hume David. “An Enquiry Concerning Human Understanding”. In: *LaSalle* (1748).

[7] Sobel M. E. “What do randomized studies of housing mobility demonstrate? causal inference in the face of interference”. In: *Journal of the American Statistical Association* 101 (2006), pp. 1398–1407.

[8] Hudgens M. G. and M. E. Halloran. “Toward Causal Inference With Interference”. In: *Journal of the American Statistical Association* 103 (482 2008). DOI: `https://doi.org/10.1198/016214508000000292`.

[9] Ruocheng Guo et al. “A Survey of Learning Causality with Data: Problems and Methods”. In: *ACM Computing Surveys* 53 (Sept. 2020), pp. 1–37. DOI: `10.1145/3397269`.

[10] Messerli F. H. “Chocolate consumption, cognitive function, and Nobel laureates”. In: *The New England journal of medicine* 367.16 (2012), pp. 1562–1564. URL: `https://doi.org/10.1056/NEJMon1211064`.

[11] M. C. Hansen et al. “High-Resolution Global Maps of 21st-Century Forest Cover Change”. In: *Science* 342.6160 (2013), pp. 850–853. ISSN: 0036-8075. DOI: `10.1126/science.1244693`. eprint: `https://science.sciencemag.org/content/342/6160/850.full.pdf`. URL: `https://science.sciencemag.org/content/342/6160/850`.

[12] M. A. Hernn and J. M. Robins. “Causal inference: What if. Boca Raton: Chapman & Hall/CRC”. In: (2020).

[13] Ardo Van Den Hout. *Statistical Design of Investigations (STAT0029).* URL: `https://www.ucl.ac.uk/module-catalogue/modules/statistical-design-of-investigations-STAT0029`. (accessed: 1.9.2021).

[14] Pearl J., Glymour M., and Jewell N. P. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[15] B. Scholkopf N. Pfister P. Buuhlmann and J. Peters. "Kernel-based tests for joint independence". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2018).

[16] MIT OpenCourseWare. *Lecture 14: Causal Inference, P., 2021. Lecture 14: Causal Inference, Part 1*. URL: `https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-s897-machine-learning-for-healthcare-spring-2019/lecture-videos/lecture-14-causal-inference-part-1/`. (accessed: 01.09.2021).

[17] Judea Pearl and Dana Mackenzie. *The book of why: The new science of cause and effect*. Basic Books, 2020.

[18] Spirtes Peter and Zhang Kun. "Causal discovery and inference: concepts and recent methodological advances". In: *Applied Informatics* (1 2016). DOI: `10.1186/s40535-016-0018-x`.

[19] Julian Quand. *Power Analysis by Data Simulation in R, Part I to IV*. 2020. URL: `https://julianquandt.com/categories/power-analysis/`. (accessed: 01.09.2021).

[20] Brian J. Reich et al. "A Review of Spatial Causal Inference Methods for Environmental and Epidemiological Applications". English. In: *International Statistical Review* (Jan. 2021). DOI: `10.1111/insr.12452`.

[21] Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

[22] Ricardo Silva. *Observational-Interventional Priors for Dose-Response Learning*. 2016. arXiv: `1605.01573 [stat.ML]`.

[23] Ralph Sundberg and Erik Melander. "Introducing the UCDP Georeferenced Event Dataset". In: *Journal of Peace Research* 50.4 (2013), pp. 523–532. DOI: `10.1177/0022343313484347`. eprint: `https://doi.org/10.1177/0022343313484347`. URL: `https://doi.org/10.1177/0022343313484347`.

[24] Henri Theil. "Repeated least squares applied to complete equation systems". In: *The Hague: central planning bureau* (1953).

[25] Pettersson Therese et al. "Organized violence 1989-2020, with a special emphasis on Syria". In: *Journal of Peace Research* 58.4 (2021).

A Appendix1: Proof Estimator SDE

The following proof shows why the SDE estimator for f proposed in section 4.2.3 estimates the ACO.

Lets first analyse the f regression for a general s

$$\begin{aligned}
\mathbb{E}[Y_s^t|x_s^t, h_s] &= \sum_{k \in dom(I_s)} \mathbb{E}[Y_s, I_s = k|x_s^t, h_s] \\
&= \sum_{k \in dom(I_s)} \mathbb{E}[Y_s|I_s^t = k, x_s^t, h_s]\mathbb{P}[I_s^t = k|x_s^t, h_s] \\
&= \sum_{k \in dom(I_s)} \mathbb{E}[Y_s|I_s^t = k, x_s^t, h_s]\mathbb{P}[k|h_s]
\end{aligned}$$

We propose the following counting estimators for $\mathbb{E}[Y_s|I_s = i, x_s, h_s]$ and $\mathbb{P}_{h_s}[i]$:

- $\hat{\mathbb{P}}_{h_s}[I_s^t = k] = \frac{1}{|R_s^{(x,k)}|}$
  Where $R_s^{(x,k)} = \{t : X_s^t = x, I_s^t = k\}$, the subset of the observations T for the location s where $X_s^t = x$ and $I_s^t = k$
- $\hat{\mathbb{E}}[Y_s|I_s^t = k, X_s^t = x, H_s = h_s] = \sum_{l \in R_s^{(x,k)}} (y_s^l) \frac{|I_s = k|}{|T|}$
  Which is the average of Y_t with $t \in R_s^{(x,k)}$

Therefore, our estimtor for $\hat{f}$ is:

$$\hat{f}_{Y|X,H}(X_s^m, I_s^m, Y_s^m)(x) = \sum_{k \in I_s} \frac{1}{|R_s^{(x,k)}|} \sum_{l \in R_s^{(x,k)}} (y_s^l) \frac{|I_s = k|}{|T|}$$

Finally by plugin in this value into the ACO estimator from equation: (9)

$$\begin{aligned}
f^{nm}_{AVE(X \to Y}(X_n^m, Y_n^m)(x) &= \mathbb{E}[Y|do(X = x)] \\
&= \frac{1}{n} \sum_s^n \mathbb{E}[Y_s|x_s, h_s] \\
&= \frac{1}{n} \sum_s^n \sum_{k \in I_s} \frac{1}{|R_s^{(x,k)}|} \sum_{l \in R_s^{(x,k)}} (y_s^l) \frac{|I_s = k|}{|T|}
\end{aligned}$$