Tutorial on Causal Inference with Spatiotemporal Data

Sahara Ali
University of North Texas
Denton, USA
sahara.ali@unt.edu

Jianwu Wang
University of Maryland Baltimore County
Baltimore, USA
jianwu@umbc.edu

Abstract

Spatiotemporal data, which captures how variables evolve across space and time, is ubiquitous in fields such as environmental science, epidemiology, and urban planning. However, identifying causal relationships in these datasets is challenging due to the presence of spatial dependencies, temporal autocorrelation, and confounding factors. This tutorial provides a comprehensive introduction to spatiotemporal causal inference, offering both theoretical foundations and practical guidance for researchers and practitioners. We explore key concepts such as causal inference frameworks, the impact of confounding in spatiotemporal settings, and the challenges posed by spatial and temporal dependencies. The paper covers synthetic spatiotemporal benchmark data generation, widely used spatiotemporal causal inference techniques, including regression-based, propensity score-based, and deep learning-based methods, and demonstrates their application using synthetic datasets. Through step-by-step examples, readers will gain a clear understanding of how to address common challenges and apply causal inference techniques to spatiotemporal data. This tutorial serves as a valuable resource for those looking to improve the rigor and reliability of their causal analyses in spatiotemporal contexts.

CCS Concepts

• **Computing methodologies → Artificial intelligence**; **Knowledge representation and reasoning**; **Causal reasoning and diagnostics**;

Keywords

spatiotemporal data, causal inference, confounding, autocorrelation, machine learning

ACM Reference Format:
Sahara Ali and Jianwu Wang. 2024. Tutorial on Causal Inference with Spatiotemporal Data. In *1st ACM SIGSPATIAL International Workshop on Spatiotemporal Causal Analysis (STCausal'24), October 29-November 1 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3681778.3698786

1 Introduction

Spatiotemporal data, which encompasses both spatial and temporal dimensions, is increasingly critical in a wide range of scientific domains, including environmental monitoring, epidemiology, transportation systems, and climate modeling [5, 9, 12]. Understanding the dynamics of how variables evolve across space and time is key to uncovering patterns, making predictions, and identifying causal relationships. Causal inference goes beyond identifying correlations to uncover the actual mechanisms by which one variable influences another. In spatiotemporal data, failing to account for these dependencies can result in biased estimates of causal effects [1]. For example, a policy intervention implemented in a city might appear to affect crime rates, but if both the policy and crime are influenced by unmeasured regional factors (such as economic conditions), the causal effect could be confounded. Thus, to draw valid conclusions from spatiotemporal data, it is essential to account for confounders, spatial autocorrelation, and temporal dynamics.

In this tutorial, we provide an accessible yet comprehensive introduction to spatiotemporal causal inference. We will start off by presenting some key terminologies in causal inference, followed by steps to generate synthetic time-series and spatiotemporal datasets and finally look at some state-of-the-art approaches based on different modeling techniques. By the end of this tutorial, participants will have a strong understanding of how to approach causal inference in spatiotemporal data, and they will be equipped with practical tools and techniques for applying these methods to their own research. The supporting code and datasets for our work can be found at Github[1]

2 Key Concepts in Spatiotemporal Causal Inference

2.1 Causal Inference Frameworks

One of the most widely used frameworks for causal inference is the potential outcomes framework (POM) [10], which defines the causal effect of a treatment or intervention by comparing the outcomes under different hypothetical scenarios—typically, one where the treatment occurs and one where it does not. For spatiotemporal data, this framework becomes more complex, as both the treatment and outcomes vary over space and time. Though POM is not the only framework for causal inference, for an in-depth understanding of this and other frameworks, we refer the audience to [5].

2.2 Spatiotemporal Dependencies and Spillover Effects

In spatiotemporal data, observations at different locations and times are often not independent, leading to complex dependencies that must be accounted for in causal inference. These dependencies can arise because nearby locations tend to share similar environmental, social, or economic conditions, and events that occur in one

[1] https://github.com/SaharaAli16/spatiotemporal-causality/tree/main/stcausal2024

STCausal'24, October 29-November 1 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1154-1/24/10
https://doi.org/10.1145/3681778.3698786

place may influence nearby areas over time. A key aspect of spatiotemporal dependencies is the presence of **spillover effects** [2]. Spillover effects occur when the impact of a treatment or event at one location spills over to affect nearby locations. For example, in public health interventions, introducing a vaccination campaign in one city may reduce disease transmission not only within that city but also in neighboring regions as people move between locations. Similarly, a policy aimed at reducing pollution in one area may have downstream effects on air quality in surrounding regions due to the diffusion of pollutants over space.

2.3 Confounding in Spatiotemporal Data

Confounding occurs when an unobserved variable influences both the treatment and the outcome, leading to biased estimates of the causal relationship. In spatiotemporal data, this problem is compounded by the presence of **spatiotemporal confounders**, which vary across both space and time. These confounders can introduce biases that distort causal inferences if not properly accounted for.

Spatial Confounding arises when a spatially structured variable affects both the exposure (or treatment) and the outcome, creating a spurious association between them [1]. For example, in an analysis of air pollution and health outcomes across different cities, socioeconomic factors like poverty or healthcare access—both of which may vary by location—can confound the relationship. Without adjusting for these spatially varying confounders, any observed relationship between air pollution and health outcomes might reflect the influence of these unmeasured factors rather than a true causal effect.

Temporal Confounding occurs when a time-varying factor influences both the treatment and the outcome [10]. For instance, in a study examining the effect of an economic policy on employment rates over time, macroeconomic trends or seasonal patterns (e.g., holidays or weather-related economic slowdowns) may simultaneously affect both the policy implementation and the employment outcomes, leading to biased causal estimates.

3 Methods for Spatiotemporal Causal Inference

This section forms the core of the tutorial, where we will present different techniques for causal inference that are suited to spatiotemporal data. We will break down this section into three categories; temporal, spatial and spatiotemporal causal inference methods.

Temporal Causal Inference Methods focus on causal relationships in data that vary over time, assuming there are temporal dependencies but no spatial structure to account for. These include time-varying and time-invariant methods. In this tutorial we will go through the following techniques from each category (i) Difference-in-Difference (DiD) - a time-invariant method [7], (ii) Marginal structure Models - a propensity-score based method for time-varying data [8], (ii) Time-series Deconfounder - a latent factor model to reduce temporal confounding bias. [6] (iii) TCINet - a deep learning based causal inference method for continuous data [3].

Spatial Causal Inference Methods account for spatial autocorrelation by incorporating a spatial lag or error term in the regression model. In this tutorial, we will look at Weather2Vec - a spatial causal inference technique for capturing spatial confounding effects [11].

Spatiotemporal Causal Inference Methods integrate both spatial and temporal dependencies, addressing the challenges of causal inference when data varies over both space and time. We will go over STCINet - a spatiotemporal model that can capture both confounding effects and spillover effects in space and time [4].

4 Synthetic Dataset Generation

To test causal inference methods for tracking information flow in spatiotemporal data, we present two variants of synthetic datasets, (i) autocorrelated time-series data, (ii) spatiotemporal data to mimic a dominant physical process found in many geo-science applications, that is, diffusion.

4.1 Time-series Dataset

We adopt the data-generation process followed by TCINet [3] to generate four non-linear time-series A, B, C, and D. Using Gaussian white noise ε, we generate these time-series given in Equations 1 to 4.

$$A_t = cos(\frac{t}{10}) + log(|A_{t-6} - A_{t-10}| + 1) + 0.1\varepsilon 1 \tag{1}$$

$$B_t = 1.2e^{\frac{A_{t-1}^2}{2}} + \varepsilon 2 \tag{2}$$

$$C_t = -1.05e^{\frac{-A_{t-1}^2}{2}} + \varepsilon 3 \tag{3}$$

$$D_t = -1.15e^{\frac{-A_{t-1}^2}{2}} + 1.35e^{\frac{-C_{t-1}^2}{2}} + 0.28e^{\frac{-D_{t-1}^2}{2}} + \varepsilon 4 \tag{4}$$

4.2 Spatiotemporal Dataset

Diffusion is a physical process that describes the movement of particles or substances from regions of higher concentration to regions of lower concentration. It is driven by the random motion of particles, and it tends to equalize the concentration of substances in a given medium over time. Following this concept, we generate three spatiotemporal variables A, B and C adapting the data-generation process of STCINet [4]. Where A is an independent variable with spatial and temporal autocorrelations. B is dependent on A and C is dependent on both A and B. The spatial domain is represented by $[i, j] \epsilon N \times M$ and t represent the notion of time. α, β, and γ are causal coefficients to incorporate causal influence in these variables. D_a, D_b, D_c are diffusion coefficients, whereas dt is the time step size. A Laplacian operation ∇^2 is performed on each variable to model its spatial diffusion in the dataset. The key steps involved in generating the dataset are given in our Github repository.

$$A_{i,j}^t = A_{i,j}^{t-1} + dt \times \left(D_a \times \nabla^2 A\right) \tag{5}$$

$$B_{i,j}^t = B_{i,j}^{t-1} + dt \times \left(D_b \times \nabla^2 B + \alpha \times \nabla^2 A_{i,j}^{t-1}\right) \tag{6}$$

$$C_{i,j}^t = C_{i,j}^{t-1} + dt \times \left(D_c \times \nabla^2 C + \beta \times \nabla^2 A_{i,j}^{t-1} + \gamma \times \nabla^2 B_{i,j}^{t-1}\right) \tag{7}$$

5 Challenges and Limitations

While spatiotemporal causal inference methods offer powerful tools for understanding relationships in data that vary across space and time, there are several inherent challenges and limitations that complicate their application. These challenges stem from the complex nature of spatial and temporal dependencies, the presence of confounders, and the difficulty of identifying causal relationships in dynamic, interconnected systems. Below are some of the key challenges:

5.1 Data-related Challenges

Spatiotemporal datasets are often large and high-dimensional, with observations across many time points and spatial units (e.g., cities, regions). However, despite the large number of observations, data can be sparse in certain contexts, especially if observations are missing for some regions or time periods. This can make it difficult to detect causal relationships and increases the risk of overfitting models.

5.2 Model-related Challenges

Modeling spatiotemporal data often requires the use of advanced statistical and computational techniques, such as **spatial econometrics**, **Bayesian hierarchical models**, or **spatiotemporal machine learning models**. These models can be computationally intensive and difficult to implement, especially for researchers without expertise in these areas. Moreover, model specification is challenging, as misspecified models may lead to incorrect causal inferences.

5.3 Conceptual Challenges

Identifying valid causal relationships in spatiotemporal settings is particularly difficult due to the presence of spatial and temporal confounders, as well as spillover effects. Many of the standard methods for causal inference (e.g., randomized controlled trials, natural experiments) are hard to implement in spatiotemporal contexts because of this reason.

5.4 Computational Challenges

Spatiotemporal data can be large and complex, making the application of advanced causal inference models computationally expensive. Many of the methods used to account for spatiotemporal dependencies, such as spatial econometrics or machine learning models, involve intensive computations, especially when dealing with high-resolution spatial data or long time series.

5.5 Uncertainty Quantification

Quantifying uncertainty in causal estimates is often difficult. This is especially true for models that account for multiple layers of dependencies, where uncertainty from spatial and temporal components can compound.

6 Audience

This tutorial is aimed at researchers, students and practitioners from a variety of domains who wish to apply causal inference techniques to their spatiotemporal data. The tutorial covers fundamental concepts, discusses common challenges, introduces key methods for causal inference in spatiotemporal settings, and demonstrates practical applications using multiple synthetic datasets. For prerequisite, basic understandings of causality and machine learning will be preferred, but the tutorial will also introduce the basic concepts for better audience engagement.

7 Conclusion

This tutorial has provided an overview of the key methods and challenges in spatiotemporal causal inference. By incorporating both spatial and temporal dependencies into causal models, researchers can make more robust inferences about the underlying mechanisms driving changes in spatiotemporal data. We have highlighted several methods, including difference-in-differences, propensity score-based, and deep learning approaches, and demonstrated their application on synthetic datasets. We hope this tutorial serves as a foundation for further exploration and innovation in the analysis of spatiotemporal data.

References

[1] Kamal Akbari, Stephan Winter, and Martin Tomko. 2021. Spatial Causality: A Systematic Review on Spatial Causal Inference. *Geographical Analysis* (2021).

[2] Kamal Akbari, Stephan Winter, and Martin Tomko. 2023. Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis* 55, 1 (2023), 56–89.

[3] Sahara Ali, Omar Faruque, Yiyi Huang, Md Osman Gani, Aneesh Subramanian, Nicole-Jeanne Schlegel, and Jianwu Wang. 2023. Quantifying causes of arctic amplification via deep learning based time-series causal inference. In *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 689–696.

[4] Sahara Ali, Omar Faruque, and Jianwu Wang. 2024. Estimating Direct and Indirect Causal Effects of Spatiotemporal Interventions in Presence of Spatial Interference. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 213–230.

[5] Sahara Ali, Uzma Hasan, Xingyan Li, Omar Faruque, Akila Sampath, Yiyi Huang, Md Osman Gani, and Jianwu Wang. 2024. Causality for Earth Science–A Review on Time-series and Spatiotemporal Causality Methods. *arXiv preprint arXiv:2404.05746* (2024).

[6] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. 2020. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*. PMLR, 884–895.

[7] Michael Lechner. 2011. The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics* 4, 3 (2011), 165–224. https://doi.org/10.1561/0800000014

[8] Bryan Lim. 2018. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in Neural Information Processing Systems* 31 (2018).

[9] Brian J Reich, Shu Yang, Yawen Guan, Andrew B Giffin, Matthew J Miller, and Ana Rappold. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review* 89, 3 (2021), 605–634.

[10] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[11] Mauricio Tec, James G Scott, and Corwin M Zigler. 2023. Weather2vec: Representation learning for causal inference with non-local confounding in air pollution and climate studies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14504–14513.

[12] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.