Operationalizing Spatial Causal Inference *

Tyler D. Hoffman[1,2][0000−0002−0012−3794], Peter Kedron[1,2][0000−0002−1093−3416]

[1] Arizona State University, Tempe AZ 85281, USA
[2] Spatial Analysis Research Center, Tempe AZ 85283, USA
tdhoffman@asu.edu, Peter.Kedron@asu.edu

Abstract. Most spatial inquiries seek to investigate causal questions about spatial processes, but many quantitative spatial methods are designed to identify correlations and spatial patterns. Studying the structure of associations that make up a spatial pattern can provide information about what the process that generated that pattern is likely to be, but it does not provide a means of testing any one explanation against alternative explanations. Causal inference provides a set of approaches to formally make comparisons between explanations. An opportunity exists to incorporate these techniques and spatialize the theory of cause in GIScience by building on recent advances in computer science and statistics. However, implementing causal inference in geography may require a shift in the design of geographic information systems.

Keywords: causality · spatial process · spatial data science

DOI: https://doi.org/10.25436/E2D30K

1 Analytical methods in GIScience

In geographic information science (GIScience), a disconnect exists between the desire to understand process and many of the quantitative methods we use to study the world (13). Although subject to debate (14), the term "spatial process" refers to the dynamics of a (spatial) system or the temporal trajectory of spatial events. Spatial processes describe the mechanisms that generate spatial phenomena (13), which makes understanding process central to the causal explanations of spatial patterns.

For decades the discipline's quantitative textbooks have presented a "pattern–process" approach to inquiry. We observe a spatial pattern, analyze the structure of associations between variables within that pattern, and use those associations to make inferences about the processes that may have given rise to the pattern. (1; 12; 32). When we employ this method of inquiry, what we often ultimately wish to understand are the causal processes that produce the patterns we observe and the associations we measure. However, the associations we build our

* This material is based on work supported by the National Science Foundation through a Graduate Research Fellowship awarded to Hoffman under NSF Grant No. **026257-001** and project grant No. **BCS-2049837** awarded to Kedron. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

inferences on are necessary but insufficient conditions for identifying the process responsible for the pattern because the same set of associations could be the result of many different processes. Developing ways to specify, examine, communicate, and infer from a set of unobserved counterfactual processes that could have given rise to a pattern we observe is one key to closing the gap between our desire to understand process and our quantitative toolkit.

Recent advances in computer science and statistics have created an opportunity to bring the presentation and investigation of counterfactuals into GIScience and to enrich the pattern–process approach. Specifically, researchers in these fields have coupled probability theory with first-order logic (see 4; 28) to create a mathematical means of testing one set of proposed causal relationships against another. These advances create an opportunity for GIScience to potentially progress up the "ladder of causation" (Table 1) (35) from investigating associations to studying interventions and/or counterfactuals. Causal inference allows for queries to move beyond correlation by modeling interventions so researchers may study their effects. However, direct interventions in geographic systems are often difficult to study as we often cannot control geographic systems and face a range of confounding factors linked to the qualities of places which cannot be easily modeled. We may be able to side step this challenge by posing and investigating counterfactuals that explore what could or would have happened if one aspect of the world was different.

Introducing these changes into geographic inquiry will require progress in at least two areas (1) the further development of theories of spatial causal inference (Section 2) and (2) a reorganization of some of our most commonly used computational and statistical tools (Section 3).

Table 1: Ladder of Causation (35).

Level	Typical Question	Definition
Association $P(y\|x, z)$	How does seeing x change my belief in y?	I took an aspirin. Will I still have a headache when I wake?
Intervention $P(y_x\|z)$	How likely is it that y happens if I fix x?	If I take an aspirin now, will I wake up with a headache?
Counterfactual $P(y_x\|x', y', z)$	Was it x that caused y, given that we observed x' and y'? What if I acted differently?	I took an aspirin. My headache is gone. Would I have a headache had I not taken that aspirin?

Note: All the probability statements include z, which represents other variables that may confound the relationship between x and y.

2 Developing theories of spatial causal inference

Development of a formal theory of space and causal inference is still in its early stages, and recent progress toward any such theory has been made indirectly through the extension of causal models into the spatial domain (for reviews, see (2; 31); for examples, see (15; 16; 19; 21; 22; 30; 31; 23)). However, extending existing models of causal inference into the spatial domain faces several key challenges. First, geographic inquiries often rely on on observational data because it is simply infeasible, or potentially unethical, to randomly intervene the system under study. Second, spatial models may be confounded by unobserved and unquantifiable location-specific phenomena. Third, observations in space may interfere with each other. The second and third challenges are linked to Anselin's principle of spatial heterogeneity and Tobler's First Law. Just as these principles challenge the foundations of conventional statistics necessitating the development and use of spatial statistics (3; 32), they also challenge the foundations of causal inference necessitating the development of frameworks for spatial causal inference (20).

The principle of spatial heterogeneity implies that unobserved spatial effects may confound inferences. Beyond those variables that can be measured, places are made up of an amalgamation of minute, unquantifiable, and highly local qualities. A perfect quantitative model of these phenomena is therefore prohibitively difficult to construct. Instead, a formalism for acknowledging spatial confounding and controlling for it in quantitative models would enable researchers to condition their inferences on these unobserved quantities, simultaneously quantifying uncertainty and codifying their modeling assumptions. Then, software based on this formalism could be developed to identify instances of spatial confounding and (when present) automatically adjust inferences.

Tobler's First Law states that spatially embedded objects tend to be more related as they get nearer to each other. In causal inference, this manifests as spatial interference: nearby units may influence each other's responses to an intervention. When present in models, spatial interference can lead to violations of the stable unit treatment value assumption (SUTVA) of non-spatial causal inference. SUTVA holds that that units (treated or untreated) will not affect each other's response to an intervention (18). In some spatial scenarios, this is easily satisfied: consider a study about factory pollution levels and lung cancer. Two individuals may not affect each other's exposure to pollution from a factory, satisfying the SUTVA[1]. But many spatial problems violate the SUTVA. Consider a study about the relationship between county-level fireworks bans and hospitalization rates: if one county bans fireworks but its neighbor does not, individuals can travel to the neighbor and acquire fireworks, potentially affecting hospitalization rates in the original county. A first step in addressing this challenge would be the development of tools to warn researchers of potential spatially driven SUTVA violations. As such violations can be driven by a variety of complex effects, introducing an formal system that helps researchers identify potential issues would be an important, initial step forward.

3 Integrating recent advances into GIS

The data models and computational foundations of geographic information systems (GIS) shaped the pattern–process approach. The cartographic origins of GIS aligned naturally with ontological foundations and data models that represents real world entities as objects located in space and time (5; 10; 36)[2]. Quantitative spatial methods inherit this object-centered representation and the overlay driven analysis used to compare analog maps. They have been designed to examine the spatial pattern of a set of objects at one time, or the difference between a set of objects at several time points. What object-oriented ontologies and spatial methods do not prioritize or directly analyze, is process.

Understanding causation is facilitated by the analysis of counterfactuals. A syntax for expressing and manipulating spatial counterfactuals is critical for the development of further theory in spatial causality. Computationally, recent innovations in probabilistic programming have made it possible to digitally implement counterfactual inference. Probabilistic programming languages (PPLs) deliver a toolchain for computationally expressing random variables and probabilistic calculation (27). Typically, these languages can perform statistical (associational) inference on purely probabilistic relationships (17), but some PPLs (like gfoRmula (26), Omega (38), MultiVerse (29), and DoWhy (34)) have novel designs that permit counterfactual inference. Constructing a taxonomy of causal spatial queries would be a first step towards representing space in a PPL. Next, building a syntax for spatial interventions in these languages or in a new PPL would allow researchers to deploy theories of spatial causal inference to answer causal questions related to locations (e.g., "which counties would experience lower hospitalization rates if a fireworks ban was implemented?" or "would species A migrate through region R if regulation S was enforced?").

4 Conclusion

Understanding causality and process has long been a central goal of research in geography and GIScience. Despite consensus about the need to understand process in GIScience, implementation of quantitative methods that can facilitate the transparent communication and formal testing of causal models has remained limited. This absence is in part the product of the ontological foundation of most geographic information systems, the slow development of the mathematical approaches to causality, and computational limitations. However, technological innovations in statistics and computer science have made it possible to begin building theory and systems to examine causation with spatial data.

By building on these innovations, we can open the door to developing a cause- and process-oriented perspective in GIScience that may align our methods with the target of our inquiries. Stepping down this path begins with designing tools to identify potential violations of underlying causal assumptions and creating ways to programmatically express and infer from spatial counterfactuals. By placing cause and process at the center of studies, researchers will be able to reframe research questions from "what is happening and where is it occurring?" to "how is this happening and what drives it?"

Notes

[1]In fact, this is a problem where satisfying the SUTVA is a modeling decision rather than an innate feature of the problem. The researcher could equivalently model individual interactions that result in exposure to pollution, in which case individual level treatments would violate the SUTVA.

[2]The distinction between objects and fields and the ontological accuracy of the representation of each been the subject of much debate (25; 24; 37; 11). GIS progressed naturally from the vector model, to the computationally quicker raster model, and finally hierarchical object-oriented programming. However, the GIScience literature now recognizes these models as exchangeable (6; 7; 8; 9; 10; 33; 39).

Bibliography

[1] Abler, R.: Spatial Organization: The Geographer's View of the World. Prentice-Hall, Inc., New Jersey (1971)

[2] Akbari, K., Winter, S., Tomko, M.: Spatial Causality: A Systematic Review on Spatial Causal Inference. Geographical Analysis p. gean.12312 (Dec 2021). https://doi.org/10.1111/gean.12312, `https://onlinelibrary.wiley.com/doi/10.1111/gean.12312`

[3] Anselin, L.: Spatial Econometrics: Models and Methods. Kluwer Academic Publishers, Norwell, MA (1988)

[4] Cartwright, N.: Probabilities and experiments. Journal of Econometrics **67**, 47–59 (1995)

[5] Gahegan, M.: Our GIS is too small: Our GIS is too small. The Canadian Geographer / Le Géographe canadien **62**(1), 15–26 (Mar 2018). https://doi.org/10.1111/cag.12434, `https://onlinelibrary.wiley.com/doi/10.1111/cag.12434`

[6] Galton, A.: A formal theory of objects and fields. In: Montello, D. (ed.) Spatial Information Theory: Foundations of Geographic Information Science: International Conference, COSIT 2001. Springer Lecture Notes in Computer Science, New York (2001)

[7] Galton, A.: Desiderata for a spatio-temporal geo-ontology. In: Kuhn, W., Worboys, M., Timpf, S. (eds.) Spatial Information Theory: Foundations of Geographic Information Science: International Conference, COSIT 2003, Kartause Ittingen, Switzerland, September 24-28, 2003: Proceedings. Springer Lecture Notes in Computer Science, New York (2003)

[8] Galton, A.: Fields and Objects in Space, Time, and Space-time. Spatial Cognition & Computation **4**(1), 39–68 (Mar 2004). https://doi.org/10.1207/s15427633scc0401$_4$, `http://www.tandfonline.com/doi/abs/10.1207/s15427633scc0401_4`

[9] Goodchild, M.F.: Geographical data modeling. Computers & Geosciences **18**(4), 401–408 (May 1992). https://doi.org/10.1016/0098-3004(92)90069-4, `https://linkinghub.elsevier.com/retrieve/pii/0098300492900694`

[10] Goodchild, M.F., Yuan, M., Cova, T.J.: Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science **21**(3), 239–260 (Mar 2007). https://doi.org/10.1080/13658810600965271, `http://www.tandfonline.com/doi/abs/10.1080/13658810600965271`

[11] Grenon, P., Smith, B.: SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition & Computation **4**(1), 69–104 (Mar 2004). https://doi.org/10.1207/s15427633scc0401$_5$, http://www.tandfonline.com/doi/abs/10.1207/s15427633scc0401_5
[12] Haining, R.: Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge (2003)
[13] Harvey, D.W.: Pattern, Process, and the Scale Problem in Geographical Research. Transactions of the Institute of British Geographers (45), 71 (Sep 1968). https://doi.org/10.2307/621393, https://www.jstor.org/stable/621393?origin=crossref
[14] Hay, A.M., Johnston, R.J.: The study of process in quantitative human geography. Espace géographique **12**(1), 69–76 (1983). https://doi.org/10.3406/spgeo.1983.3801, https://www.persee.fr/doc/spgeo_0046-2497_1983_num_12_1_3801
[15] Herrera, M., Mur, J., Ruiz, M.: Detecting causal relationships between spatial processes: Detecting causal relationships between spatial processes. Papers in Regional Science **95**(3), 577–594 (Aug 2016). https://doi.org/10.1111/pirs.12144, https://onlinelibrary.wiley.com/doi/10.1111/pirs.12144
[16] Herrera-Gomez, M., Matilla-Garcia, M., Ruiz-Marin, M.: Spatial partial causality. The European Physical Journal Special Topics (Dec 2021). https://doi.org/10.1140/epjs/s11734-021-00378-5, https://link.springer.com/10.1140/epjs/s11734-021-00378-5
[17] Ibeling, D., Icard, T.: Probabilistic Reasoning Across the Causal Hierarchy. Proceedings of the AAAI Conference on Artificial Intelligence **34**(06), 10170–10177 (Apr 2020). https://doi.org/10.1609/aaai.v34i06.6577, https://aaai.org/ojs/index.php/AAAI/article/view/6577
[18] Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)
[19] Jean H.P. Paelinck, Jesús Mur: Some Issues on the Concept of Causality in Spatial Econometric Models. Estudios de Economia Aplicada **36**(1), 107–118 (2018)
[20] Kedron, P.: Can reproducible and replicable research facilitate causal explanation in geography? In: 2019). Accepted Short Papers and Posters from the 22nd AGILE Conference on Geo-information Science. Limassol, Chipre. Editorial, Stichting AGILE (2019)
[21] Kim, C., Daniels, M.J., Hogan, J.W., Choirat, C., Zigler, C.M.: Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls. The Annals of Applied Statistics **13**(3) (Sep 2019). https://doi.org/10.1214/19-AOAS1260, https://projecteuclid.org/journals/annals-of-applied-statistics/volume-13/issue-3/Bayesian-methods-for-multiple-mediators--Relating-principal-stratification-and/10.1214/19-AOAS1260.full
[22] Kim, C., Henneman, L.R.F., Choirat, C., Zigler, C.M.: Health effects of power plant emissions through ambient air quality. Journal of the Royal

Statistical Society: Series A (Statistics in Society) **183**(4), 1677–1703 (Oct 2020). https://doi.org/10.1111/rssa.12547, https://onlinelibrary.wiley.com/doi/10.1111/rssa.12547

[23] Kolak, M., Anselin, L.: A Spatial Perspective on the Econometrics of Program Evaluation. International Regional Science Review **43**(1-2), 128–153 (Jan 2020). https://doi.org/10.1177/0160017619869781, http://journals.sagepub.com/doi/10.1177/0160017619869781

[24] Kuhn, W.: Ontologies in support of activities in geographical space. International Journal of Geographical Information Science **15**(7), 613–631 (Oct 2001). https://doi.org/10.1080/13658810110061180, http://www.tandfonline.com/doi/abs/10.1080/13658810110061180

[25] Mark, D.M., Freksa, C., Hirtle, S.C., Lloyd, R., Tversky, B.: Cognitive models of geographical space. International Journal of Geographical Information Science **13**(8), 747–774 (Dec 1999). https://doi.org/10.1080/136588199241003, http://www.tandfonline.com/doi/abs/10.1080/136588199241003

[26] McGrath, S., Lin, V., Zhang, Z., Petito, L.C., Logan, R.W., Hernán, M.A., Young, J.G.: gfoRmula: An R Package for Estimating the Effects of Sustained Treatment Strategies via the Parametric g-formula. Patterns **1**(3), 100008 (Jun 2020). https://doi.org/10.1016/j.patter.2020.100008, https://linkinghub.elsevier.com/retrieve/pii/S2666389920300088

[27] van de Meent, J.W., Paige, B., Yang, H., Wood, F.: An introduction to probabilistic programming (2018). https://doi.org/10.48550/ARXIV.1809.10756, https://arxiv.org/abs/1809.10756

[28] Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)

[29] Perov, Y., Graham, L., Gourgoulias, K., Richens, J.G., Lee, C.M., Baker, A., Johri, S.: MultiVerse: Causal Reasoning using Importance Sampling in Probabilistic Programming. In: 2nd Symposium on Advances in Approximate Bayesian Inference. pp. 1–36 (2019)

[30] Pollmann, M.: Causal Inference for Spatial Treatments (Oct 2020), http://arxiv.org/abs/2011.00373, number: arXiv:2011.00373 arXiv:2011.00373 [econ, stat]

[31] Reich, B.J., Yang, S., Guan, Y., Giffin, A.B., Miller, M.J., Rappold, A.G.: A review of spatial causal inference methods for environmental and epidemiological applications (Jul 2020), http://arxiv.org/abs/2007.02714, number: arXiv:2007.02714 arXiv:2007.02714 [stat]

[32] Rogerson, P.: Spatial Statistical Methods for Geography. SAGE Publications, Ltd. (2021)

[33] Schuurman, N.: Critical gis: Theorizing an emerging science. Cartographica Monograph **53**(36), 4 (1999)

[34] Sharma, A., Kiciman, E., et al.: DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy (2019)

[35] Shpitser, I., Pearl, J.: Complete Identification Methods for the Causal Hierarchy. Journal of Machine Learning Research **9**, 1941–1979 (2008)

[36] Sinton, D.: The inherent structure of information as a constraint to analysis: mapped thematic data as a case study. In: Dutton, G. (ed.) Harvard Papers on Geographic Information Systems. vol. 6. Addison-Wesley, Reading, MA (1978)

[37] Smith, B., Mark, D.M.: Geographical categories: an ontological investigation. International Journal of Geographical Information Science **15**(7), 591–612 (Oct 2001). https://doi.org/10.1080/13658810110061199, http://www.tandfonline.com/doi/abs/10.1080/13658810110061199

[38] Tavares, Z., Zhang, X., Minaysan, E., Burroni, J., Ranganath, R., Lezama, A.S.: The random conditional distribution for higher-order probabilistic inference (2019)

[39] Worboys, M.F.: Object-oriented approaches to geo-referenced information. International Journal of Geographical Information Systems **8**(4), 385–399 (1994)