Journal of Geographical Systems (2024) 26:451–460
https://doi.org/10.1007/s10109-024-00452-1

ORIGINAL ARTICLE

Introduction to the special issue on spatial machine learning

Kevin Credit[1]

Received: 30 October 2024 / Accepted: 30 October 2024 / Published online: 15 November 2024

Abstract
While, many of the machine learning (ML) and artificial intelligence (AI) methods that are now commonly being used to answer questions across scientific disciplines have been around for some time, their widespread application to *spatial* data and *spatially-explicit* research questions is much more recent. The large number of excellent review papers and special issues in leading journals published in the last few years—which this issue of the *Journal of Geographical Systems* takes its place among—attest to the growing interest in the application and development of cutting-edge methodologies for spatial data. This editorial begins by proposing a new inclusive definition for spatial ML, then provides a brief overview of each of the six papers in this special issue, and ends with a suggestion of several possible directions for future research in spatial ML.

Keywords Spatial machine learning · Spatial data · Spatially-explicit models · GeoAI · Random forest

JEL Classification C14 · C18 · C21 · C45

1 Background

While many of the machine learning (ML) and artificial intelligence (AI) methods that are now commonly being used to answer questions across scientific disciplines have been around for some time (Rosenblatt 1958; Amari 1967; Openshaw and Openshaw 1997; Breiman 2001), their widespread application to *spatial* data and *spatially-explicit* research questions is much more recent. The large number of excellent review papers and special issues in leading GIScience journals published in the last few years (e.g., Janowicz et al. 2019; Nikparvar and Thill 2021; Kopczewska 2022; Papadakis et al. 2022)—which this issue of the *Journal of Geographical*

✉ Kevin Credit
kevin.credit@mu.ie

[1] Maynooth University, National University of Ireland Maynooth, Maynooth, Ireland

Systems takes its place among—attest to the growing interest in the application and development of cutting-edge methodologies for spatial data. This is of course in part due to the increasing volume, velocity, and variety of spatial data available for analysis, which require more powerful computation tools to analyse (Kitchin and McArdle 2016); it is also likely due to the demonstrated improvement in predictive performance for these methods compared to traditional statistical techniques (Hagenauer et al. 2019; Yoshida and Seya 2021; Credit 2022).

2 Defining spatial machine learning

However, despite (or, perhaps, due to) the increasing attention paid to new methods in the literature, we lack a coherent conceptual paradigm for discussing or defining *what we mean when we talk about spatial machine learning:* in other words, what methods and domains are included (and excluded)? How is spatial machine learning different than (or the same as) non-spatial ML or geographic artificial intelligence (GeoAI)? While, others have already effectively traced the history of the development of AI and ML methods in a spatial context (Janowicz et al. 2019; Hu et al. 2024) and classified and categorised the use of existing ML and AI methods for analysing spatial data (Nikparvar and Thill 2021; Kopczewska 2022), there is still a need to develop a cohesive and inclusive definition for "spatial machine learning" that provides a framework for describing the wide range of new methodological work in GIScience and allied fields, and, more specifically, the content and goals of this special issue.

To provide a useful definition of spatial ML, several terms need to be disentangled. First, machine learning vs. artificial intelligence: in common technical parlance, "artificial intelligence" refers to the very broad domain of developing "artificial system[s]…[that can] successfully achieve a novel goal through computational algorithms" (Gignac and Szodorai 2024), while "machine learning" is typically understood as a subset of AI that specifically deals with the "technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience and data" (Columbia Engineering 2024). In other words, ML concerns the development and use of statistical methods and algorithms that contribute to making machines "intelligent" or helping us better identify and explain patterns in data.

Of course, this definition for ML is exceptionally broad, and basically applies to all kinds of traditional statistical and econometric techniques. Indeed, in this context, ML has been famously referred to by statistician Robert Tibshirani as "glorified statistics," and the boundaries between these two disciplines are certainly fluid and, in many cases, overlapping (Hastie et al. 2016; Bennett et al. 2022). Undoubtedly, some of the ostensible distinction between machine learning and statistics is due more to contrasting perspectives rather than substance. These include semantic differences—in referring to, e.g., "labels," "algorithms," and "learning" as opposed to "dependent variables", "models", and "fitting"—and disciplinary perspectives, as packages for the implementation of ML are often designed by computer scientists (rather than statisticians) and intended to be used on particular types of problems—a

difference which can still be readily felt when assessing the output of, e.g., a linear regression in Python vs. R.

In addition to these somewhat stylistic differences, we can also categorise the methods themselves along two primary dimensions, which I believe capture the meaning of "machine learning" in the way the phrase is most commonly used in the literature and professional practice: *recency* and *specification*: recency relating to when each particular method was first developed and/or moved into widespread use (i.e., pre- and post-1990), and specification relating to the role that statistical assumptions about data distributions and other (embedded) parameters play in the method (i.e., parametric and non-parametric). In this case, "non-parametric" does not refer to a complete lack of parameters or assumptions in a model, as every statistical/ML model must make particular assumptions and specify parameters (or hyperparameters), often with very important implications for the results. It also is not a direct proxy for *unsupervised* learning methods. Instead, it is meant to capture methods that don't require a strong prior assumption about the distribution of the data (e.g., normality), which then strongly informs the patterns of prediction. Table 1 shows a rough categorisation of existing methods according to this two-dimensional typology, with primary methods employed by papers in this special issue marked in bold.

While, some may (rightfully) quibble with this typology (for instance, it ignores Bayesian approaches, as they can be applied to many of these methods) and my specific classification of the recency and specification for certain methods, I believe it provides a useful starting point for developing a framework for differentiating ML from

Table 1 Typology of statistical and machine learning methods. Primary methods used by papers in this special issue marked in bold

		Recency	
		Pre-1990	*Post-1990*
Specification	*Parametric*	Linear regression Principle Components Analysis (PCA) Logistic regression Spatial econometric methods	LASSO regression Geographically-weighted regression (GWR)
	Non-parametric	*k*-nearest neighbours Decision trees Hierarchical clustering Self-Organising Map (SOM) *k*-means clustering Early artificial neural networks (ANN)	Convolutional Neural Networks (CNN) Support Vector Machines (SVM) Random forest Contemporary deep learning approaches XGBoost Geographically-weighted ANN (GWANN) word2vec and contemporary natural language processing (NLP) methods Geographical random forest (GRF) (Spatial) Meta-learners Causal forest

"traditional" statistics. Based on the way the terms tend to be used, I would confidently associate the top left-hand quadrant with "traditional" statistics or econometrics, i.e., those methods developed primarily before 1990 with strong parametric and distributional assumptions. The bottom left-hand and top right-hand quadrants, then, can perhaps be thought of as methods working at the boundaries of ML, while the bottom right-hand quadrant contains "core" ML approaches, i.e., *recently-developed non-parametric* statistical methods or algorithms.

The final question, then, is what makes a particular ML method spatial? In this case, Kopczewska (2022) provides a useful distinction between approaches that use standard ML approaches *on spatial data* and new, *spatially-explicit methods* that are expressly designed to deal with the unique properties of spatial data. I am inclined to provide a similarly inclusive definition of spatial ML that includes, essentially, any kind of approach that explicitly considers (in the data or methodological design) spatial location, spatial relationships (operationalised, e.g., through the spatial weights matrix or other forms), or other properties inherent to spatial data such as spatial dependence and heterogeneity. This includes—as demonstrated by several contributions to this special issue—methods that incorporate spatial diagnostics and/or accuracy testing to account for, or measure, the influence of the unique properties of spatial data (e.g., spatial dependence). In this way, *spatiality* can be viewed as an additional layer on top of the typology displayed in Table 1: while, spatially-explicit methods have their own entries (and thus could be grouped together in a third dimension), *any* of these methods could be used *on spatial data* or with spatially-engineered features, which, in my view—as long as proper consideration and care is made for the unique properties of these data—equally constitutes a "spatial" model or analysis approach. In fact, as Kolak & Anselin (2020) importantly point out, in terms of understanding the true nature of spatial processes and effects, it can often be *more useful* to go "beyond the uncritical implementation of spatial tools or methods…[to] consider the inherently spatially and temporally dynamic, interactive nature of the populations being studied [with spatial data], and, as such, inform the initial design of the model" (p. 132), even if that involves using "simpler," "older," or "exploratory" methodological approaches. Thus we should not discount "standard" or "non-spatial" methods if they are applied in thoughtful (and novel) ways to spatial data, i.e., by leveraging the spatial dimension of the data explicitly in some way.

Thus, we can now come to the final definition which frames the purpose for this special issue: spatial machine learning approaches are statistical methods or algorithms for analysing patterns in spatial data that explicitly incorporate or leverage location, spatial relationships, spatial diagnostics/testing procedures, or other features inherent to spatial data—such as spatial dependence and heterogeneity. These ML approaches tend to be recently-developed (post-1990) and non-parametric (i.e., do not rely on strong assumptions about the distribution of the data).

3 Overview of the special issue

The idea for this special issue originated at the 2021 and 2022 Meetings of the North American Regional Science Council (NARSC), in special sessions co-organised by the author on "Machine Learning in Regional Science: Perspectives, Methods, and Applications," which form part of an ongoing series (co-)organised by the author at NARSC from 2020 to 2024, and, beginning in 2024, the European Regional Science Association (ERSA). We received a number of high-quality submissions to the special issue, several of which originated in these NARSC sessions, including Carruthers and Wei's "What Drives Urban Redevelopment Activity? Evidence from Machine Learning and Econometric Analysis in Three American Cities" (*current issue*), Credit and Lehnert's "A structured comparison of causal machine learning methods to assess heterogeneous treatment effects in spatial data" (2023), and Tepe's "A random forests-based hedonic price model accounting for spatial autocorrelation" (2024). In addition to these papers, three additional submissions were included in this special issue: Kim et al.'s "Beyond visual inspection: capturing neighbourhood dynamics with historical Google Street View and deep learning-based semantic segmentation" (2023), Lotfata and Georganos's "Spatial machine learning for predicting physical inactivity prevalence from socioecological determinants in Chicago, Illinois, USA" (2023), and Kilic et al.'s "Unveiling the impact of machine learning algorithms on the quality of online geocoding services: a case study using COVID-19 data" (2024).

Interestingly, all of the papers in this issue are, at some level, focused on *testing the performance* of spatial ML approaches by comparing them to "traditional" statistical, spatial, or non-spatial ML methods. This includes (1) examining extensions to existing methods to incorporate ML estimation procedures and account for the unique properties of spatial data (Lotfata and Georganos 2023; Credit and Lehnert 2023), (2) the evaluation of spatially-informed validation measures (Tepe 2024), and (3) applications of "non-spatial" ML methods to spatial data in novel contexts (Kim et al. 2023; Carruthers and Wei 2024; Kilic et al. 2024). Taken as a whole, the results of the papers in this issue point to high levels of performance and potential usefulness for spatial ML approaches when compared to "traditional" approaches. In addition, these papers provide useful empirical results on a range of topics, including the social and environmental determinants of health, the effect of public transport infrastructure on reducing emissions, hedonic house price analysis, urban redevelopment, and geocoding for COVID-19 patient locations.

In the first paper, Lotfata and Georganos (2023) apply the newly-developed geographical random forest (GRF) method to census tract-level data related to physical inactivity and its social and environmental determinants in Chicago. The GRF works in a similar fashion to geographically-weighted regression (GWR), estimating a local random forest (RF) model at each tract based on an optimal kernel, and thus provides a window into spatially-heterogeneous relationships, while leveraging the (non-linear) estimation characteristics of the RF. Most importantly, the paper demonstrates the usefulness of the GRF approach when

analysing urban areal data compared to other methods: the GRF provides a higher overall predictive performance than multiscale GWR (MGWR) and RF while, nearly reaching the accuracy of the geographically-weighted artificial neural network (GWANN)—however, in terms of interpretability, the GRF surpasses the GWANN in that it produces *local feature importances* and goodness-of-fit measures, which can help researchers understand the specific nature of the spatial heterogeneity in their data.

The second paper, by Credit and Lehnert (2023), is similar in that it spatializes existing ML approaches and tests their performance against traditional and non-spatial methods. In this case, the focus is on creating and examining *spatially-informed approaches to causal inference* using ML. The paper develops the "spatial" T-learner (STL) and causal forest (CF) methods by including spatial lags of the dependent and independent variables, i.e., ML approximations of traditional spatial econometric specifications, and tests their performance against non-spatial versions of ordinary least squares (OLS) regression, CF, and the T-learner. Performance in this case means the ability of the model to ascertain the overall average treatment effect and the individual unit-level treatment effects of some urban intervention; however, since these effects are not directly observed in empirical data, the paper develops a unique method for simulating spatially-realistic data, where the treatment (and confounding) effects are known. Using this framework, the spatial ML methods generally outperform the others, with the spatial "Durbin" CF performing the best. This specification is then applied to the case of estimating the treatment effects from building a new light rail line on neighbourhood-level CO_2 emissions, finding a substantial and spatially-varying effect—which is then further explained by regressing the unit-level treatment effects against the independent variables—demonstrating the usefulness of the spatial CF for understanding fine-grained spatial causal effects.

The third paper, from Tepe (2024), focuses expressly on the application of spatial validation procedures for ML methods. The problem posed by the paper—in the context of using an RF model for hedonic house price analysis at the parcel level in Miami-Dade County, FL—is that standard cross-validation metrics provide overly-optimistic estimates of model performance when spatial dependence is present in the data. This has prompted the development and use of spatial cross-validation approaches such as the spatial blocking *k*-fold and spatial "leave-one-out." However, as the paper importantly demonstrates, these approaches need to be paired with a model specification that explicitly accounts for spatial dependence (again, when it is present in the data), e.g., by including the spatial lag of the dependent variable, to improve the accuracy and stability of the predictions provided by spatial cross-validation measures. In other words, non-spatial cross-validation measures are *too optimistic* when spatial dependence is present, but spatial cross-validation measures are *not optimistic enough*—with high variance—when spatial dependence is present but unaccounted for in the model.

In the fourth paper, Kim et al. (2023) test the accuracy of a deep learning-based semantic segmentation approach for classifying changes in the built environment from historical Google Street View (GSV) imagery. This paper uses the Deeplabv3 + CNN model, which has been pre-trained to identify the number of pixels associated with a number of urban environmental features, including buildings,

sky, pavement, vegetation, etc. It then applies this model to GSV images of areas in Santa Ana, CA in which buildings have been newly constructed or demolished (i.e., substantially changed), collecting a pre- to post-construction % change in building pixels and comparing that to the same pre- to post-construction time period in various control areas, where buildings were not constructed or demolished. The paper finds that the method—with a particular set of input parameters—is generally pretty good at identifying built environment changes (75% true positive rate), there remains room to improve, particularly when it comes to false positive identification. Beyond this specific finding, this paper is particularly valuable for its novel application to spatialized imagery data, i.e., built environment change detection using GSV, and its development of a systematic, spatially-grounded accuracy testing procedure that provides more useful conclusions than true positive/false negative alone.

Continuing an interest in identifying and measuring urban redevelopment activity, Carruthers and Wei (2024) compare the predictive performance and explanatory information provided by ML models—including *k*-nearest neighbours (KNN), RF, and a cost-sensitive RF—and a traditional probit model for predicting changes in parcel boundaries in Seattle, Chicago, and Boston. In general, the results indicate that the ML models (in particular, the cost-sensitive RF) slightly outperform the probit model, especially in terms of recall; however, the probit model produces regression coefficients and statistical significances that are the useful in interpreting the direction of the relationship between individual covariates and the dependent variable, and thus in explaining and understanding the processes that drive urban redevelopment in the three cities. For this reason, the authors suggest combining these methodological approaches—along with theoretical knowledge of the drivers of urban redevelopment processes—in future research.

Finally, Kilic et al. (2024) evaluate the effectiveness of using ML techniques for enhancing the accuracy of geocoding results in Turkey. This paper takes a very interesting approach by training a range of ML models (e.g., RF, support vector machines, XGBoost, and others) using a selection of text-matching metrics (e.g., character-based, term-based, and hybrid-based) on a set of known addresses for COVID-19 patients in order to predict address matches (coded as a dummy variable, with 1 = an exact match for neighbourhood, street name, number, district, and province name). The results indicate that the ML classifiers in general—and, in particular, the RF model with every type of text feature included—increased the accuracy of conventional geocoders substantially, from 79 to 86%, 52 to 91%, 80 to 82%, and 83 to 85% for ArcGIS Online, Bing Maps, Google Maps, and HERE maps, respectively (based on mean area under the curve scores).

4 Looking to the future

Spatial ML is a rapidly-growing field with significant avenues for continued development by researchers across the geographically-aligned sciences. Fundamentally, as the articles in this special issue attest, these methods appear to offer, at the very least, incremental—and in some cases, truly substantial—improvements in predictive performance compared to "traditional" statistical and

"non-spatial" ML methods across a wide range of spatially-informed applications, data, and contexts. This potential to improve existing quantitative analysis of social, economic, and physical phenomena—and, through that, foster more equitable and sustainable decision-making—should drive continued interest in creating, extending, and using these approaches for both explanatory and predictive purposes. At the same time, many of the most cutting-edge methods, including large language models (LLM) and other recent AI approaches, are mostly being developed and applied outside of a spatially-explicit context or awareness, despite the inherently spatial nature of many of the world's most pressing scientific concerns (and the data used to study them). This presents a remarkable—and consequential—opportunity for geographers, regional scientists, urban planners, and other spatially-aligned researchers to contribute to the development of new state-of-the-art tools that explicitly integrate spatial features and spatial ways of thinking. This opportunity also calls for an increase in cross-disciplinary partnerships with computer science and other disciplines, as the field of spatial ML has an inherently interconnected and symbiotic relationship with the development and application of "core" ML. There is certainly much to be gained from such collaborations—in both directions—including better integrating spatial ways of thinking into the design of new methods (from the outset), and better contextualising research questions and applications with spatially-embedded critical, theoretical, and substantive knowledge.

While, the range of topics for future research related to spatial ML is vast—and advancing so quickly that any list is likely to be outdated in a short time—I will end this editorial by providing a sample of some general possible future directions for research related to spatial ML from across the methodological and disciplinary spectrum:

- The integration of spatial data and approaches into natural language processing (NLP) and large language models (LLM), including for enhanced spatial querying, mapping, and geocoding;
- New methods for optimizing spatial pattern prediction and spatial transfer learning to improve prediction accuracy, particularly in situations with little existing context for a particular pattern (e.g., prediction of climate change induced-flooding and land use change in places with little previous history of flooding);
- Further development and use of graph embedding approaches for understanding and analysing spatial linkages, relationships, and networks;
- Development of new indicators of spatial association;
- Further development of explanatory ML metrics in a spatial context, e.g., Local Interpretable Model-agnostic Explanation (LIME) and Shapley Additive Explanations (SHAP);
- Further development of visualization methods for non-linear relationships in ML models, e.g., partial dependence (PD) and accumulated local effects (ALE) plots; and
- Further development of causal and explanatory spatial ML methods, e.g., causal forests (CF), meta-learners, deep gravity models, and feature representation learning approaches.

References

Amari S (1967) A theory of adaptive pattern classifier. IEEE Trans 16:279–307

Bennett, M., Hayes, K., Kleczyk, E. J., & Mehta, R. (2022). Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. arXiv. https://arxiv.org/abs/2201.02469.

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Carruthers J, Wei H (2024, current issue) What drives urban redevelopment activity? evidence from machine-learning and econometric analysis in three American cities. J Geogr Syst. https://doi.org/10.1007/s10109-024-00451-2

Columbia Engineering (2024) Artificial intelligence (AI) vs. machine learning. Columbia University. https://ai.engineering.columbia.edu/ai-vs-machine-learning/.

Credit K (2022) Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. Geogr Anal 54(1):58–83

Credit K, Lehnert M (2023) A structured comparison of causal machine learning methods to assess heterogeneous treatment effects in spatial data. J Geogr Syst. https://doi.org/10.1007/s10109-023-00413-0

Gignac GE, Szodorai ET (2024) Defining intelligence: bridging the gap between human and artificial perspectives. Intelligence 104:101832

Hagenauer J, Omrani H, Helbich M (2019) Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany. Int J Geogr Inf Sci 33:1399–1419

Hastie T, Tibshirani R, Friedman J (2016) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, Berlin

Hu Y, Goodchild M, Zhu AX, Yuan M, Aydin O, Bhaduri B et al (2024) A five-year milestone: reflections on advances and limitations in GeoAI research. Ann GIS 30(1):1–14. https://doi.org/10.1080/19475683.2024.2309866

Janowicz K, Gao S, McKenzie G, Hu Y, Bhaduri B (2019) GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. Int J Geogr Inf Sci 34(4):625–636. https://doi.org/10.1080/13658816.2019.1684500

Kilic B, Bayrak OC, Gülgen F et al (2024) Unveiling the impact of machine learning algorithms on the quality of online geocoding services: a case study using COVID-19 data. J Geogr Syst. https://doi.org/10.1007/s10109-023-00435-8

Kim JH, Ki D, Osutei N et al (2024) Beyond visual inspection: capturing neighborhood dynamics with historical Google Street View and deep learning-based semantic segmentation. J Geogr Syst. https://doi.org/10.1007/s10109-023-00420-1

Kitchin R, McArdle G (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. Big Data Soc 1–10

Kolak M, Anselin L (2020) A spatial perspective on the econometrics of program evaluation. Int Reg Sci Rev 43(1–2):128–153

Kopczewska K (2022) Spatial machine learning: New opportunities for regional science. Ann Region Sci 68:713–755

Lotfata A, Georganos S (2023) Spatial machine learning for predicting physical inactivity prevalence from socioecological determinants in Chicago, Illinois, USA. J Geogr Syst. https://doi.org/10.1007/s10109-023-00415-y

Nikparvar B, Thill J-C (2021) Machine learning of spatial data. Int J Geo-Inf. https://doi.org/10.3390/ijgi10090600

Openshaw S, Openshaw C (1997) Artificial intelligence in geography. Wiley

Papadakis E, Adams B, Gao S, Martins B, Baryannis G, Ristea A (2022) Explainable artificial intelligence in the spatial domain (X-GeoAI). Transactions in GIS 26:2413–2414. https://doi.org/10.1111/tgis.12996

Rosenblatt F (1958) The perceptron—a probabilistic model for information-storage and organization in the brain. Psychol Rev 65(6):386–408

Tepe E (2024, current issue), A random forests-based hedonic price model accounting for spatial autocorrelation. J Geogr Syst. https://doi.org/10.1007/s10109-024-00449-w

Yoshida, T., & Seya, H. (2021). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. arXiv. https://arxiv.org/abs/2107.12539.